ICCR 2024

XXth International Conference on the use of Computers in Radiation therapy



Proceedings

8 - 11 July 2024

Lyon - France

20th edition

www.iccr2024.org

Table of Contents

Treatment Planning and Dose Monitoring - Co-chairs: *Peter Remeijer, Wolfgang Tomé* **18**

Treatment Error Classification With Time-Resolved EPID Images With Multiple Instance Learning.
Vyacheslav Yarkin, Evelyn De Jong, Rutger Hendrix, Frank Verhaegen, Cecile Wolfs 18

A Novel Iterative Reconstruction Algorithm for Imaging of Multi-Field Proton Treatments With PET.
Keegan McNamara, Marina Béguin, Günther Dissertori, Judith Flock, Cristian Fuentes, Jan Hrbacek, Tony Lomax, Shubhangi Makkar, Christian Ritzer, Damien C. Weber, Carla Winterhalter 22

Evaluation of Algorithms for Dose Reconstruction From Prompt-Gamma Radiation in Proton Therapy.
Beatrice Foglia, Chiara Gianoli, Takamitsu Masuda, Elisabetta De Bernardi, Thomas Bortfeld, Katia Parodi, Marco Pinto 26

An Analytical Approach to Predict 3D Positron Emitter Distribution in Carbon Ion Therapy.
Tianxue Du, Julia Bauer, Katia Parodi, Marco Pinto 30

Simultaneous Optimization of Multiple Plans Within One Treatment Course for Temporally Feathered Radiation Therapy.
Jenny Bertholet, Julius Arnold, Chenchen Zhu, Gian Guyer, Silvan Mueller, Marco F M Stampanoni, Peter Manser, Michael K Fix 34

The Relationship Between Treatment Plan Complexity and Dosimetry Quality: An Analysis of Three Stereotactic Treatment Planning Challenges.
Nick Hardcastle, Mathieu Gaudreault, Alisha Moore, Benjamin Nelms, Jordi Saez, Victor Hernandez 38

Addressing the Gap in Oncologist Training: Generating Clinically Realistic, Suboptimal Dose Distributions Through Geometry Aware Convolutions.
Skylar Gay, Mary Gronberg, Raymond Mumme, Christine Chung, Meena Khan, Chelsea Pinnix, Sanjay Shete, Brent Parker, Tucker Netherton, Carlos Cardenas, Laurence Court 42

Image Segmentation I - Co-chairs: *Andre Dekker, Gerd Heilemann* **46**

Feasibility of Training Federated Deep Learning Oropharyngeal Primary Tumour Segmentation Models Without Sharing Gradient Information.
Leroy Volmer, Ananya Choudhury, Aiara Lobo Gomes, Andre Dekker, Leonard Wee 46

Follow the Leader - a General Template-Guided Deep Learning Framework for Adaptive Radiotherapy Target Delineation.
Mahdieh Kazemimoghadam, Qingying Wang, Mingli Chen, Xuejun Gu, Weiguo Lu 50

Clarifying the Dosimetric Impact of Autosegmentation Inaccuracies.
Joëlle Van Aalst, Tomas Janssen, Ilse Van Bruggen, Jelmer Wolterink, Roel Steenbakkers, Stefan Both, Johannes Langendijk, Charlotte Brouwer 54

Team Work Makes the Dream Work: An Ensemble Learning Approach for Segmentation With Convolutional Neural Networks and Vision Transformers.
Anne Andresen, Slavka Lucakova, Yasmin Lassen-Ramshad, Christian Rønn Hansen, Nadine Vatterodt, Jesper Kallehauge 58

Data Harvesting vs Data Farming: A Study of the Importance of Variation vs Sample Size in Deep Learning-Based Auto-Segmentation for Breast Cancer Patients.
Emma Skarsø Buhl, Else Maae, Louise Wichmann Matthiessen, Mette Holck Nielsen, Maja Maraldo, Mette Møller, Stine Elleberg, Sami Aziz-Jowad Al-Rawi, Birgitte Vrou Offersen, Stine Korreman 62

Development and External Validation of a DL-Based CTV Segmentation Model on a Large Multicentric Dataset in the Case of Whole-Breast Radiotherapy Treatment.
Maria Giulia Ubeira Gabellini, Gabriele Palazzo, Martina Mori, Alessia Tudda, Luciano Rivetti, Elisabetta Cagni, Roberta Castriconi, Valeria Landoni, Eugenia Moretti, Aldo Mazzilli, Caterina Oliviero, Lorenzo Placidi, Giulia Rambaldi Guidasci, Cecilia Riani, Andrei Fodor, Nadia Gisella Di Muzio, Robert Jeraj, Antonella Del Vecchio, Claudio Fiorino 66

Leveraging Network Uncertainty to Identify Regions in Rectal Cancer CTV Auto-Segmentations Likely Requiring Manual Edits.
Federica Carmen Maruccio, Rita Simões, Charlotte Brouwer, Jan-Jakob Sonke, Tomas Janssen 70

Prediction Models I - Co-chairs: *Angela Davey, Stine Korreman* **74**

Virtual Patient-Specific Quality Assurance: a Prospective Two-Year Analysis.
Alon Witztum, Phillip Wall, Emily Hirata, Gilmer Valdes 74

Model Size Reduction in CNNs Through Sparse Regularization: Balancing Efficiency and Diagnostic Accuracy in Medical Imaging.
Carlos Garcia Fernandez, Wilmer Arbelo-Gonzalez, Jerome Friedman, Gilmer Valdes 77

The Implementation of Gaussian Process Regression to Inform Voxel-Based Analysis When Dealing With Sparse and Irregular Time-Series Longitudinal Data.
Jane Shortall, Eliana Vasquez Osorio, Andrew Green, David Wong, Tanuj Puri, Peter Hoskin, Ananya Choudhury, Marcel Van Herk, Alan McWilliam 81

Voxel-Wise Causal Inference to Detect Location and Dose-Response of Organs at Risk in Lung Cancer.
Marcel Van Herk, Azadeh Abravan, Alan McWilliam, Eliana Vasquez Osorio 85

Pelvic Bone Marrow Dose-Volume Predictors of Late Lymphopenia Following Pelvic Lymph Node Radiation Therapy for Prostate Cancer.
Maddalena Pavarini, Lisa Alborghetti, Stefania Aimonetto, Angelo Maggio, Valeria Landoni, Paolo Ferrari, Antonella Bianculli, Edoardo Petrucci, Alessandro Cicchetti, Bruno Farina, Paolo Salmoiraghi, Eugenia Moretti, Barbara Avuzzi, Tommaso Giandini, Fernando Munoz, Alessandro Magli, Giuseppe Sanguineti, Justyna Magdalena Waskiewicz, Luciana Rago, Domenico Cante, Giuseppe Girelli, Vittorio Vavassori, Nadia Gisella Di Muzio, Tiziana Rancati, Cesare Cozzarini, Claudio Fiorino 88

Image Registration - Co-chairs: *Kristy Brock, Jamie McClelland* **92**

Inter-Patient Registration Methods for Voxel-Based Analysis in Lung Cancer.
Chloe Min Seo Choi, Jue Jiang, Maria Thor, Joseph Deasy, Harini Veeraraghavan 92

Modelling Progressive Anatomical Changes and Their Variability During Radiotherapy Treatment.
Poppy Nikou, Eliana M. Vasquez Osorio, Marcel Van Herk, Teresa Guerrero Urbano, Anna Thompson, Andrew Nisbet, Sarah Gulliford, Jamie McClelland 96

An Anatomically-Informed Correspondence Initialisation Method to Improve Learning-Based Registration for Radiotherapy.
Edward G. A. Henderson, Marcel Van Herk, Andrew F. Green, Eliana M. Vasquez Osorio 100

DiffuseRT: Learning Likely Deformations of Head and Neck Radiotherapy Patients From Data.
Luciano Rivetti, Andreas Smolders, Nadine Vatterodt, Stine Korreman, Antony Lomax, Manju Sharma, Andrej Studen, Damien Weber, Francesca Albertini, Jeraj Robert 104

Choice of Colourmaps for Radiotherapy.
Alex Davies, Mark Fleckney, Marcel Van Herk 108

Radiomics and Voxel-Based-Analysis - Co-chairs: *Mathieu Hatt, Charlotte Robert* **112**

Genetically-Based Analysis of Dose Surface Maps for Assessing Toxicity Post-RT.
Tiziana Rancati, Eliana Gioscio, Michela Carlotta Massi, Nicola Rares Franco, Petra Seibold, Barbara Avuzzi, Alessandro Cicchetti, Ananya Choudhury, Barry Rosenstein, David Azria, Dirk De Ruysscher9, Maarten Lambrecht, Elena Sperk, Christopher J Talbot, Ana Vega, Liv Veldeman, Adam Webb, Paolo Zunino, Anna M Paganoni, Francesca Ieva, Andrea Manzoni, Sarah L Kerns, Alison Dunning, Jenny Chang-Claude, Catharine L West 112

Prediction of Survival Outcomes and Relevant Tumor Immune Infiltrates Using Radiomic Network Analysis.
Jung Hun Oh, Aditya Apte, Harini Veeraraghavan, Amita Shukla-Dave, Joseph Deasy 116

Radiomic Extraction and Analysis for DICOM Images to Refine Objective Quality Control (READII-2-ROQC): An Open-Source Foundation for Head and Neck Radiomics.
Katy L. Scott, Sejin Kim, Jermiah J. Joseph, Matthew Boccalon, Mattea Welch, Mogtaba Alim, Umar Yousafzai, Ian Smith, Chris McIntosh, Katrina Rey-McIntyre, Shao Hui Huang, Tirth Patel, Tony Tadic, Brian O'sullivan, Scott V. Bratman, Andrew J. Hope, Benjamin Haibe-Kains . . . 119
Predicting Radiation Pneumonitis Using Pre-Radiotherapy CT Scans by Radiomic and a Pre-Trained CNN Based Feature Extraction.
Matthew Gil, William Nailon, Stephen Harrow, Paul Murray, Sorcha Campbell, Ai Wain Yong, John Murchison, Gillian Ritchie, Stephen Marshall . . . 123
Integrating Radiomics Harmonization and Deep Survival Analysis for Enhanced Predictive Modeling in a Multicentric Context.
Nassib Abdallah, Jean-Marie Marion, Pierre Chauvet, Clovis Tauber, Thomas Carlier, Mathieu Hatt . . . 127
Can Bootstrapping Increase Robustness of Voxel-Based-Analysis?
Alan McWilliam, Marcel Van Herk, Chloe Pratt, Eliana M. Vasquez Osorio . . . 131

Hadrontherapy Planning - Co-chairs: *Francesca Albertini, George Dedes* **135**
Trading Robustness? Introducing Total Variance Objectives to Pareto Surface Approximation Techniques for Radiotherapy.
Remo Cristoforetti, Philipp Süss, Niklas Wahl . . . 135
Mid-Range Proton Arc Therapy.
Qingying Wang, Mingli Chen, Mahdieh Kazemimoghadam, Xuejun Gu, Weiguo Lu . . . 139
From Concept to Clinic: How Our In-House Automated Planning System Augmented the Commercial TPS for Treating Over 10,000 Patients.
Masoud Zarepisheh, Gourav Jhanwar, Ying Zhou, Qijie Huang, Hai D. Pham, Jie Yang, Laura Cervino, Jean M Moran, Joseph O. Deasy, Linda Hong . . . 143
Fully-Automated Integrated Multi-Criterial Optimisation of Transmission Pencil Beams in IMPT for Penumbra Sharpening.
Wens Kong, Merle Huiskes, Steven Habraken, Eleftheria Astreinidou, Coen Rasch, Ben Heijmen, Sebastiaan Breedveld . . . 147
Interactive Dose Modification: a Novel Approach to Proton Beam Therapy Treatment Planning.
Matthew Lowe . . . 151
Simultaneous Reduction of Number of Spots and Energy Layers in Intensity Modulated Proton Therapy for Rapid Spot Scanning Delivery.
Anqi Fu, Vicki Taasti, Masoud Zarepisheh . . . 155

Poster Session, Odd IDs **159**
1 - Ameliorate Immune Suppression During Radiation Therapy to Advanced Stage Lung Cancer via a Predictive Algorithm.
Krishni Wijesooriya, Cam Nguyen, Grant Williams, Marco Baessler, Constance Hodge, John Morgan, Ryan Gentzler, Nilanga Liyanage . . . 159
3 - Comparing Two Methods of Using Toxicity Outcomes to Study Dose-Response Mapping of Bladder and Rectum in Prostate Cancer Patients Undergoing Radiotherapy.
Tanuj Puri, Marcel Van Herk, Eliana Vasquez Osorio, Tiziana Rancati, Petra Seibold, Eliana Gioscio, Andrew Green, Andrew P. Morris, Jane Shortall, Sarah Kerns, David Azria, Marie-Pierre Farcy-Jacquet, Jenny Chang-Claude, Alison Dunning, Maarten Lambrecht, Barbara Avuzzi, Dirk De Ruysscher, Elena Sperk, Christopher Talbot, Adam Webb, Ana Vega, Liv Veldeman, Barry Rosenstein, Catharine West, Ananya Choudhury, Peter Hoskin, Alan McWilliam . . . 163
5 - A Technological Pipeline for Detecting Radiation-Induced Heart Damage Mechanisms and Prevent Major Cardiac Adverse Events.
Alessandro Cicchetti, Alessandra Catalano, Silvia Meroni, Anna Cavallo, Albina Allajbej, Chiara Veronese, Paola Vallerio, Claudia Sangalli . . . 167
7 - Whole-Body Organ Dose Assessment of Cancer Patients Undergoing External Beam Radiotherapy Using Body-Size-Dependent Phantom Library for Secondary Cancer Risk Prediction.
Kyeong Hyeon Sung, Bo-Wi Cheon, Hyung-Joo Choi, Soo Min Lee, Saerom Seong, Yoon Seo, Donghyouk Kim, Hyun Joon Choi, Chul Hee Min . . . 171

9 - Towards Gaussian Processes Modelling to Study the Late Effects of Radiotherapy in Children and Young Adults With Brain Tumours..
Angela Davey, Arthur Leroy, Eliana M. Vasquez Osorio, Kate Vaughan, Peter Clayton, Marcel Van Herk, Mauricio A Alvarez, Martin McCabe, Marianne Aznar . . . 175
11 - Explainable Artificial Intelligence for the Prediction of Non-Small Cell Lung Carcinoma Recurrence.
Lorna Tu, Herve Choi, Haley Clark, Samantha Lloyd . . . 179
13 - Is Comorbidity Associated With Interruption of Radiotherapy in Patients With HNSCC Cancer?
Golnoosh Motamedi-Ghahfarokhi, Azadeh Abravan, Eliana M. Vasquez Osorio, Marcel Van Herk, James Price, Hitesh Mistry, Gareth Price . . . 183
15 - Prediction of Acute Pulmonary Toxicity Events With 3D Convolutional Neural Networks From Radiotherapy Dose Maps.
Pedro Juan Soto Vega, Vincent Bourbonne, Wistan Marchadour, Gustavo Andrade-Miranda, François Lucia, Martin Rehn, Ulrike Schick, Dimitris Visvikis, Franck Vermet, Mathieu Hatt . . . 186
17 - Deep Learning Based Three-Dimensional Radiation Dose Prediction With Dose Level Guidance - a Spine SBRT Example.
Tushar Shinde, Gregory Arthur Szalkowski, Cynthia Chuang, Lei Wang, Lianli Liu, Scott Soltys, Erqi Pollom, Elham Rahimy, Quoc-Anh Ho, Neelan Marianayagam, David Park, Ulas Yener, Antonio Meola, Hao Jiang, Mingli Chen, Zi Yang, Weiguo Lu, Xuejun Gu . . . 190
19 - Can Input Reconstruction Be Used to Directly Estimate Uncertainty of a Dose Prediction U-Net Model?
Margerie Huet Dastarac, Dan Nguyen, Steve Jiang, John Lee, Ana Barragan-Montero . . . 193
21 - Investigation of Unsupervised Race-Based Predictive Modeling Strategies to Improve Prediction Power of Tumor Recurrence in Head and Neck Cancer Patients in Diverse Racial Groups.
Hassan Bagher-Ebadian, Marissa V. Gilbert, Farzan Siddiqui, Benjamin Movsas, Kundan Thind, Indrin J. Chetty . . . 197
23 - Robust Models for Esopohageal Toxicity Prediction From Radiation Dose Maps.
Matthew Gil, William Nailon, Stephen Harrow, Paul Murray, Stephen Marshall . . . 201
25 - Improving Dose Volume Histogram (DVH) Based Analysis of Clinical Outcomes Using Modern Statistical Techniques. A Systematic Answer to Multiple Comparisons Concerns.
Mirek Fatyga, Jiuyun Hu, Jing Li . . . 205
27 - On the Implementation and Evaluation of Loss Functions for Robust Multiple Anatomy Segmentation on CT Images.
Chengyin Li, Rafi Ibn Sultan, Hassan Bagher-Ebadian, Yao Qiang, Kundan Thind, Indrin Chetty . . . 209
29 - Scripted Dosimetric Validation of AI Auto-Contouring in Prostate Radiotherapy.
Daniel Warren, Adam Chalkley, Ian Stronach . . . 213
31 - Implementing a Simple, Dynamic and Controllable AI Delineation Tool in Clinical Settings.
Nis Sarup, Maximilian Konrad, Marie Louise Madsen Lange Sevdal, Simon Long Krogh, Christian Rønn Hansen, Irene Hazell, Ebbe Laugaard Lorenzen, Carsten Brink . . . 216
33 - A Robust Auto-Segmentation Approach to Accelerate Adaptive MRI-Guided Radiotherapy for Pancreatic Cancer Patients.
Mehdi Shojaei, Bjoern Eiben, Jamie McClelland, Simeon Nill, Alex Dunlop, Uwe Oelfke . . . 220
35 - A National Infrastructure for Sharing AI Segmentation Tools.
Simon Long Krogh, Ruta Zukauskaite, Ebbe Lorenzen, Nis Sarup, Maximilian Konrad, Janne Nørlykke Drudgaard, Mette Klüver-Kristensen, Eva Samsøe, Mohammad Farhadi, Thomas Hinz-Berg Johansen, Carsten Brink, Christian Rønn Hansen . . . 224
37 - Use of Complementary Multispectral Data for Improving the Accuracy of Deep Learning Solution for Bioluminescence Tomography Reconstruction in Glioblastoma Tumor Models.
Behzad Rezaeifar, Cecile Wolfs, Natasja Lieuwes, Rianne Biemans, Brigitte Reniers, Ludwig Dubois, Frank Verhaegen . . . 228
39 - Monte Carlo Photons Simulation From Alpha-Emitter Radionuclides.
David Sarrut, Ane Etxebeste, Jean Létang . . . 232
41 - A Full GATE/Geant4 Model of a Helical Radiotherapy Device.
Sheraz Blas, Luc Simon . . . 235
43 - Impact of Temporal Structure of Electron, Proton and Photon on (H2O2) Concentration Using Montecarlo Simulation Toolkit (TOPAS-NBio).
Mustapha Chaoui, Othmane Bouhali, Tayalati Yahya . . . 239

45 - Log File-Based Patient Specific QA Using GPU-Accelerated Monte Carlo Simulation for Proton Pencil Beam Line Scanning Therapy.
Chanil Jeon, Sung Hwan Ahn, Jungwook Shin, Hoyeon Lee, Youngyih Han . . . 243
47 - A Fast 3D Range-Modulator Delivery Approach: Sensitivity of the Resulting Dose Distribution on Potential Setup Misalignments.
Yuri Simeonov, Miriam Krieger, Michael Folkerts, Gerard Paquet, Pierre Lansonneur, Petar Penchev, Klemens Zink, Uli Weber . . . 245
49 - Benchmark of Geant4 Hadronic Models for Carbon-Therapy Within the CLINM Project.
Marie Vanstalle, Lévana Gesson, Claire Reibel, Nicolas Arbor, Aurélia Arnone, Christian Finck, Catherine Galindo, Stéphane Higueret, Thé-Duc Lê, Arshiya Sood, Quentin Raffy, Philippe Peaupardin, Marco Pullia 248
51 - Monte Carlo Modeling of Elekta VERSA HD for Out-of-Field Dose Characterization.
Maxime Jacquet, Thomas Baudier, Eric Deutsch, Charlotte Robert, David Sarrut . . . 252
53 - Deformable Voxel Geometry for Photon and Electron Monte Carlo Dose Calculations.
Björn Zobrist, Hannes A Loebner, Jenny Bertholet, Daniel Frei, Werner Volken, Florian Amstutz, Marco F M Stampanoni, Peter Manser, Michael K Fix . . . 256
55 - A GPU-Based Patient-Specific CT Dosimetry Tool Combining Monte Carlo Calculations and ML-Based Segmentations.
Ronan Lefol, Yannick Lemaréchal, Jonathan Boivin, Philippe Despres . . . 260
57 - Accuracy Comparison Between GATE and TOPAS for 6MV Varian CLINAC IX.
Zakaria Ait Elcadi, Othmane Bouhali . . . 264
59 - Can Low-Cost Magnetic Resonance Distortion Phantoms Be 3D Printed?
Haley Clark . . . 268
61 - Automatic Contour Quality Assurance (CAT-QA) to Speed up Online Adaptive Radiotherapy.
Lisa Fankhauser, Andreas Smolders, Renato Bellotti, Damien Charles Weber, Antony John Lomax, Francesca Albertini . . . 271
63 - QATrack+: An Open-Source Routine Quality Assurance Platform.
Randle Taylor, James Kerns . . . 275
65 - Big Data Methods for CT Doses Analysis and Monitoring.
Pierre-Luc Asselin, Philippe Despres, Jonathan Boivin, Yannick Lemaréchal, Gabriel Couture, Samuel Ouellet 279
67 - AutoMRISimQA: an Automated System for Daily Quality Control of a 3T MRI Simulator.
Aitang Xing, Gary Goozee, Gary Liney, Sankar Arumugam, Shrikant Deshpande, Anthony Espinoza, Alison Gray, Vasilis Kondilis, Doaa Elwadia, Robba Rai, Phillip Chlap, Lois Holloway . . . 283
69 - NoPAUSE: Feasibility of a General-Purpose, Physician-Initiated Automatic VMAT Treatment Planning System for Curative Treatments..
Nick Chng, Stacy Miller, Robert Olson, Maryam Golshan, Fred Cao, Quinn Matthews . . . 287
71 - A Complete Portfolio for Automated Treatment Planning for Whole Breast Radiotherapy.
Hana Baroudi, Leonard Che Fru, Deborah Schofield, Dominique Roniger, Donald Hancock, Kent Gifford, Tucker Netherton, Joshua Niedzielski, Adam Melancon, Manickam Muruganandham, Simona Shaitelman, Sanjay Shete, Melissa Mitchell, Laurence Court . . . 290
73 - An Open Source Multi-Criteria Optimization Framework for Radiotherapy.
Tobias Becher, Amit Ben Antony Bennan, Remo Cristoforetti, Niklas Wahl . . . 294
75 - Zero Click Treatment Preparation: From CT to Plan.
Geert Wortel, Rita Simões, Johan Trinks, Peter Remeijer, Eugène Damen, Mark Gooding, Tomas Janssen . . 298
77 - Development of Bolus Optimization System for Electron Conformal Therapy.
Sang-Won Kang, Woong Cho, Dongsan Kang, Moo-Jae Han, Jin-Beom Ching . . . 302
79 - A Python Script to Analyse Treatment Planning Throughput.
Ezhlalan Ramalingam . . . 305
81 - Development of Dynamic Collimator Mixed Beam Radiotherapy to Explore the Dosimetric Benefits of Non-Coplanar and Mixed Photon-Electron Beam Setups.
Chengchen Zhu, Gian Guyer, Jenny Bertholet, Silvan Mueller, Hannes A Loebner, Marco F M Stampanoni, Michael K Fix, Peter Manser . . . 308
83 - Online Adaptive Proton Therapy for Esophageal Cancer: Adaptation Rate and Automated Strategies Comparison.
Camille Draguet, Pieter Populaire, Macarena Chocan Vera, Karin Haustermans, John A. Lee, Edmond Sterpin, Ana Barragan-Montero . . . 312

85 - AI-Guided Unattended Plan Generation.
Christian Velten, Michael Bowers, William Martin, Madhur Garg, Julie Shade, Todd McNutt, Wolfgang Tomé 316
87 - Impact of Artificial Intelligence Solution Aiming to Accelerate MRI Acquisition on the Therapeutic Care in Oncology in Radiology and Radiotherapy Departments.
Raphaëlle Lemaire, Charlotte Raboutet, Thomas Leleu, Cyril Jaudet, Loïse Dessoude, Ferdinand Missohou, Yoann Poirier, Pierre-Yves Deslandes, Joëlle Lacroix, Ilyass Moummad, Stéphane Bardet, Juliette Thariat, Dinu Stefan, Aurélien Corroyer-Dulmont 320
89 - The Automatic Body Composition (ABC) Toolkit.
Donal McSweeney, Craig Jones, Michael Brown, Ashwin Sachdeva, Noel Clarke, Marcel Van Herk, Alan McWilliam 324
91 - OSIRIS-RT : a Minimum Data Set for Data Sharing and Interoperability in Radiation Oncology.
Vincent Le Texier, Thomas Baudier, Sébastien Guihard, Thomas Lacornerie, David Pasquier, Charlotte Robert 328
93 - Towards a Fully Automatic Workflow for Investigating the Dynamics of Lung Cancer Cachexia During Radiotherapy Using Cone Beam Computed Tomography.
Lars Daenen, Wouter Van De Worp, Behzad Rezaeifar, Joël De Bruijn, Peiyu Qiu, Justine Webster, Ramon Langen, Frank Verhaegen, Cecile Wolfs 331
95 - Enhancing CBCT Image Quality for Radiotherapy Using Multi-Channel CycleGAN for CBCT-to-CT Synthesis.
Chelsea Sargeant, Robert Chuter, Jane Shortall, Andrew Green, Jacqui Bridge, Phil Fendall Amaro, Rachael Bailey, Alan McWilliam 335
97 - Evaluating Treatment Delivery Performance of SCINTIX Biology-Guided Radiotherapy Under Non-Periodic In-Treatment Motions.
Chenyang Shen, Thomas Banks, Sagar Ghimire, Yang Kyun Park, Andrew Godley, Steve Jiang, Bin Cai . . . 339
99 - Detection and Correction of Rigid Patient Motion in SPECT With Exponential Data Consistency Conditions.
My Hoang Hoa Bui, Antoine Robert, Simon Rit, Etxebeste Ane 343
101 - Preliminary Study on Residual Motion Evaluation During Breath Hold SBRT.
Weihua Mao, Amol Narang 347
103 - Super Resolution Dose Calculation Via Implicit Neural Representation Based Blind Deconvolution.
Siqi Ye, Sheng Liu, Yu Gao, Lei Xing 350
105 - Evaluation of Organ Dose Reduction Effectiveness in Liver 3D/4D Computed Tomography Simulation Through Radiation Shielding: an Anthropomorphic Phantom Study.
Jongin Oh, Minsik Lee, Bonghyo Lee, Doohyun Lee, Min-Young Kang, Semie Hong, Jun-Bong Shin, Jeong-Woo Lee 354
107 - Artificial Intelligence Dose Calculation Engine Based Pencil-Beam for Intensity Modulated Radiation Therapy.
Zirong Li, Yaoying Liu, Xuying Shang, Huashan Sheng, Wei Zhao, Gaolong Zhang, Qichao Zhou, Shouping Xu 358
109 - Development and Evaluation of Novel Parameters for Describing Anatomical Changes and Predicting Radiotherapy Replanning for Head and Neck Cancer Patients.
Odette Rios-Ibacache, James Manalad, Aixa X. Andrade Hernandez, Kayla O'sullivan-Steben, Emily Poon, Luc Galarneau, John Kildea 362
111 - Software-Assisted Optimisation of Linear Accelerator Energy Consumption.
Genotan Reggian, Catherine Stanford-Edwards 366
113 - Deep Reinforcement Learning for Custom Applicator Design in Personalised High Dose Rate Brachytherapy.
Cedric Loiseau, Philippe Berejny, Marlon Silva, Juliette Thariat, Alain Batalla 370
115 - Angle Dependence on Positional Radioisotope Source Using Monte Carlo Simulation in Brachytherapy.
Moo-Jae Han, Sang-Won Kang, Woong Cho, Dongsan Kang, Jin-Beom Jung 374
117 - State-of-the-Art of CT Enhancement Methods for RT Planning: a Review.
Özgür Özer, Juliette Thariat, Alexandre Haut, David Gibon 378
119 - A Tool for Individualised Dose Mapping Uncertainty Quantification in the Reirradiation Setting: Head and Neck Cancer.
Chelmis Muthoni Thiong'o 381
121 - Robustness of Radiomics Features to Reconstruction Kernel Batch Effect: How Bad Is the Situation?
Elfried Salanon, Aditya Apte, Anqi Fu, Usman Mahmood, Belkhatir Zehor, Amita Shukla-Dave, Joseph Deasy 385
123 - Leveraging Hybrid FLT-PET/MR Imaging to Distinguish Tumour Progression From Radionecrosis in Brain Metastasis Patients Post-Radiosurgery.
Michael Maddalena, Adam Farag, Noha Sinno, Paula Alcaide-Leon, Benjamin Schultz, Catherine Coolens . 389

125 - Towards Fast Plan Adaptation for Proton Arc Therapy.
Benjamin Roberfroid, Ana Barragan, John A. Lee, Edmond Sterpin . . . 393
127 - Fast Robust Dose Evaluation for IMPT Using Deep Learning.
Hazem Nomer, Franziska Knuth, Joep Van Genderingen, Dan Nguyen, Margriet Sattler, Uwe Oelfke, Steve Jiang, Linda Rossi, Ben Heijmen, Sebastiaan Breedveld . . . 396
129 - Inverse Robustness to Control Uncertainties in the RBE Parameter in Proton Therapy Planning.
Mara Schubert . . . 400
131 - Machine Learning for Dose Prediction in Carbon Ion Therapy.
Miriam Schwarze, Leo Thomas, Hans Rabus, Ali Kamal, Pichaya Tappayuthpijarn . . . 404
133 - Investigating Associations Between Radiomics Features and Biomarkers in an Orthotopic Non-Small Cell Lung Cancer Mouse Model.
Gabriel Giampa, Ranjan Yadav, Liming Wang, John Kildea, Martin Vallières, Norma Ybarra . . . 406
135 - Towards Patient-Specific Computational Models of Microvasculature to Investigate Radiation Toxicity.
Sophie Materne, Luca Possenti, Alessandro Cicchetti, Francesco Pisani, Alessandra Catalano, Piermario Vitullo, Paolo Zunino, Tiziana Rancati . . . 409
137 - Modeling the Geometric Distortion and Fan Beam Imaging of an On-Rails Computed Tomography Scanner for 2D/3D Registration.
Giovanni Fattori, Riccardo Via, Antony John Lomax, Sairos Safai . . . 413
139 - Bladder Cancer Hypoxia: Combining OE-MRI and Genomics for Biomarker Discovery.
Rekaya Shabbir, Brian A Telfer, Ben Dickie, Muhammad Babur, Kaye Williams, Catharine Ml West, Ananya Choudhury, Tim Ad Smith . . . 417

Rising Stars Competition - Co-chairs: *Magdalena Bazalova, Marc Kessler* **421**

Quantifying Changes in Ventilation for Photon- and Proton-Treated Patients According to Dose and Pneumonitis Development.
Rebecca Lim, Tien Tang, Austin Castelo, Caleb O'connor, Yulun He, Zhongxing Liao, Radhe Mohan, Kristy Brock . . . 421
Introducing a Co-Creation Design Methodology for AI-Based Error Detection in Radiotherapy.
Luca Heising, Frank Verhaegen, Viola Mottarella, Yin-Ho Chong, Mariangela Zamburlini, Bas Nijsten, Ans Swinnen, Michel Öllers, Maria Jacobs, Carol Ou, Cecile Wolfs . . . 425
Dose Driven Gaze Optimization in Ocular Proton Therapy.
Daniel Björkman, Riccardo Via, Tony Lomax, Maria De Prado, Alessia Pica, Guido Baroni, Damien C. Weber, Jan Hrbacek . . . 429
Neural Graphics Primitives-Based Deformable Image Registration for On-the-Fly Motion Extraction.
Xia Li, Fabian Zhang, Muheng Li, Damien Weber, Antony Lomax, Joachim Buhmann, Ye Zhang . . . 433
Robust Optimization for Contouring Uncertainties in Adaptive Radiation Therapy.
Ivar Bengtsson, Andreas Smolders, Anders Forsgren, Antony Lomax, Damien Weber, Albin Fredriksson, Francesca Albertini . . . 437
Short-Term Mortality Prediction in SIII NSCLC: What Do Clinicians Want and Can Models Provide It?
Matt Craddock, Azadeh Abravan, Alan McWilliam, Corinne Faivre-Finn, Gareth Price . . . 441

Breathing Motion - Co-chairs: *Jenny Bertholet, Jan-Jakob Sonke* **445**

Deep Learning Autosegmentation Performance on Mid-Position and Time-Averaged Lung CT Scans.
Joren Brunekreef, Jan-Jakob Sonke . . . 445
Surrogate-Optimized Respiration Motion Model to Estimate Motion for Each Projection of 3D CBCT Scan.
Yuliang Huang, Kris Thielemans, Jamie McClelland . . . 449
Monte Carlo Robustness Analysis of Breath-Hold and Free-Breathing Photon and Proton Treatments of Small Lung Tumors Using Poisson Displacement Fields.
Nils Olovsson, Kenneth Wikström, Anna Flejmer, Anders Ahnesjö, Alexandru Dasu . . . 453
Real-Time CBCT Imaging via a Single Arbitrarily-Angled X-Ray Image Using a Joint Dynamic Reconstruction and Motion Estimation (DREME) Framework.
Hua-Chieh Shao, Tielige Mengke, Tinsu Pan, You Zhang . . . 457

Enhancing Clinical Utility of 0.55T 4DMRI in Radiotherapy Planning: Exploring Compressed Sensing, Denoising, and Auto-Segmentation.
Austen Curcuru, Young Hun Yoon, Shanti Marasini, H Michael Gach, Jianing Pang, Xinzhou Li, Xin Miao, Jin Sung Kim, Taeho Kim 461

Internal Radiation Therapy - Co-chairs: *Philippe Després, Ane Etxebeste* **465**

Reconstruction of Time-Activity Curves in MRT Therapy Using Bayesian Unfolding.
Francesca Nicolanti, Carlo Mancini Terracciano, Riccardo Faccini, Elena Solfaroli Camillocci, Silvio Morganti, Riccardo Mirabelli, Lorenzo Campana, Francesco Collimati 465

Rapid Auto-Planning for Cervical Brachytherapy: Direct Dwell Time vs. Dose Prediction With Deep Learning Networks.
Lance Moore, Karoline Kallis, Kelly Kisling, Nuno Vasconcelos, Sandra Meyers 469

Near-to-Target Organs-at-Risk Segmentation in Cervical High Dose-Rate Brachytherapy via Deep Learning.
Ruiyan Ni, Elizabeth Chuk, Kathy Han, Jennifer Croke, Anthony Fyles, Jelena Lukovic, Michael Milosevic, Benjamin Haibe-Kains, Alexandra Rink 473

Automatic Target Propagation Strategies for MRI-Guided Cervical Brachytherapy.
Rita Simões, Eva E. Schaake, Eva C. Rijkmans, Marlies E. Nowee, Tomas M. Janssen 477

Geant4 Simulation of DNA Damage Surrounding a DaRT Seed.
Laura Ballisat, Chiara De Sio, Lana Beck, Susanna Guatelli, Dousatsu Sakata, Yuyao Shi, Jinyan Duan, Jaap Velthuis, Anatoly Rosenfeld 481

Poster Session, Even IDs **485**

2 - A Pipeline for Generating Longitudinal Head and Neck Cancer Radiotherapy Datasets for Machine Learning.
Kayla O'sullivan-Steben, Odette Rios-Ibacache, Luc Galarneau, John Kildea 485

4 - Demonstration of Current Best Practice for Prognostic Model Validation Using the Manchester Score.
Charlie Cunniffe, Gareth Price, Matthew Sperrin, Fiona Blackhall 489

6 - A Non-Invansive and Self-Managed Microscope for Predicting Radiation-Induced Acute Toxicities in Cancer Patients.
Alessandra Catalano, Sophie Veronique Materne, Fabio Badenchini, Eliana La Rocca, Eliana Gioscio, Francesco Pisani, Luca Possenti, Barbara Avuzzi, Barbara Noris Chioda, Carlotta Giandini, Riccardo Roy Colciago, Maria Carmen De Sanctis, Tommaso Giandini, Alessandro Cicchetti, Tiziana Rancati 493

8 - Data Preparation Pipeline for Artificial Intelligence in Radiation Oncology: Lessons From a Large Prospective Imaging Trial.
Hannah Thomas, Amal Joseph Varghese, Manu Mathew, Devadhas Devakumar, Aparna Irodi, Joel T, Timothy Peace, Henry Finlay Godson, Rajesh Isiah, Simon Pavamani, Balu Krishna Sasidharan 497

10 - Transfer Learning for the Detection of Lung Lesions in 18F FDG PET/CT Images Using GraphNet.
Maryam Fallahpoor, Biswajeet Pradhan 500

12 - Optimising Physics-Informed CRNNs for Longitudinal Volumetric Medical Image Data.
Denis Page, Eliana M. Vasquez Osorio, Robert Chuter, Andrew Green, Alan McWilliam 502

14 - i-SART: A Prototype AI-Based Tool to Enhance Patient Safety in Radiotherapy.
Anastasia Sarchosoglou, Natalia Silvis-Cividjian, Yijing Zhou, Periklis Papavasileiou, Athanasios Bakas, Evangelos Pappas 506

16 - Deep Learning-Based Tumor Shape Variation Prediction of Lung Cancer Patient's Pre-Radiation Treatment Response.
Sangwoon Jeong, Kyungmi Yang, Youngyih Han, Hee Chul Park, Sang Hee Ahn 510

18 - Deep Learning for in Vivo EPID Dosimetry Classification: Relating Gamma Analysis and DVH Metrics for Lung Cancer Patients.
Femke Vaassen, Sebastiaan Nijsten, Richard Canters, Frank Verhaegen, Cecile Wolfs 514

20 - Adaptive Radiotherapy Dose Prediction on Head and Neck Cancer Patients With a Multi-Headed U-Net Deep Learning Architecture.
Hui-Ju Wang, Austen Maniscalco, David Sher, Mu-Han Lin, Steve Jiang, Dan Nguyen 518

22 - Prediction of Stress in Radiation Therapy Patients Using Artificial Intelligence and Biological Signals.
Sangwoon Jeong, Hongryull Pyo, Won Park, Youngyih Han 522

24 - Habitat Imaging Based on Voxel-Wise GLCM Joint Energy Clustering: Prediction of Disease-Free Survival in Localized Soft-Tissue Sarcoma..
Benoît Allignet, Benjamin Leporq, Amine Bouhamama, Frank Pilleul, Alexandra Meurgey, Vaz Gualter, Marie Pierre Sunyach, Waisse Waissi, Olivier Beuf 526
26 - A Deep Learning-Driven Framework for Automatic Target Volume Localization and Delineation in Spinal Metastases Stereotactic Body Radiotherapy.
Zi Yang, Mahdieh Kazemimoghadam, Gregory Szalkowski, Cynthia Chuang, Lei Wang, Lianli Liu, Scott Soltys, Erqi Pollom, Elham Rahimy, Hao Jiang, David Park, Amit Persad, Yusuke Hori, Jie Fu, Ignacio Romero, Laszlo Zalavari, Mingli Chen, Weiguo Lu, Xuejun Gu 530
28 - Safeguarding Clinical AI-Based Auto-Segmentation: Open-Source Models as Quality Assurance Against Catastrophic Failures.
Daniel Alexander, Rafe McBeth 534
30 - Automatic Target Segmentation and Beam Setup for Electron Skin Radiotherapy.
Rik Westendorp, Miranda Beumer, Tristan Van Heijst, Kelvin Ng Wei Siang 538
32 - Comparative Analysis of Deep Learning Models for Flap Segmentation in Head and Neck Computed Tomography Images: A Generalization Study.
Abir Fathallah, Alice Blache, Zacharia Mesbah, Romain Modzelewski, Juliette Thariat, Mathieu Hatt . . . 542
34 - An Automated Framework of Accurate Dominant Intraprostatic Lesion Delineation Based on Multimodal Images for MRI-Guided Adaptive Prostate SBRT.
Zirong Li, Xiao Liu, Qichao Zhou, X. Sharon Qi 546
36 - Medical Image Segmentation Assisted With Clinical Inputs via Language Encoder in A Deep Learning Framework.
Hengrui Zhao, Biling Wang, Deepkumar Mistry, Jing Wang, Michael Dohopolski, Daniel Yang, Weiguo Lu, Steve Jiang, Dan Nguyen 549
38 - Influence of Electron Transport Algorithm on Monte Carlo Based Treatment Planning.
Grisel Mora Paula, A Eldib, J Li, Charles Ma 553
40 - GMC [gímɪk] Then and Now: (r)evolution of a One-Variable Proton Monte Carlo Dose Algorithm..
Nicolas Depauw, Michael Williams, Benjamin Clasie, Kyung-Wook Jee, Thomas Madden, Hanne Kooy . . . 557
42 - Assessing Neutron-Induced Carcinogenesis Risk in High-Energy Radiation Therapy: a Monte Carlo and Track Structure Study Using Realistic Neutron Spectra and DNA Damage Clusters.
Sachin Dev, Felix Mathew, John Kildea 560
44 - Recent Progress in GMicroMC Package for Microscopic Monte Carlo Simulations.
Youfang Lai, Yuting Peng, Meiying Xing, Yujie Chi, Xun Jia 563
46 - Evaluation of a New Radiosurgery System (infinityARC) Using Geant4.
Patrick Johan Ngontie Ngounou, Philippe Despres, Yannick Lemaréchal 567
48 - Oxygen Effect on Tumor Response to Irradiation: Insights From GATE Simulation Studies.
Zakaria Ait Elcadi, Othmane Bouhali 571
50 - Monte-Carlo Modeling of Intratumoral Radionuclide Distribution in TAT: Comparison to Experiment.
Victor Levrague, Mario Enrique Alcocer-Avila, Etienne Testa, Michaël Beuve, Rachel Delorme 573
52 - DINo: Deep Learning Intelligence for Nuclear Reactions.
Lévana Gesson, Greg Henning, Uli Weber, Marie Vanstalle 577
54 - Development of the Low Dose Radiotherapy Machine for the Osteoarthritis: Optimization of Design Based on the Monte Carlo Simulation.
Hyojun Park, Jung-In Kim, Sang Hoon Heo, Song Yi Heo, Minjae Choi, Ju Yeol Shin, Chang Heon Choi . . 580
56 - Prompt-Gamma Timing With a Track-Length Estimator in Proton Therapy.
Jean Létang, Oreste Allegrini, Etienne Testa 584
58 - Identification of Suitable Probing Locations for a Proton Radiography-Based Quality Control of CBCT-Based Synthetic-CTs.
Måns Lundberg, Arturs Meijers, Antony Lomax, Damien Weber, Kevin Souris, Antje Knopf 588
60 - External Validation of an AI-Based Tool for the Early Detection of Radiotherapy Treatment Planning Errors-An Updated Multicentric Study.
Petros Kalendralis, Samuel Luk, Alan Kalet, Andre Dekker, Inigo Bermejo, Tomasz Piotrowski, Adam Ryczkowski, Joanna Kazmierska 592
62 - Development of an Automated Radiotherapy Daily QA System.
Marcus Tyyger, Matthew Carlisle, Jack P. C. Baldwin, Richard Homer, Phillip Rixham, Steve Weston 596

64 - Isocenter Deviation Assessment and Model for Halcyon Radiotherapy Systems.
Linda Lankinen, Antti Kulmala, Jouko Lehtomäki, Ari Harju 599
66 - Patient-Specific Quality Assurance in Radiotherapy: Prediction With Deep Hybrid Learning for VMAT Plans on Rapid Arc and IMRT Plans on Halcyon Machines.
Noémie Moreau, Christine Boutry, Laurine Bonnor, Cyril Jaudet, Nadia Falzone, Alain Batalla, Laetitia Lechippey, Cindy Bertaut, Aurélien Corroyer-Dulmont 603
68 - Analysis of a Decade's Worth of Oncology Information System Data From a Large US Healthcare System.
Christian Velten, Ping Yan, Madhur Garg, Wolfgang Tomé 607
70 - Fast Many-Scenario Robust IMPT Planning Through Constrained Dose-Mimicking.
Franziska Knuth, Michelle Oud, Wens Kong, Hazem Nomer, Joep Van Genderingen, Linda Rossi, Steven Habraken, Ben Heijmen, Sebastiaan Breedveld 611
72 - DeepTuning: Tune Radiotherapy Plan Dose Distribution in Real Time by Deep Learning.
Lin Ma, Huan Liu, Mingli Chen, Weiguo Lu 614
74 - Automating Physics Plan Checks With Monaco Scripting.
Hill Paul 618
76 - Optimization in Radiotherapy Integrating a Multidisciplinary Framework With GEANT4.
Elisabetta Sergi, Mara Severgnini, Enrico Rigoni, Francesco Longo 621
78 - A Personalized Automatic Radiotherapy Plan Scoring Mechanism Based on Multi-Organ Individually-Objective Optimization.
Zirong Li, Zhenjiang Li, X. Sharon Qi, Kuo Li, Huashan Sheng, Zhigang Yuan, Yong Yin, Qichao Zhou . . . 623
80 - Dynamic Collimator Rotation Treatment Plan Optimization for O-Ring Treatment Systems.
Fix Michael, Werner Volken, Daniel Frei, Silvan Mueller, Gian Guyer, Daniel Schmidhalter, Peter Manser . 626
82 - Dosimetry for Head and Neck Radiotherapy Using Personalized Quantitative MRI.
Laura Sayaque, Benjamin Leporq, Charlène Bouyer, Frank Pilleul, Vincent Gregoire, Olivier Beuf 630
84 - IOeRT Conventional and FLASH TPS Implementation Exploiting Fast GPU Monte Carlo: the Case of Breast Cancer.
Gaia Franciosini, Angelica De Gregorio, Massimo Di Francesco, Fabio Di Martino, Giuseppe Felici, Jake Harold Pensavalle, Michela Marafini, Annalisa Muscato, Vincenzo Patera, Philip Poortmans, Adalberto Sciubba, Angelo Schiavi, Marco Toppi, Giacomo Traini, Antonio Trigilio, Alessio Sarti 634
86 - Exploring VHEE Treatment Efficacy Through In-Silico Investigations in Conventional and FLASH Irradiation Regimes.
Annalisa Muscato, Angelica De Gregorio, Daniele Carlotti, Michele Fiore, Gaia Franciosini, Teresa Insero, Valerio Marè, Sara Ramella, Alessio Sarti, Angelo Schiavi, Marco Schwarz, Vincenzo Patera 638
88 - Reconstruction Toolkit (RTK) V2, an Insight Toolkit (ITK) Module for Tomographic Reconstruction.
Simon Rit, Sébastien Brousmiche, Julien Finet, Gregory C. Sharp, Philipp Steininger 641
90 - Validation of an Open Source Contouring Tool.
Haley Clark 645
92 - Harmonizing Multicentric T1-Weighted Magnetic Resonance Imaging of Children With Ependymomas: Can Deep Learning Based Image Synthesis Help?
Soumia Bengoufa, Fatima Tensaouti, Pedro Juan Soto Vega, Stéphanie Bolle, Alexandre Escande, Xavier Muracciole, Claire Alapetite, Line Claude, Julie Leseur, Aymeri Huchet, Jérôme Doyen, Georges Noel, Stéphane Supiot, Valerie Bernier-Chastagner, Julien Welmant, Pierre Leblond, Dimitris Visvikis, Vincent Jaouen, Anne Laprie, Mathieu Hatt 649
94 - Performance Evaluation of a Pix2Pix Synthetic Xray Image Generator for Preclinical Applications.
Vasileios Eleftheriadis 653
96 - Fast and Automated Dose Verification Procedure for Reinforced Online Adaptive MRI-Guided RT.
Teo Stanescu 655
98 - Optimizing the Contrast to Noise Ratio for Tumor Tracking in the Lung Using Dual Energy Imaging Strategies for Soft Tissue Enhancement.
Nawal Alqethami, Wentao Xie, Prasannakumar Palaniappan, Felix Ginzinger, Philipp Steininger, Marco Riboldi 657
100 - Performance Evaluation of Virtual Laser-Based Surface-Guided Radiation Therapy Setup System Using Humanoid Head Phantom.
Jeong-Woo Lee, Byoung-Moon Park, Yong-Ki Bae, Min-Young Kang, Semie Hong 661

102 - Preliminary Evaluation of a Software Application for Target Tracking Without Fiducials During Cyberknife Treatment of Pancreas.
James Bedford, Simeon Nill, Uwe Oelfke . . . 664

104 - The Impact of Training Dataset Size and Model Size on the Accuracy of Deep Learning Dose Prediction.
Joep Van Genderingen, Dan Nguyen, Franziska Knuth, Hazem Nomer, Luca Incrocci, Abdul Sharfo, Uwe Oelfke, Steve Jiang, Linda Rossi, Ben Heijmen, Sebastiaan Breedveld . . . 668

106 - Deep Learning for Instantaneous Estimation of the Out-of-Field Dose in External Beam Photon Radiotherapy. A Proof of Concept.
Nathan Benzazon, Mohammed El Aichi, Alexandre Carré, François De Kermenguy, Stéphane Niyoteka, Pauline Maury, Julie Colnot, Meissane M'hamdi, Cristina Veres, Rodrigue Allodji, Florent De Vathaire, David Sarrut, Neige Journy, Claire Alapetite, Vincent Gregoire, Eric Deutsch, Ibrahima Diallo, Charlotte Robert . . . 672

108 - Accurate Selection of Calculation Grid Size for Radiation Plans With Small Brain Tumors in Intracranial Stereotactic Radiotherapy.
Tagreed Alalawi, Mazen Al-Gaoud, Elham Rashaidi, Nesreen Shorbaji . . . 676

110 - To Calibrate or Not to Calibrate: a Decision Support Approach Using Microcontroller-Based Tiny Machine Learning.
Nicolas Boussion, Lou Bosc, Gaëlle Goasduff, Yoann Jegat, Anne-Sophie Lucia, Emmanuelle Martin, Laetitia Lann, Dimitris Visvikis . . . 679

112 - Experiment of Proof-of-Principle on C-Arm CT/SPECT System for In-Vivo Dose Verification in Online Image-Guided Adaptive Brachytherapy.
Saerom Sung, Kim Hyemi, Min Chul Hee, You Sei Hwan, Choi Hyun Joon . . . 683

114 - Multiphysics Modeling and Simulations for Y-90 Liver Radioembolization.
Emilie Roncali, Carlos Ruvalcaba, Gustavo Costa . . . 686

116 - Partial-Ring PET Image Correction Using Implicit Neural Representation Learning.
Shubhangi Makkar, Siqi Ye, Marina Béguin, Günther Dissertori, Jan Hrbacek, Antony Lomax, Keegan McNamara, Christian Ritzer, Damien C. Weber, Lei Xing, Carla Winterhalter . . . 689

118 - Artificial-Intelligence Enhanced Image Reconstruction for Proton Range Verification.
Lena Setterdahl, Ilker Meric, William Lionheart, Sean Holman . . . 693

120 - Head and Neck Deformable Registration Using Self-Attention With Positional Encoding.
Elizabeth McKenzie, Taman Upadhaya, Hengjie Liu, Ke Sheng . . . 697

122 - The Prognostic Value of Foundational Radiomics and Clinical Features in Locally Advanced Non-Small Cell Lung Cancer in NRG Oncology Trial RTOG 0617.
Upadhaya Taman, Elizabeth M. McKenzie, Indrin J. Chetty, Katelyn M. Atkins . . . 701

124 - Exploring How Prostate Cancer Treatment Choices Affect Radiomic Features: A Look at Healthy Tissue vs. Dominant Lesion in MR-Guided Radiotherapy.
Aaron G. Rankin, Swinton Martin, Jim Zhong, Priya Maitre, Angela Davey, Marianne Aznar, Ananya Choudhury, Alan McWilliam . . . 705

126 - On Computational Challenges in Proton Therapy Treatment Planning.
Mohammadhossein Mohammadisiahroudi, Zou Wei, Lei Dong, Yuriy Zinchenko, Tamás Terlaky . . . 709

128 - A Neural Network-Based Dose Calculation Engine for Proton Spots Dose: a Clinical Implement Study.
Xuying Shang, Yaoying Liu, Zishen Wang, Nan Li, Chunfeng Fang, Yue Zou, Wei Zhao, Xiaoyun Le, Gaolong Zhang, Shouping Xu . . . 713

130 - JulianA.jl - A Flexible Open Source Julia Package for Pencil Beam Scanning Proton Therapy.
Renato Bellotti, Tony Lomax, Andreas Adelmann, Jan Hrbacek, Damien C. Weber . . . 716

132 - Monte Carlo Verification of Proton Treatment Planning.
Klara Badraoui Cuprova . . . 720

134 - RadiSeq: A Whole-Genome DNA Sequencing Simulation Pipeline to Investigate Radiation-Induced Carcinogenesis.
Felix Mathew, John Kildea . . . 723

136 - Detectability of Iodine Contrast Agent With Ultra-Low Concentration by Photon-Counting CT: A Simulation Study.
Xiaoyu Hu, Yuncheng Zhong, Kai Yang, Xun Jia . . . 727

138 - Unraveling Interstitial Fluid Pressure: Analyzing Drug Transport Through Cross Voxel Exchange Model and Dynamic Contrast Enhanced MRI.
Yeyoung Kim, Noha Sinno, Michael Milosevic, Catherine Coolens . . . 731

140 - Revealing Head and Neck Tumors Spatial Heterogeneity Through Quantitative MRI Protocol: Initial Findings..
Walid Dandachly, Benjamin Leporq, Laura Sayaque, Chloé Miran, Benoit Allignet, Frank Pilleul, Olivier Hamelin, Charlène Bouyer, Vincent Gregoire, Olivier Beuf . . . 735

Image Segmentation II - Co-chairs: *Charlotte Brouwer, Catherine Coolens* **739**

Segment Anything Model for Head and Neck Tumor Segmentation With CT, PET and MRI Multi-Modality Images.
Jintao Ren, Mathis Rasmussen, Jasper Nijkamp, Jesper Eriksen, Stine Korreman . . . 739

Test-Time Adaptation for Medical Image Segmentation Based on a Joint Framework With an Adaptive Wavelet-VNet and a Refine Model (WaVNet-Refine).
Xiaoxue Qian, You Zhang . . . 743

MAGIC: Multitask Adversarial Generator of Images and Contours From CBCT for Adaptive Radiotherapy.
Adam Szmul, Kamalram Thippu Jayaprakash, Rajesh Jena, Andrew Hoole, Catarina Veiga, Yipeng Hu, Jamie R. McClelland . . . 747

Improving Clinical Target Volume Segmentation Throughout a Gastric Cancer Radiotherapy Clinical Trial.
Phillip Chlap, Mark Lee, Trevor Leong, Matthew Field, Jason Dowling, Hang Min, Julie Chu, Jennifer Tan, Phillip K Tran, Tomas Kron, Annette Haworth, Martin A Ebert, Shalini K Vinod, Lois Holloway . . . 751

OAR-Weighted Dice Score: A Spatially Aware, Radiosensitivity Aware Metric for Target Structure Contour Quality Assessment.
Lucas McCullum, Kareem Wahid, Barbara Marquez, Clifton Fuller . . . 755

Towards Automatic Identification of Missing Tissues Using a Geometric-Learning Correspondence Model.
Eliana M Vasquez Osorio, Edward Henderson . . . 759

Fully Automatic Flap Segmentation Through Deep Learning Methods to Refine Head and Neck Cancer Postoperative Radiotherapy, a GORTEC Study.
Juliette Thariat, Zacharia Mesbah, Youssef Chahir, Abir Fathallah, Vianney Bastit, Ingrid Breuskin, Arnaud Beddok, Alice Blache, Jean Bourhis, Mathieu Hatt, Romain Modzelewski . . . 763

Radiobiology and Microscopic Disease - Co-chairs: *Sarah Gulliford, Lydia Maigne* **767**

A Rat Model of Lymphocyte Recirculation and the Effect of Irradiation.
Chris Beekman, Natalia Carrasco-Rojas, Julia Withrow, Wesley E. Bolch, Harald Paganetti . . . 767

An AI-Based Approach for Modeling the Synergy Between Radiotherapy and Immunotherapy.
Hao Peng, Casey Moore, Yuanyuan Zhang, Debabrata Saha, Steve Jiang, Robert Timmerman . . . 771

Integration of White Matter Tracts in Stereotactic Brain Radiotherapy Treatment Planning With a Fully Automated Atlas-Based Pipeline.
Sophie Bockel, Mohammed El Aichi, Roger Sun, Nathan Benzazon, Alexandre Carré, François De Kermenguy, Eric Deutsch, Charlotte Robert . . . 775

Comparison of Geant4-DNA and RITRACKS/RITCARD: Microdosimetry, Nanodosimetry and DNA Break Predictions.
Victor Levrague, Rachel Delorme, Mathieu Roccia, Ngoc Hoang Tran, Sébastien Incerti, Michaël Beuve, Etienne Testa, Ianik Plante, Shirin Rahmanian, Floriane Poignant . . . 779

Whole-Body Blood Circulation Modeling and Dose to Circulating Blood in Intensity Modulated Radiation Therapy.
Bingqi Guo, Ping Xia . . . 783

Using Fluid Flow Simulations to Model Microscopic Tumor Invasion in Muscle Tissue for Radiotherapy Treatment.
Gregory Buti, Ali Ajdari, Yen-Lin Chen, Christopher Bridge, Gregory Sharp, Thomas Bortfeld . . . 787

Characterization of Acute Radiation-Induced Vascular Changes in Brain Tumors Using Wavelet-Based Analysis of Dynamic-Contrast-Enhanced MRI Information.
Hassan Bagher-Ebadian, Stephen L. Brown, Tavarekere N. Nagaraja, Prabhu Acharya, James R. Ewing, Indrin J. Chetty, Benjamin Movsas, Kundan Thind . . . 791

Open Source Software - Co-chairs: *Hermann Fuchs, Julien Jomier* **795**

OpenTPS: an Open-Source Treatment Planning System to Foster Research and Innovation.
Ana María Barragan Montero, Damien Dasnoy, Sophie Wuyckens, Wei Zhao, Eliot Peeters, Romain Schyns, Margerie Huet Dastarac, Estelle Löyen, Gauthier Rotsart De Hertaing, Guillaume Janssens, Valentin Hamaide, Sylvain Deffet, Kevin Souris, Edmond Sterpin, Benoit Macq, John Lee 795
GATE 10: A New Versatile Python-Driven Geant4 Application for Medical Physics.
Nils Krah, Lydia Maigne, Martina Favoretto, Andreas Resch, Carlotta Triglia, Guneet Mummaneni, Emilie Roncali, Maxime Jacquet, Thomas Baudier, Alexis Pereda, Hermann Fuchs, Aurélien Coussat, David Sarrut 799
PortPy: An Open-Source Initiative for Radiotherapy Treatment Planning Optimization Including Benchmark Data and Algorithms and Integration With Commercial TPS.
Gourav Jhanwar, Mojtaba Tefagh, Qijie Huang, Vicki Trier Taasti, Seppo Tuomaala, Saad Nadeem, Masoud Zarepisheh 803
Pylinac: An Open-Source Quality Assurance Image Analysis Engine.
James Kerns 807

Dose Calculation - Co-chairs: *James Bedford, Irène Buvat* **811**
A Long Short-Term Memory Network-Based Proton Dose Calculation Method in a Magnetic Field.
Fan Xiao, Domagoj Radonic, Niklas Wahl, Luke Voss, Ahmad Neishabouri, Nikolaos Delopoulos, Stefanie Corradini, Claus Belka, George Dedes, Christopher Kurz, Guillaume Landry 811
Deep Learning-Based Dose Prediction for INTRABEAM.
Mojahed Abushawish, Arthur V Galapon, Joaquin L. Herraiz, José M Udías, Paula Ibanez 815
GPU-Based Monte Carlo Method for Mixed Beam Radiation Therapy Electron Beamlet Calculation..
Veng Jean Heng, Marc-André Renaud, Jan Seuntjens 819
Revisiting Over Three Years of Clinical Application of Monte Carlo Based Patient Quality Assurance at the Maastro Proton Therapy Center.
Ilaria Rinaldi, Jan Gajewski, Angelo Schiavi, Nils Krah, Antoni Rucinski, Vincenzo Patera, Sebastiaan Nijsten 823
A Stochastic Model to Estimate Dose to Circulating Lymphocytes During Brain Radiotherapy.
François De Kermenguy, Nathan Benzazon, Pauline Maury, Rémi Vauclin, Meissane M'hamdi, Vjona Cifliku, Elaine Limkin, Ibrahima Diallo, Daphné Morel, Candice Milewski, Céline Clémenson, Michele Mondini, Eric Deutsch, Charlotte Robert 826
Deep Learning Dose Engine for Electron FLASH Radiotherapy.
Lorenzo Arsini, Barbara Caccia, Andrea Ciardiello, Angelica De Gregorio, Gaia Franciosini, Stefano Giagu, Jack Humphreys, Annalisa Muscato, Angelo Schiavi, Christopher White, Markus Hagenbuchner, Susanna Guatelli, Carlo Mancini Terracciano 830
Skin Dose in Breast Radiation Therapy: Monte Carlo Calculations From Deformed Vector Fields (DVF)-Driven CT Images.
Nicolas Arbor, Laurent Bartolucci, Botao Dai, Halima Elazhar, Pierre Galmiche, Delphine Jarnet, Michel De Mathelin, Philippe Meyer, Hyewon Seo 834

Treatment Planning - Co-chairs: *Marie-Claude Biston, Peter Remeijer* **837**
Peeking Under the Hood: Addressing Biomarker-Driven Uncertainty in Machine Learning-Based Radiotherapy Personalization.
Ali Ajdari, Zhongxing Liao, Thomas Bortfeld 837
pyanno4rt: a Toolkit for Machine Learning (N)TCP Model-Based Inverse Radiotherapy Treatment Plan Optimization.
Tim Ortkamp, Patrick Salome, Oliver Jäkel, Martin Frank, Niklas Wahl 841
Intelligent Automatic Treatment Planning for Radiation Therapy.
Yin Gao, Yang Kyun Park, Daniel Yang, Xun Jia 845
An Exploration of the Complexities of CTV Mapping.
Marcel Van Herk, Matt Craddock, Mairead Daly, Eliana M Vasquez Osorio 849
Biologically-Guided Multi-Criteria Optimization and Exploration to Generate Clinical Plans for Functionality-Sparing Radiation Therapy Trials.
Dimitre Hristov, Matthew Kim, Jen-Yeu Kim, Elizabeth Kidd 852
Deep Learning Guided Radiotherapy Modality Selection via Multi-Task Dose Prediction.
Austen Maniscalco, Ezek Mathew, David Parsons, Justin Visak, Mona Arbab, Prasanna Alluri, Xingzhe Li, Narine Wandrey, Mu-Han Lin, Asal Rahimi, Steve Jiang, Dan Nguyen 856

Producing Deliverable Treatment Plans With Deep Learning: A Fully Automated Pipeline Outside the Treatment Planning System.
Gerd Heilemann, Tufve Nyholm, Gregor Goldner, Joachim Widder, Dietmar Georg, Peter Kuess 860

Language Models and Ontologies - Co-chairs: *Lei Xing, John Kildea* **864**

Enhancing Domain-Specific Knowledge of LLMs Through Retrieval Augmented Multi-Process Prompting (RAMP).
Jace Grandinetti, Rafe McBeth 864

An Open Source PyMedPhys Harness for Interfacing MOSAIQ®With Claude®.
Stuart Swerdloff, Simon Biggs 868

Iterative Prompt Refinement for Radiation Oncology Symptom Extraction Using Teacher-Student Large Language Models.
Reza Khanmohammadi, Ahmed Ghanem, Kyle Verdacchia, Ryan Hall, Mohamed Elshaikh, Benjamin Movsas, Hassan Bagher-Ebadian, Indrin Chetty, Mohammad Ghassemi, Kundan Thind 872

A BFO-Compliant Ontology for Radiation Therapy (RTO).
Mark Phillips, Andre Dekker, Jonathan Bona 876

Patient and Public Involvement and Engagement on Natural Language Processing for Real World Cancer Medical Notes.
Wuraola Oyewusi, Eliana M. Vasquez Osorio, Goran Nenadic, Issy Macgregor, Gareth Price 880

Decision Support Systems - Co-chairs: *Thomas Kron, Cecile Wolfs* **884**

An AI-Driven Radiation Therapy Workflow Management Platform.
Anyi Li, Doori Rose, Grace Chow, Laura Cervino, Joseph Deasy, Sharif Elguindi, Gregory Niyazov, Jean Moran, Darian Pazgan-Lorenzo, Ellen Pinto, Lakshmi Santanam, Niral Shah, Pengpeng Zhang 884

A Mixture of Hidden Markov Models to Predict the Lymphatic Spread in Head and Neck Cancer Depending on Primary Tumor Location.
Roman Ludwig, Julian Brönnimann, Yoel Pérez Haas, Esmée Lauren Looman, Panagiotis Balermpas, Sergi Benavente, Adrian Schubert, Dorothea Barbatei, Laurence Bauwens, Jean-Marc Hoffmann, Olgun Elicin, Matthias Dettmer, Bertrand Pouymayou, Roland Giger, Vincent Gregoire, Jan Unkelbach 887

Data Collection Platform for Radiation Oncology Clinical Use and Medical Research in a Multivendor Setting.
Diego García, David Brito 891

PARROT: A Web Platform for AI-Driven Radiation Therapy Planning.
Margerie Huet Dastarac, Wei Zhao, Sylvain Deffet, Eléonore Longton, Sophie Cvilic, Edmond Sterpin, John Lee, Ana Barragan-Montero 895

Dose Evaluation on Daily CBCTs for Breast Cancer Patients.
Peter Stijnman, Maureen Groot Koerkamp, Anette Houweling, Cornel Zachiu, Bas Raaymakers 899

Software Development and Interoperability - Co-chairs: *Natalia Silvis-Cividjian, Marcel van Herk* **903**

Harnessing Radiotherapy Data for Quality Care and Outcome Research.
Rishabh Kapoor, William Sleeman, Jatinder Palta 903

TrajectoryLog.NET Applications of an Integrated Trajectory Log File Analysis API.
Matthew Schmidt, Frank Marshall, Sarah Cradick, Justin Mikell, Phillip Wall, Nels Knutson, Thomas Mazur, Yao Hao 907

PARADIM: a Platform to Support Research at the Interface of AI and Medical Imaging.
Yannick Lemaréchal, Gabriel Couture, François Pelletier, Pierre-Luc Asselin, Ronan Lefol, Samuel Ouellet, Leonardo Di Schiavi Trotta, Philippe Despres 911

pyCERR – A Python-Based Computational Environment for Radiological Research.
Aditya P. Apte, Aditi Iyer, Eve Locastro, Sharif Elguindi, Jue Jiang, Jung Hun Oh, Harini Veeraraghavan, Amita Shukla-Dave, Joseph O. Deasy 915

A System of Governance for Use of Scripts in a Commercial Treatment Planning System.
Kenton Thompson, Keith Offer, Glen Osborne, Catherine Lawford, Alan Turner, Sandro Porceddu, Tomas Kron 919

QuAAC: An Interoperable Routine Quality Assurance Data Format.
James Kerns, Randle Taylor 922

Monte-Carlo Simulation - Co-chairs: *Indrin Chetty, David Sarrut* **926**

MCMimic: a Machine Learning Approach for Fast Prediction of Monte Carlo Proton Dose and LET.
Samuel Ingram, Karen Kirkby, Ranald Mackay, Matthew Clarke . . . 926

Evaluation of the GATE-RTion/Geant4 Simulation Speed and Dosimetric Commissioning for Independent Dose Calculation in Carbon Ion Pencil Beam Scanning Radiotherapy.
Yihan Jia, Lisa Hartl, Martina Favaretto, Markus Stock, Loic Grevillot, Andreas F. Resch . . . 930

New Global Reconstruction for Charged Particles With the FOOT Experiment.
Christian Finck . . . 934

Deep-Learning Powered Denoising of Monte Carlo Dose Distributions.
Hannes A. Loebner, Raphael Joost, Jenny Bertholet, Stavroula Mougiakakou, Michael K Fix, Peter Manser 938

Development of a Range Probe to Improve Dose Delivery Accuracy in Proton Beam Therapy.
Josephine Jones, Michael Taylor, Adam Aitkenhead, Samuel Manger . . . 942

Implementation of the Mayo Clinic Florida Microdosimetric Kinetic Model and AMDM Into GATE/Geant4: Investigation of the Impact of Nozzle Configurations.
Hermann Fuchs, Alessio Parisi, Keith Furutani, Dietmar Georg, Chris Beltran . . . 946

Image Reconstruction and Generative Models - Co-chairs: *Guillaume Landry, Simon Rit* **950**

Diffusion Schrödinger Bridge Models for High-Quality MR-to-CT Synthesis for Head and Neck Proton Treatment Planning.
Muheng Li, Xia Li, Sairos Safai, Damien Weber, Antony Lomax, Ye Zhang . . . 950

Self Super-Resolution of Anisotropic Volumes in Prostate MRI With Normalized Edge Priors.
Vincent Jaouen, Pierre-Henri Conze, Julien Bert, Dimitris Visvikis . . . 954

Patient-Specific Simultaneous CT and MR Reconstruction From Sparse Projection Measurements Using Implicit Neural Representation.
Oscar Pastor-Serrano, Liang Qiu, Siqi Ye, Lei Xing . . . 958

FDDM: Unsupervised MR-to-CT Translation With a Frequency-Decoupled Diffusion Model.
Yunxiang Li, Hua-Chieh Shao, Xiaoxue Qian, You Zhang . . . 962

Enhancing CBCT Image Quality Using a Novel SCT-Based Shading Correction.
Francesco Maria Russo, Lukas Lamminger, Gabriele Belotti, Katia Parodi, Heinz Deutschmann, Philipp Steininger . . . 966

Comparison of Saddle and Helical Trajectories of a Mobile Cone Beam Computed Tomography Scanner With Online Calibration Correction.
Chengtao Wei, Simon Rit, Matthieu Laurendeau, Philipp Steininger, Felix Ginzinger, Marco Riboldi, Christopher Kurz, Guillaume Landry . . . 970

Prediction Models II - Co-chairs: *Vincent Grégoire, Stephen Tozer-Loft* **974**

Deep Learning-Based Normal Tissue Complication Models for Head and Neck Cancer Patients.
Lisanne V. Van Dijk, Suzanne P.m. Vette, Hendrike Neh, Luuk Van Der Hoek, Hung Chu, Johannes A Langendijk, Peter M.a. Ooijen, Nanna M. Sijtsema . . . 974

Causal Targeted Machine Learning for Real-World Evidence Generation in Oncology: Assessing Causal Effects of Radiation Toxicity on Overall Survival in Stage III Non-Small Cell Lung Cancer.
Rachael Phillips, Tomi Nano, Jessica Scholey, Efstathios Gennatas, Nasir Mohammed, Jing Zeng, Rupesh Kotesha, Pranshu Mohindra, Lane Rosen, John Chang, Henry Tsai, James Urbanic, Carlos Vargas, Nathan Yu, Lyle Ungar, Eric Eaton, Charles Simone, Mark Van Der Laan, Gilmer Valdes . . . 978

Rethinking the Elective Target Volume: Modelling Bilateral Lymphatic Head and Neck Tumour Progression Using Probability Theory.
Kristoffer Moos, Jesper Eriksen, Anne Holm, Stine Korreman . . . 983

Three-Dimensional Visualization of DEBI-NNs for Supporting Interpretability of Predictive Models Relying on Neural Networks.
Amine Boukhari, Ecsedi Boglarka, Nassib Abdallah, David Haberl, Clemens Spielvogel, László Papp, Mathieu Hatt . . . 987

Robustness In Image-Based Data Mining: Bias In Training Data Can Affect The Identification Of Significant Regions.
Tanwiwat Jaikuna, Fiona Wilson, Marcel Van Herk, Peter Hoskin, Catharine West, David Azria, Jenny Chang-Claude, Maria Carmen De Santis, Sara Gutiérrez-Enríquez, Maarten Lambrecht, Petra Seibold, Elena Sperk, Paul R Symonds, Christopher J Talbot, Tiziana Rancati, Tim Rattay, Victoria Reyes, Barry S Rosenstein, Dirk De Ruysscher, Ana Vega, Liv Veldeman, Adam Webb, Marianne Aznar, Eliana Vasquez Osorio 991
Modelling Occult Lymph Node Metastases in HNSCC Patients With a Trinary State Hidden Markov Model.
Yoel Pérez Haas, Roman Ludwig, Esmée Lauren Looman, Vincent Gregoire, Laurence Bauwens, Panagiotis Balermpas, Jan Unkelbach 995

Treatment error classification with time-resolved EPID images with Multiple Instance Learning

Vyacheslav Yarkin[1], Evelyn E.C. De Jong[2], Rutger Hendrix[3], Frank Verhaegen[1] and Cecile J.A. Wolfs[1]

[1]Department of Radiation Oncology (Maastro), GROW School for Oncology and Reproduction, Maastricht University Medical Center, the Netherlands

[2]Instituut Verbeeten, Tilburg, the Netherland

[3]Medical Image Analysis group, Department of Biomedical Engineering, Eindhoven University of Technology, the Netherlands

Abstract

Recent research has investigated the classification of radiotherapy errors in time-integrated (TI) in vivo Electronic Portal Imaging Device (EPID) images with convolutional neural networks, showing promise. However, it has been observed that TI images may cancel out the error presence on γ-maps during dynamic treatments. To address this limitation, time-resolved (TR) γ-maps for each VMAT angle were used to detect realistically simulated treatment errors caused by complex patient geometries and beam arrangements. Typically, such images can be interpreted as a set of frames where only set class labels are provided, making manual analysis difficult and time-consuming. To effectively aggregate individual segment information and evaluate its contribution to error classification we applied a transformer based multiple instance learning (MIL) approach on TR EPID images.

The proposed algorithm showed excellent performance on classification of anatomical, positioning and mechanical errors. Accuracy on the hold-out test set reached 0.94 for 11 error classes.

1 Introduction

Modern radiation techniques like volumetric modulated arc radiotherapy (VMAT) enable precise dose delivery, reducing excessive dose to organs at risk. To ensure clinically acceptable doses, Electronic Portal Imaging Devices (EPID) are used for pre-treatment verification and for in vivo dosimetry (IVD) [1]. However, per-patient transit dosimetry analysis is resource-intensive. Hence, automatic error identification could improve treatment outcome and clinical workflow.

There are only a few studies that proposed to use AI-methods on IVD for VMAT [2-3]. The challenges for IVD 2D transit portal dose verification, is associated with resource-intensive per-patient dosimetry analysis, especially for time-resolved (TR) dose verification. TR γ-maps provide a significant increase of contextual and spatial information about the VMAT dose distribution but cannot be assessed manually in clinical routine due to its quantity of information complexity. Therefore, automatic error identification could benefit this process, and consequently aid further decision making about the patients' treatments.

In this study, we propose a novel approach of treatment error classification with 2D TR γ-maps, based on a deep multiple instance learning (MIL) approach with a transformer model [4] and convolutional neural network (CNN) multi-scale feature extraction.

2 Materials and Methods

Data. The dataset consisted of 26659 simulated 2D TR γ-maps acquired from 53 VMAT treatment plans of 47 lung cancer patients. TR γ-maps were attained by comparing simulated (original) portal dose images (PDI_o) and PDI_e (with only one known error added to treatment plans or computed tomography (CT) images) using gamma analysis. Each VMAT treatment plan contained 2 half-arcs, and each half-arc consisted of 97 beam segments. PDIs were calculated with a model based on pencil beam scatter kernels using in-house developed software [5]. Gamma analysis criteria were set to 3% dose difference and 3 mm distance-to-agreement. In this work, we simulated a wide range of clinically realistic errors, that can occur during radiotherapy treatment (Table 1). Not all treatment errors are expected to manifest for each of the VMAT arc angles, so each of the 97 segments of the whole 2D TR γ-maps could contain a different error representation (Figure 1).

Error Class	Error magnitudes, increment and units
Mediastinum shift	[2 .. 16, 2] mm
Pleural effusion	[5 .. 50, 5] mm
Tumor regression	[10 .. 90, 10] %
Tumor shift	[3 .. 30, 3] mm
Collimator rotation	[0.2 .. 2, 0.2] degrees
MLC shift Random	[0.2 .. 2, 0.2] mm
MLC shift systematic	[0.2 .. 2, 0.2] mm
MU scaling random	[1 .. 10, 1] %
MU scaling systematic	[1 .. 10, 1] %
Positioning rotation	[0.2 .. 1, 0.2] degrees
Positioning translation	[3 .. 21, 3] mm

Table 1: Simulated error classes, and magnitudes of errors applied on CT and treatment plans. Minimum and maximum value provided with the increment size. For instance, anatomical mediastinum shift was acquired with shifting the relative voxels on CT image from its original position to right or left with a magnitude from 2 to 16 mm with a step of size 2 mm.

Deep learning model design. In this study we focused on treatment error classification (multi-class classification) and modified the multiple instance learning (MIL) framework

[6] with vision transformers (ViT) [7]. A pre-trained convolutional neural network (CNN) [1] was adopted for per-segment feature extraction from TR γ-maps. Segment features were treated as tokens for a transformer encoder, to account for the dependencies among all segment representations. MIL exploits an attention-based operator to weight the contributions of each segment embedding.

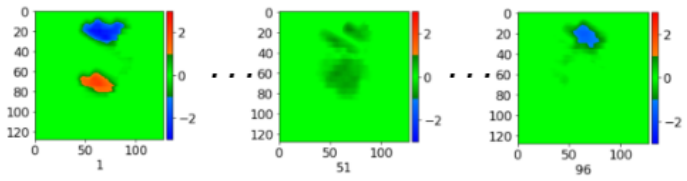


Figure 1: Examples of three individual TR segments of a simulated 21 mm upward tumour shift along the axial plane. Higher γ values coincide with the regions from where the tumor was shifted, causing a higher dose difference (γ value >1, red), in contrast to blue regions with lower dose values than the plan (γ value < -1). The segments with the index 51 does not contain pixels with absolute γ values higher than 1 and these are considered segments where the simulated error has less influence on the dose distribution.

Vision Transformers. ViT use a multi-head self-attention mechanism (MHSA) to process image patches, that are non-overlapping image sections. These patches are transformed into learnable matrices. Multiple heads of the self-attention mechanism are applied in parallel to capture dependencies between different image patch representations. The ViT includes adding a classification token that summarizes class features, as well as a positional token for position-dependent information, with the final step involving the separation of the classification token for prediction.

Multiple Instance Learning. To allow each individual segment to contribute to error classification, we propose a MIL framework. In the context of 2D TR γ-maps, the core principle of MIL revolves in gathering information from the individual segments to make the whole γ-map prediction with an aggregation operator, e.g. a max operator that selects the maximum value among segments and assigns an error class to whole γ-map based on this segment. In this work, to avoid using only one segment to base the classification of the whole TR γ-map on, we used attention-based aggregation to weight each segment representation [8]. Implementation of the transformer MIL involves the following steps: 1) generate feature embeddings of individual segments for the transformer encoder using a pre-trained CNN, 2) model the dependencies along all features with the transformer encoder, 3) aggregate the encoder embeddings with size $n \times d$, where the n = 97 segments segment embeddings of size d, to obtain a whole γ-map representation and 4) transform the aggregated representation to error class probabilities with a linear layer (Figure 2).

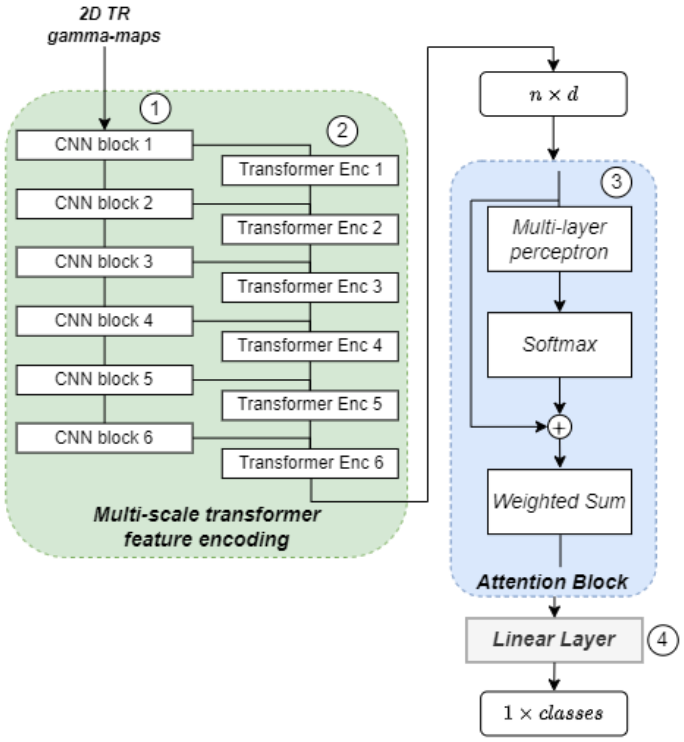


Figure 1: Schematic representation of the transformer MIL architecture. Features are extracted with a CNN from each segment of TR γ-map (1); a transformer encoder models the dependencies among low-dimensional segment embeddings (2); MIL attention aggregates all segment representations (3); whole γ-map level classification is performed (4).

Experimental design and evaluation metrics.
The dataset was partitioned in training, validation and test datasets, so that each patient's simulated TR γ-maps belonged to the same split. Split proportions were therefore set to 70%, 12%, 18% for training, validation and test sets, respectively. The training and validation sets were used for model optimization and tuning hyperparameters. The hold-out test set was only used for model inference, to evaluate the resulting classification performance.

The model was trained with Cross Entropy loss with weight regularization and the Adam [9] optimizer with weight decay set to 0.02. We set an initial learning rate of 6×10^{-5} and gradually decreased it using a cosine learning rate scheduler for 30 epochs. The transformer dropout rate was set to 0.2. The model capacity had approximately 11.4×10^{6} parameters.

For the evaluation metrics, we used accuracy and receiver-operating characteristic (ROC) area under the curve (AUC) to evaluate the treatment error classification. The resulting performance is evaluated on a hold-out test set.

3 Results

Training time for 30 epochs was ~60 hours, training and validation accuracy reached 0.98 and 0.97, respectively. Excellent performance on the hold-out test set on classification of 11 error classes is supported with 0.94 average class accuracy score. The AUC score was 0.98. The majority of misclassifications were seen for mediastinum shift and pleural effusion (Figure 3). In the latter case, 86% of samples that belong to pleural effusion were correctly predicted, while 14% were incorrectly predicted as other anatomical errors. The model demonstrated highest per class accuracy on machine related errors, with MU random and MU systematic accuracy scores of 1.0.

4 Discussion

In this paper we implemented a novel approach of utilizing ViT and MIL for IVD treatment error classification on 2D TR γ-maps. Results show excellent performance on the hold-out test set for a set of simulated errors. The lowest per-class accuracy was 0.86 for pleural effusion, but the model misclassified this error as other anatomical error classes. For tumor regression and tumor shift, misclassifications were found only between these two classes. However, precision for tumor regression was the lowest (0.71) among other classes, with 29 % percent of false positives related to other anatomical error classes. This possibly means that introduced anatomical errors form specific pattern on γ-maps, that the model was able to take into account.

In this study, we focused on TR data, and we assume that additional per-segment spatial information could resolve the issue of averaged out errors in time-integrated (TI) dosimetry [10]. To overcome the limitations of TI data, enormous amount of TR data is effectively leveraged with proposed model architecture. Transformer MIL focuses on error patterns on each individual segments. This is achieved through a transformer encoder, that models the dependencies among segment representations on multiple feature levels. By doing so, the model is effectively capturing the interrelationship between segments that could show different error patterns. Individual patches contribution is enabled with attention aggregation, which is frequently overlooked by transformer based models. Consequently, TR data is effectively utilized and enhance the error classification models on TI data.

Further research could be connected with model interpretation. The possible solution could be to study the transformer's attention matrix due to its ability to capture the dependencies among all segments. Existing approaches focus on visualization of raw attention weights or functions of attention weights. Intuitively, we can extract raw attention matrix of transformer's multi-head self-attention and identify the index of the segment the most contributed to the error classification. However, within the number of encoder layers, the complexity of interactions also raise, and could provide less reliable explanations [11]. Accordingly, the research on more advanced interpretation methods beyond attention-based techniques is needed. This would help to acquire interpretable descriptions on Transformer MIL model decisions.

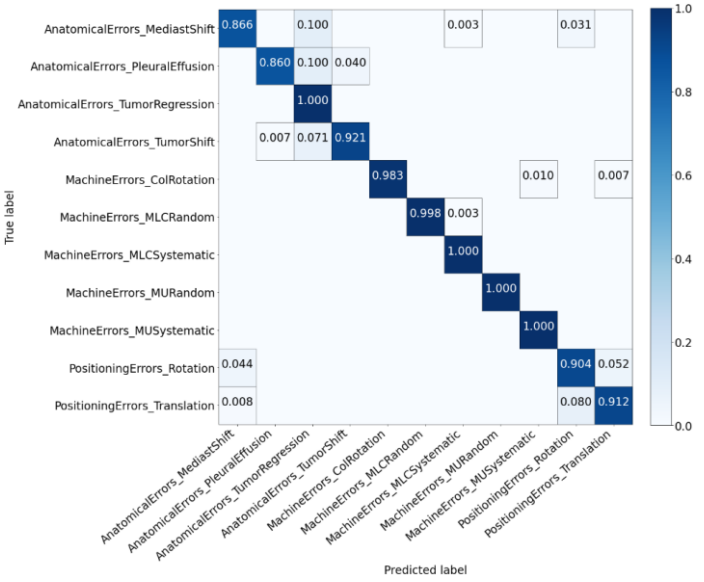


Figure 3: Confusion matrix for treatment error classification. Diagonal values represent the fraction of correctly classified samples of a class out of samples of that class. Empty spaces show no misclassifications for a particular class.

5 Conclusion

In this work we adopted a MIL approach that can automatically enhance dose verification during radiotherapy. Transformer MIL was implemented to explore the spatial information between each of the arc segments of TR images. The following novel contribution was made: we adopted pre-trained CNN for per-segment feature extraction, implemented the transformer-based MIL to account individual segments contribution for error classification. We evaluated the method on a synthetic sample of 2D TR γ-maps with introduced 11 classes of treatment errors with an average accuracy of 0.94.

Funding

This work was partially funded by a research grant from Varian Medical Systems.

References

[1] I. Olaciregui-Ruiz, S. Beddar, P. Greer, N. Jornet , B. McCurdy, G. Fonseca, B. Mijnheer, F. Verhaegen. In vivo dosimetry in external beam photon radiotherapy: Requirements and future directions for research, development and clinical practice. Phys Imag Radiat Oncol. (2020) 15, pp. 108-16

[2] C. Wolfs, R. Canters, F. Verhaegen. "Identification of treatment error types for lung cancer patients using convolutional neural networks and EPID dosimetry." Radiother Oncol (2020), pp. 243-249. DOI: 10.1016/j.radonc.2020.09.048

[3] J. Bedford, I. Hanson. "A recurrent neural network for rapid detection of delivery errors during real-time portal dosimetry." Radiother Oncol (2022) Physics and Imaging in Radiation Oncology pp. 36-43. DOI: 10.1016/j.phro.2022.03.004

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser and I. Polosukhin. (2017). Attention is all you need. Advances in neural information processing systems, 30. DOI: 10.48550/arXiv.1706.03762

[5] W.J. van Elmpt, S. M. Nijsten, B. J. Mijnheer, & A. W. Minken. Experimental verification of a portal dose prediction model. Medical physics (2005). *32*(9), 2805–2818. DOI: 10.1118/1.1987988

[6] A. Myronenko, & Z. Xu, & D. Yang, & H. Roth, & D. Xu, Daguang. Accounting for Dependencies in Deep Learning Based Multiple Instance Learning for Whole Slide Imaging. (2021) DOI: 10.48550/arXiv.2111.01556

[7] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly and J. Uszkoreit. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020). DOI: 10.48550/arXiv.2010.1192

[8] M. Ilse, J. Tomczak, & M. Welling. Attention-based Deep Multiple Instance Learning. Proceedings of the 35th International Conference on Machine Learning, in Proceedings of Machine Learning Research. (2018) DOI: 10.48550/arXiv.1802.04712

[9] I. Loshchilov, & F. Hutter. Decoupled weight decay regularization. (2018) *arXiv preprint arXiv:1711.05101*. DOI: 10.48550/arXiv.1711.05101

[10] L. E J R Schyns, L. C G G Persoon, M. Podesta, W J C van Elmpt, and F. Verhaegen. Time-resolved versus time-integrated portal dosimetry: the role of an object's position with respect to the isocenter in volumetric modulated arc therapy. Physics in Medicine and Biology, (2016) p.61(10): p. 3969-84. DOI: 10.1088/0031-9155/61/10/3969

[11] S. Abnar, & W. Zuidema. Quantifying attention flow in transformers. (2020). arXiv preprint arXiv:2005.00928. DOI: 10.48550/arXiv.2005.00928

A novel iterative reconstruction algorithm for imaging of multi-field proton treatments with PET

Keegan McNamara[1,2], Marina Béguin[2], Günther Dissertori[2], Judith Flock[2], Cristian Fuentes[2], Jan Hrbacek[1], Antony J. Lomax[1,2], Shubhangi Makkar[1,2], Christian Ritzer[2], Damien C. Weber[1,3,4], and Carla Winterhalter[1]

[1]Centre for Proton Therapy, Paul Scherrer Institute, Villigen, Switzerland
[2]Department of Physics, ETH Zurich, Zurich, Switzerland
[3]Department of Radiation Oncology, University Hospital of Zürich, Switzerland
[4]Department of Radiation Oncology, Inselspital, Bern University Hospital, University of Bern, Switzerland

Abstract Treatment verification of multi-field proton therapy using PET imaging may allow for comprehensive verification capabilities through range measurements for each delivered field. We introduce a novel maximum-likelihood expectation maximization (MLEM) algorithm which combines PET data recorded following delivery of each field in a multi-field treatment plan and reconstructs the expected activity distribution induced by each field. We showcase an example use case through simulation of a 3-field treatment plan delivered to a brain tumour, imaged using the open-ring PETITION PET detector. Our novel MLEM approach produced activity distributions with better agreement to the expected positron annihilation distributions for each field, with a decrease in normalised root mean square error by 0.4%-9.7% and an increase in structural similarity index by 4.4%-17.3% when compared to a simplified calculation of the expected activity. We show reduced noise, and the potential for better range predictions for each field using iso-activity analysis along the beams-eye-view.

1 Introduction

Positron emission tomography (PET) imaging of proton therapy patients during treatment provides a potential method for the validation and verification of delivered dose, with the imaging of positron emitting isotopes generated within the patient during treatment serving as a surrogate for delivered dose [1]. The move towards in-beam imaging with novel detector designs, such as the recent open-ring PET detector developed as part of the PETITION (PET for InTensive care units and Innovative protON therapy) project [2], creates the unique opportunity to image multiple treatment fields as they are delivered to the patient. The complex composition, timing structure, and overlap of fields, creates challenges for using the data recorded during a treatment for verification purposes.

The primary goal of this work is to image the activity induced by individual treatment fields delivered to a patient, with the goal being the validation of proton range for each field. In the planning target volume the lateral fall off of one field may overlap with the distal edge of another, or fields may be delivered anti-parallel from one other. As such, disentangling the activity induced by each field presents a unique problem. Low count rates due to the low overall induced activity leads to noisy images, such that a simple subtraction of individual images is not necessarily reliable for the determination of activity induced in each field.

We introduce a novel maximum-likelihood expectation maximization (MLEM) method for reconstruction of the activity induced in patients by multiple treatment fields and imaged using multiple imaging positions. We present an example use case for a brain tumour patient treated with 3 fields. We consider a scenario in which the open-ring PETITION PET scanner is capable of rotating around the patient head, such that the opening of the scanner aligns with the beam direction during delivery of each treatment field.

2 Computational implementation

We consider delivery of N treatment fields to a patient. The delivery of each field induces a particular isotope distribution within the patient geometry for a series of positron emitting isotopes, e.g. ^{10}C, ^{11}C, ^{13}N, and ^{15}O. We denote the activity induced by field f in voxel v by the isotope i by A^f_{iv}. The total number of decays expected in voxel v during imaging period p, from time t_{p0} to t_{p1}, due to isotopes generated by field f is then

$$\lambda^f_{vp} = \sum_i A^f_{iv} \int_{t_{p0}}^{t_{p1}} dt\, e^{-(t-t_f)\ln 2/t_{i,1/2}}, \tag{1}$$

where t_f is the time that field f finished delivery, $t_{p0} > t_f$, and $t_{i,1/2}$ is the half life of isotope i. We make the simplifying assumption that the activity distribution of all isotopes generated by a treatment field are identical up to a scale factor, i.e. $A^f_{iv} = w_i A^f_v$, such that we can write the total number of decays expected during imaging period p to be

$$\lambda_{vp} = \sum_f A^f_v d_{pf} \tag{2}$$

where $d_{pf} = 0$ if $t_{p0} < t_f$, otherwise

$$d_{pf} = \sum_i w_i \int_{t_{p0}}^{t_{p1}} dt\, e^{-(t-t_f)\ln 2/t_{i,1/2}}. \tag{3}$$

We consider imaging of the induced activity by a maneuverable PET scanner, such that during each imaging period p the PET scanner is in a different imaging position, denoted by α_p. In our case the open-ring PET scanner rotates during the time between delivery of each treatment field into a new

imaging position. The number of expected coincidences in line of response (LOR) j acquired during each period is

$$\tilde{y}_{jp} = \sum_v c_{jv}^{\alpha_p} \lambda_{vp}, \quad (4)$$

where $c_{jv}^{\alpha_p}$ is the system matrix describing the probability of detecting a positron annihilation in voxel v in LOR j while in imaging position α_p. Comparison to the recorded number of coincidences y_{jp} using the standard Poissonian likelihood function, along with the standard derivation of the MLEM method, leads us to the iterative update procedure

$$A_v^{f,(k+1)} = A_v^{f,(k)} \frac{1}{\sum_{pj} d_{pf} c_{vj}^{\alpha_p} a_j^{\alpha_p}} \sum_{pj} \frac{y_{jp} d_{pf} c_{vj}^{\alpha_p}}{\sum_{v'f'} A_{v'}^{f',(k)} d_{pf'} c_{v'j}^{\alpha_p}}, \quad (5)$$

where we also include attenuation correction via the term $a_j^{\alpha_p} = \exp\left(-\sum_v l_{jv}^{\alpha_p} \mu_v\right)$, which is the attenuation of LOR j in imaging position p through the patient geometry, defined by the mu map μ_v. This algorithm is unique, in that the activity induced by all fields influences the determination of the activity of all others.

The computational implementation of our algorithm now requires the system matrix of the scanner in multiple positions. We use the rotational, mirror, and translational symmetries of the scanner to generate the system matrix at each position α_p from only 3 stored system matrices, reducing the storage requirements by a factor of 25.5. The matrix elements $c_{vj}^{\alpha_p}$ are generated using a multi-ray approach, which accounts for the possible interaction points within each crystal. Inter-crystal scatter, scatter correction, and random corrections are not applied.

Data: $y_{jp}, c_{vj}^{\alpha_p}, w_i, t_{p0}, t_{p1}, \sum_j a_{vj}^{\alpha_p} c_{vj}^{\alpha_p}$

Calculate d_{pf} from t_{p0}, t_{p1};
$s_v^f = \sum_{pj} d_{pf} a_{vj}^{\alpha_p} c_{vj}^{\alpha_p}$;
Initialise all A_v^f;
for $k \leftarrow 1$ **to** N_{it} **do**
  $m_v^f \leftarrow 0$;
  for j, p; *s.t.* $y_{jp} \neq 0$ **do**
    Load $c_{vj}^{\alpha_p}$;
    $c \leftarrow \sum_{vf} A_v^f d_{pf} c_{vj}^{\alpha_p}$;
    $m_v^f \leftarrow m_v^f + \frac{1}{c} y_{jp} d_{pf} c_{vj}^{\alpha_p}$;
  end
  $A_v^f \leftarrow \frac{A_v^f m_v^f}{s_v^f}$;
end

The algorithm is implemented as indicated above. The measured and time stamped list mode data, y_{jp}, are read in, indicating which matrix elements $c_{vj}^{\alpha_p}$ must be loaded into memory. The assumed isotope composition along with the timing for the fields and imaging periods are used to calculate d_{pf}. The time and isotope weighted sensitivity maps are calculated. Following this, N_{it} MLEM iterations are applied.

3 Results

We simulated treatment of a brain tumour using 3 treatment fields, performed at a fixed Gantry angle of 90°, and couch angles -75°, 0°, and 90°. Following the delivery of each field the open-ring PET scanner is rotated by 90° in order to align with the next treatment delivery position. Treatment delivery and activation of the patient are simulated using the GPU accelerated Monte Carlo engine FRED [3, 4]. We scored production of ^{10}C, ^{11}C, ^{13}N, ^{14}O, ^{15}O, ^{30}P, and ^{38}K. The scored isotopes are then used as input in the Monte Carlo tool Gate [5], which simulates the decay of isotopes and the detection of produced gammas by the PET detector. The coincidences per second are shown in Fig. 1. It is clear that a significant portion of the counts recorded in the second and third imaging periods are due to the previously delivered fields. We also note that the count rate does not follow a typical exponential decay, as the rotation of the scanner moves the field of view of the PET scanner such that the higher sensitivity region overlaps with the activated region of the patient.

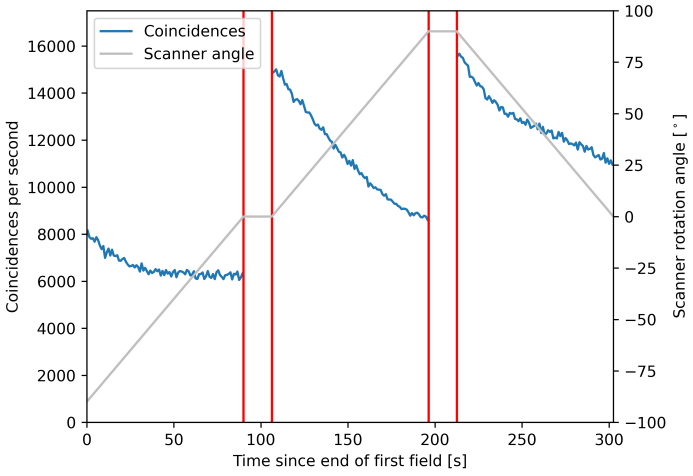


Figure 1: Coincidence rate and position of scanner over time. The red lines indicate the start and end of beam-on time.

We compare our novel MLEM algorithm to a simplified method which may be used to determine the activity induced by each individual field. The simplified method uses an image reconstructed from each imaging period, i.e. a direct reconstruction of λ_{vp}. Inversion of Eq. 2 then gives an estimation of A_v^f. The inversion of Eq. 2 is not strictly positive definite. 10 MLEM iterations were used for reconstructions using the simplified solution, while 30 iterations were used for the novel MLEM method. As the novel approach solves all 3 fields simultaneously, this leads to a comparable number of iterations for both techniques. We compare deduced images of A_v^f based on our technique

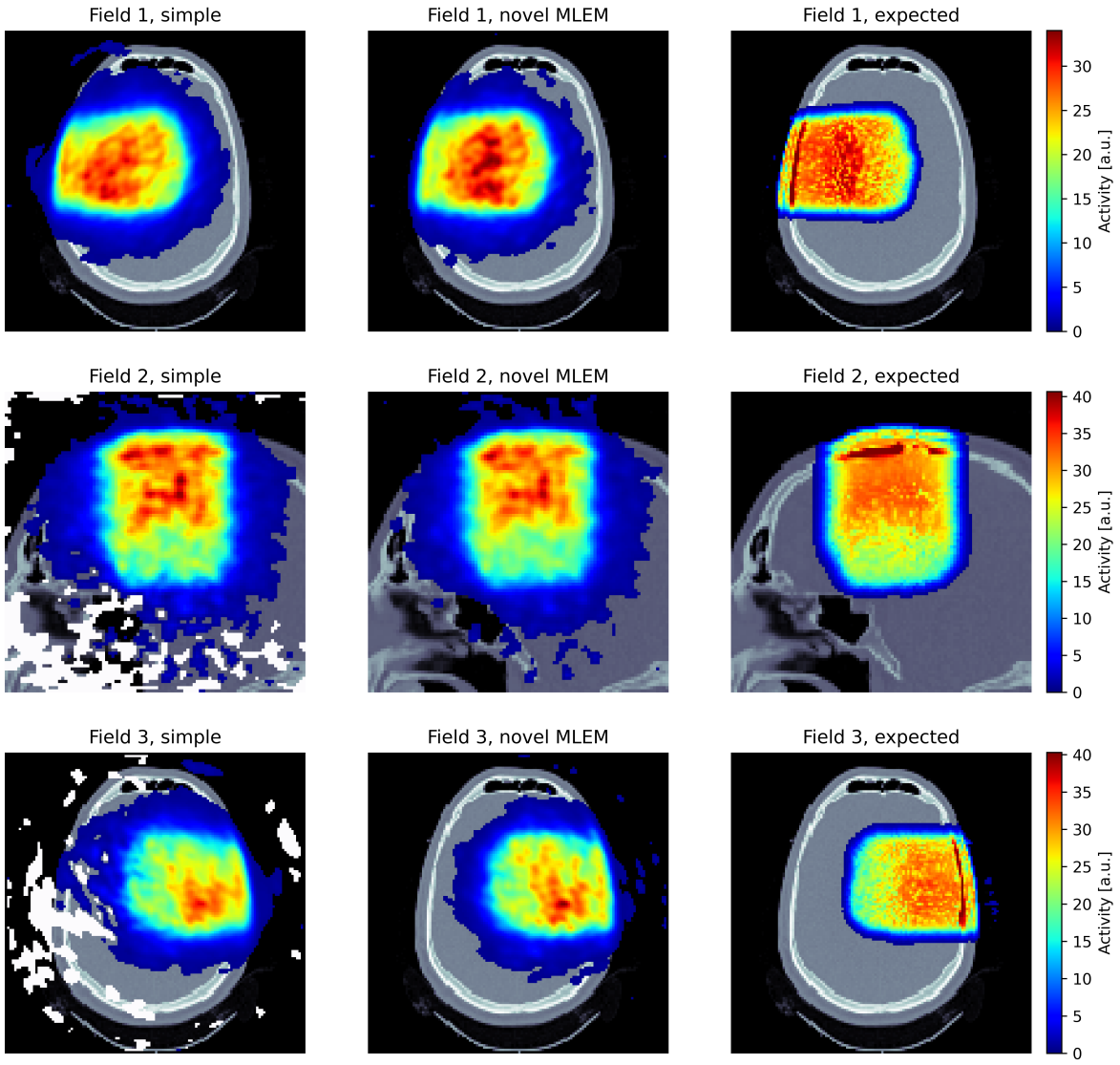


Figure 2: Reconstruction of A_v^f based on (left) a simplified solution to Eq. 2, and (middle) the novel MLEM algorithm proposed. The reconstructions are compared to the normalised predicted activity induced by each field (right). The activity predicted by the simplified solution for field 2 and 3 is negative in some regions, shown in white.

and a simplified approach in Fig. 2. The expected results shown in Fig. 2 are calculated using the distribution of the positron annihilation locations predicted by GATE.

Fig. 2 shows a reduced signal to noise ratio in all images using the novel MLEM, while showing comparable predicted distal fall off for all fields. In table 1 we the consider normalised root mean square error (NRMSE) and structural similarity index (SSIM) between the reconstructions and the predicted activity. For all fields the novel MLEM method has lower NRMSE by 0.4%-9.7% than the simplified method, while the SSIM is higher by 4.4%-17.3%.

As the intention of the reconstructed images is range verification in proton therapy, we also consider the range that may be assigned for each field. In order to assign a range we transformed the reconstructed images into beams eye view (BEV), and calculated the isoactivity surface along the beam direction. This gives a range value for each point in the plane tangent to the beam direction. Due to a lack of scatter correction, activity between 0-5 a.u. distal to the beam did not agree well with the expected activity distribution. We found that an isosurface at 16 a.u. gave good agreement across all fields, corresponding to approximately 50% of the high activity region in brain. The difference between the range assigned by the reconstructions and the expected distribution is shown in Fig. 3. In table 2 the average shift of the BEV

Table 1: NRMSE and SSIM between the expected activity distribution and the reconstructed activity using the novel MLEM approach and a simplified subtraction.

	Field 1	Field 2	Field 3
Novel MLEM (NRMSE)	0.253	0.216	0.266
Simple (NRMSE)	0.254	0.239	0.281
Novel MLEM (SSIM)	0.439	0.466	0.411
Simple (SSIM)	0.421	0.398	0.390

images are shown. For field 1 and 3 the range predicted by the novel MLEM is closer on average to the expected result, while for all fields the standard deviation in the shift is lower with the novel MLEM.

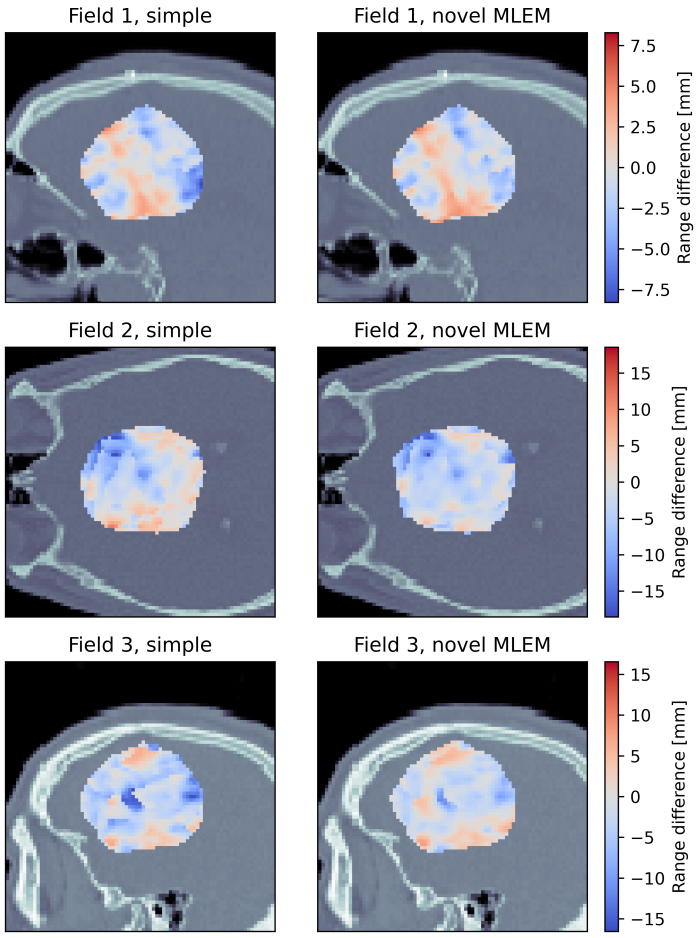


Figure 3: Difference in range assigned along the BEV at a fixed isoactivity curve. Difference is between the reconstructions and predicted total activity as shown in Fig. 2.

Table 2: Average (standard deviation) of the difference in range assigned for the isoactivity curves as shown in Fig. 3. All values in mm.

	Novel MLEM	Simple
Field 1	-0.2 (1.9)	-0.7 (2.0)
Field 2	-3.4 (3.6)	-2.1 (4.6)
Field 3	-1.0 (3.0)	-2.7 (4.0)

4 Discussion

Our algorithm introduces a possible approach for validation of the range of each field delivered in a multi-field proton therapy plan via PET imaging. The ability to disentangle the activity induced by each field may allow for verification of stopping power predictions, CT calibration, or even detection of anatomical changes between fractions.

We considered imaging of a 3 field proton plan, with imaging taking place over a 90 second period following each field delivery. This imaging scheme would allow for minimal interruption to the treatment workflow, as imaging occurs during the time required for couch and gantry movements. We found that our novel MLEM generated images with less noise than those generated via a simplified calculation of the activity. The NRMSE was between 0.4%-9.7% lower than the simplified approach, while the SSIM was higher by 4.4%-17.3%. The range predicted by both methods was comparable, however thanks to the lower noise we saw a lower standard deviation in the assigned range for all fields using the novel approach.

Throughout we have only considered a single patient example, and note that further work is required to assess the capabilities in different scenarios. Fields delivered at comparable angles may have significant overlap, and further work is required in order to test the capabilities of our algorithm to disentangle the activity of each field.

5 Conclusion

We have presented a novel MLEM algorithm for reconstruction of the PET activity induced by individual fields in multi-field proton therapy treatments. Use of a rotating open-ring scanner, such as the one recently built as part of the PETITION project, gives a unique opportunity for such in-beam and inter-field measurements. The determination of range on a field by field basis may allow for greater sensitivity and further applications of PET for treatment verification in proton therapy.

6 Acknowledgements

This work was supported by the Swiss National Science Foundation, Grant No. CR-SII5189969.

References

[1] A.-C. Knopf and A. Lomax. “In vivo proton range verification: a review”. *Phys. Med. Biol.* 58.15 (2013), R131–R160.

[2] C. E. Ritzer, V. Commichau, J. Flock, et al. “The PETITION Project - Initial Monte-Carlo Simulations”. *Joint Annual Meeting of the Austrian Physical Society and the Swiss Physical Society* (2021), p. 137.

[3] A. Schiavi, M. Senzacqua, S. Pioli, et al. “Fred: a GPU-accelerated fast-Monte Carlo code for rapid treatment plan recalculation in ion beam therapy”. en. *Phys. Med. Biol.* 62.18 (2017). Publisher: IOP Publishing, pp. 7482–7504.

[4] K. McNamara, A. Schiavi, D. Borys, et al. “GPU accelerated Monte Carlo scoring of positron emitting isotopes produced during proton therapy for PET verification”. *Phys. Med. Biol.* 67.24 (2022), p. 244001.

[5] D. Sarrut, N. Arbor, T. Baudier, et al. “The OpenGATE ecosystem for Monte Carlo simulation in medical physics”. *Phys. Med. Biol.* 67.18 (Sept. 2022), p. 184001.

Evaluation of algorithms for dose reconstruction from prompt-gamma radiation in proton therapy

Beatrice Foglia[1], Chiara Gianoli[1], Takamitsu Masuda[2], Elisabetta De Bernardi[3], Thomas Bortfeld[4], Katia Parodi[1], and Marco Pinto[1]

[1]Department of Experimental Physics - Medical Physics, Ludwig-Maximilians-Universität (LMU) Müenchen, Garching bei München, DE
[2]National Institutes for Quantum Science and Technology, Chiba, JP
[3]School of Medicine and Surgery, University of Milano-Bicocca, Monza, IT
[4]Department Of Radiation Oncology, Massachusetts General Hospital (MGH), Boston (MA), USA.

Abstract Dose and range monitoring with secondary radiation, like prompt-gammas, can help minimizing the range uncertainties in proton therapy. This work aims at the reconstruction of the delivered proton dose from prompt-gamma distributions. Some techniques have already been proposed in the literature to reconstruct the dose from secondary radiation, mostly positron emitters. Among them, very promising methods are the analytical deconvolution approach, the evolutionary algorithm and the maximum-likelihood expectation-maximization algorithm. Only the evolutionary algorithm has been applied to prompt-gamma distributions so far. Within this work, the applicability of these approaches to prompt-gamma distributions was assessed with simulated mono- and polynergetic proton beams irradiating homogeneous and heterogeneous phantoms. The reconstructed depth-dose profiles were compared with their corresponding simulated dose distributions using different metrics. Overall, the reconstructed dose profiles match the simulated ones, and range differences evaluated at different distal dose levels were found to be lower than or in the order of 1 mm.

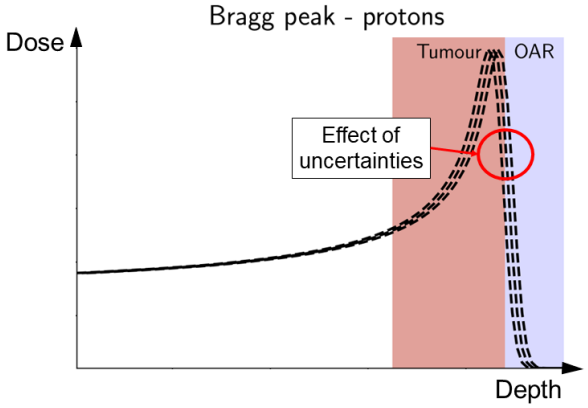


Figure 1: Effect of uncertainties on the dose deposited by protons.

1 Introduction

In cancer therapy, the characteristics of protons are in principle favourable compared to the ones of photons. The dose deposited by protons is lower in the entrance channel, it increases until a maximum, known as Bragg peak, is reached, and then it rapidly fall-off at the end of the particle path, better known as range [1]. The Bragg peak position can be tuned changing the initial beam energy. These properties allow the delivery of high doses to the tumor, while sparing the surrounding healthy tissues.

However, many factors can cause proton range uncertainties in clinical practice, among them the patient positioning, possible anatomical changes and the conversion from X-ray computed tomography (CT) numbers (measured in Hounsfield units) into tissue stopping power ratio relative to water. These uncertainties limit the achievable maximum potential of proton therapy [2, 3], as shown in figure 1. To achieve a more conformal dose distribution in the tumour target, it is necessary to monitor and, where needed, correct for such uncertainties. This is why *in vivo* range verification techniques were proposed in the past few years.

One possibility for range verification is through the detection of secondary prompt-gammas (PG), first proposed by [4] and then experimentally verified by [5]. The emission of PG is correlated with the penetration path of protons in tissue, and the measurement of PG can be used to draw conclusions on the proton range.

For range verification purposes, the expected PG signals related to planned dose distributions are needed. The expected PG distributions can be, for example, modelled by means of Monte Carlo simulations or by a filtering approach. The latter was introduced by Parodi and Bortfeld [6] to relate the positron emission tomography (PET) activity profile (i.e. 1D signal) to the depth-dose profile of proton irradiation. The main idea behind the filtering is the possibility to determine a PET profile from the convolution of a dose distribution with a so-called filter function f. Validation of this approach and its extension to 3D distributions and PG monitoring came with successive works [7–10].

Getting a direct information on the deposited dose from the PG emission is not straightforward: the underlying mechanisms of these two physical phenomena are different, governed by electronic and nuclear interactions, respectively. Using an inverse filter function f^{-1} to obtain the actual delivered dose from the PG distribution represents an ill-posed deconvolution problem. To face this challenge, several deconvolution techniques have been proposed in the literature. In this regard, three different approaches were chosen for their inherent use of the filtering formalism: the analytical deconvolution method [11], the evolutionary algorithm [12–14] and the maximum-likelihood expectation-maximization (MLEM) algorithm [15, 16]. These strategies were all applied to reconstruct dose delivered by protons from PET data, while the evolutionary algorithm also explored PG distribu-

tions [12, 14]. In the present work, the mentioned algorithms are applied to reconstruct dose deposited in homogeneous and inhomogeneous (slab) phantoms, for simulated pristine Bragg peaks and spread out Bragg peaks (SOBPs). The perfomance of the techniques are evaluated and compared in terms of different metrics.

2 Materials and Methods

The 1D signal $P(z)$ of PG emitted along the proton beam propagation depth z in a homogeneous target can be obtained from the corresponding 1D dose distribution $D(z)$ by convolving (i.e. filtering) the dose profile itself with a filter function $f(z)$:

$$P(z) = D(z) * f(z) \tag{1}$$

The filter functions can be calculated exploiting the so-called $\tilde{Q}_\nu$ functions, defined as convolutions of Gaussian and power-law functions. For the filter calculation, its properties and the extension to inhomogeneous targets, please refer to [6, 9, 10]. The analytical deconvolution approach [11], consists of an iterative maximum *a posteriori* approach (MAP) based on the minimization of the functional J:

$$J[D(z)] = S[D(z)] + \lambda R[D(z)] \tag{2}$$

S is a similarity term and its choice depends on the underlying noise distribution. For this work, a least-squares (LS) term was chosen:

$$S_{LS}[D(z)] = \int |D(z) * f(z) - P(z)|^2 dz \tag{3}$$

R is a regularization term, weighted against S by a factor λ. R considers a prior knowledge of the signal to stabilize the ill-posed deconvolution problem by adding some constraints to restrict the amount of the possible solutions. The Tikhonov-Miller (TM) regularization was chosen for this work, since its use was well motivated in the literature [11, 13]:

$$R_{TM}[D(z)] = \int |\nabla D(z)|^2 dz \tag{4}$$

In [11] it is pointed out that the noise amplitude in the reconstructed dose reflects the noise in PET signals. The same concept remains valid for PG signals. In this work, an attempt in trying to eliminate the noise was made exploiting the $\tilde{Q}_\nu$ functions formalism. One or more $\tilde{Q}_\nu$ functions can be used to fit dose and PG profiles. So, instead of optimizing bin by bin in the dose distribution as in [11], the optimization is carried out only in terms of $\tilde{Q}_\nu$ functions' parameters.
To define convergence of these deconvolution algorithms, in this work a tolerance value was defined, which sets the acceptable error on both the optimized variable (i.e. $D(z)$) and the minimized cost function (i.e. $J[D(z)]$). A tolerance equals to 10^{-10} was used in the deconvolution approach implemented as in [11], while a value of 10^{-6} was used for the optimization carried out only in terms of $\tilde{Q}_\nu$ functions' parameters.

The evolutionary algorithm consists of iteratively mutating a population of N depth-dose profiles $D_j(z)$, which are convolved with the filter functions, in order to obtain predictions of depth-PG profiles $P_j(z)$. Each of these j-th depth-PG predictions is compared to the ground-truth one, potentially measured. The agreement of each comparison j at every iteration i is assessed by a quantity called fitness value f_j^i:

$$f_j^i = \frac{1}{1 + NRMSE[D_j^i(z)] + \lambda R[D_j^i(z)]} \tag{5}$$

$NRMSE$ is the normalized root mean squared error for the j-th individual, R is the regularization term and λ the balancing factor. To select which individuals will be propagated in the following iteration $i+1$, a fitness selection is applied. In this work, for the evolutionary algorithm a maximum number of iterations equals to 1500 was enough to reach convergence. The algorithm was anyway stopped earlier if the difference between the NRMSE calculated on the depth-PG profiles between successive iterations was lower than a defined tolerance value, set to 10^{-8}.
In the MLEM algorithm, the convolution used to calculate the filter is expressed in its equivalent matrix product form. A system matrix C based on the filter functions is built. The dose reconstruction algorithm is based on the maximization of the likelihood, defined as the probability that a certain PG distribution at the voxel i obtained from applying the filtering to a dose distribution corresponds to the related measured signal P_i. If the dose in a certain voxel j at the iteration k is defined as δ_j^k, the regularization term as R and the balancing factor as λ, the likelihood maximization brings to the MLEM formula:

$$\delta_j^{k+1} = \frac{\delta_j^k}{\left[1 + \lambda \left(\frac{dR}{d\delta^k}\right)_j\right] \sum_i C_{ij}} \sum_i \frac{P_i C_{ij}}{\sum_j C_{ij'} \delta_{j'}^k} \tag{6}$$

The convergence criterion for the MLEM algorithm applied in this work was similar to the one used for the evolutionary algorithm. The maximum number of iterations was set at 150, while the tolerance value was equals to 10^{-6}.
The simulated depth-dose and PG emission profiles were obtained with the general purpose MC toolkit Geant4 [17], version 10.07.p03. The performances of the algorithms were compared via evaluation of different metrics to assess the match between the simulated and the reconstructed dose profiles: the NRMSE, evaluated over the entire curve and also only on the distal area; the $|\Delta R_{80}|$, $|\Delta R_{50}|$ and $|\Delta R_{10}|$, where $|\Delta R_{\%}|$ is defined as the absolute difference between the distal positions corresponding to the defined percentage (%) of the dose of the two compared curves; the peak (for single Bragg peak cases) or plateau (for SOBP) width (pw) percentage difference $\Delta pw(\%)$ between the reconstructed curve (rec) and the simulated (ground truth, GT) one, defined as

$$\Delta pw(\%) = 100 \cdot \frac{pw_{rec} - pw_{GT}}{pw_{GT}} \tag{7}$$

3 Results

Examples of obtained results related to the reconstruction of SOBP curves in the slab phantom are shown in figure 2. Table 1 collects the different metrics used to compare the reconstructed distributions in figure 2 with the simulated one.

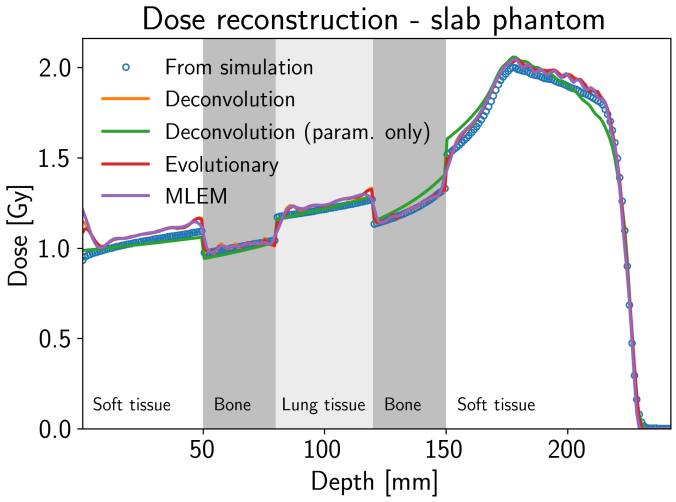


Figure 2: SOBP reconstruction in the slab phantom.

Table 1: Metrics used to compare the SOBP reconstructed in the slab phantom with the simulated one.

Metric	Deconvolution	Deconvolution (param. only)	Evolutionary	MLEM
NRMSE ($\times 10^{-2}$)	3.2	3.3	3.1	3.5
NRMSE plateau ($\times 10^{-2}$)	2.7	3.92	2.6	2.8
$\lvert\Delta R_{80}\rvert$ (mm)	0.37	1.18	0.39	0.58
$\lvert\Delta R_{50}\rvert$ (mm)	0.07	0.19	0.12	0.32
$\lvert\Delta R_{10}\rvert$ (mm)	0.18	0.07	0.17	0.36
Δpw (%)	-0.79	-17.75	0.24	-4.29

4 Discussion

Three algorithms already applied in the literature to reconstruct depth-dose profiles from PET signals were applied in this work to depth-PG distributions: the deconvolution approach, the evolutionary algorithm and the MLEM algorithm. The feasibility of the methods in the context of PG monitoring was validated via application to simulated depth-dose and depth-PG profiles in homogeneous and inhomogeneous (slab) phantoms, for pristine Bragg peaks and SOBPs. Looking at the metrics, the range differences $|\Delta R_{\%}|$ at different distal percentages of the dose maximum are generally lower than 1 mm (only one value is higher).

The reconstructed dose profiles exhibit noise, which reflect the noise present in the PG distributions. The analytical deconvolution approach in terms of $\tilde{Q}_v$ functions parameters was introduced to eliminate the noise in the reconstructed curves. This method has also the advantage of speed. However, the approximations introduced in this method makes it less accurate than the other methods, as can be noticed comparing the shown plots and the values of the different metrics, especially the NRMSE. This technique can anyway still be improved and the advantages are worthwhile being futher investigated.

The evolutionary algorithm was already applied to PG data [12, 14]. However, this work is still innovative for aspects related to the generation of the filters, in particular to how their small energy dependence is tackled, and how they are applicable to inhomogeneous targets. In fact, differently from [12, 14], in this work the filters found in a homogeneous target were optimized for energy values within the entire clinically used range. Also, to deal with inhomogeneous targets, instead of performing a double evolutionary algorithm as in [14], the filter is first generated in a homogeneous space, then range conversions and factors related to tissue composition are applied.

In the MLEM application, the regularization was applied differently than in [15, 16], where the MLEM algorithm was implemented in two steps and the total-variation (TV) regularization was applied to the dose itself, in an algorithm referred to as EM-TV algorithm. In this work, a TM regularization was applied as in the other methods in order to make a meaningful comparison among them. For the same reason, instead of working on the signal itself, the regularization was applied to the objective function, i.e. the likelihood, by implementing a Bayesian reconstruction. To carry this out, the one-step-late (OSL) approximation, introduced first by [18], was used.

Future work foresees the reconstruction from detected PG data instead of emitted ones. This would allow to study more realistic scenarios, where PG noise has a major impact compared to simulations. At first, a simulated detection response is going to be implemented, following the behaviour of already commerciable available PG cameras. Measurements will be then used for algorithm validation. A study of the algorithms behaviour when different levels of noise are present is planned.

Difference metrics and criteria (i.e. number of iterations and tolerances), which were described in section 2, were applied to define the convergence of the algorithms. A comparative convergence analysis within the different algorithms, based on the same metrics and criteria, is foreseen.

A work in progress is related to the extension of the approaches to 3D distributions, and to the consideration of lateral heterogeneities, in order to be able to reconstruct the entire dose distributions from emitted PG in more complex phantoms and patients.

The emission of PG happens in a timescale below the nanosecond. This makes PG suitable for the final purpose of real-time adaptive particle therapy (RT-APT) applications, towards which the interest of the radiotherapy community is increasing. Based on benefits seen from online adaptive

proton therapy [19, 20], RT-APT aims at reducing proton therapy uncertainties not only between treatment fractions (inter-fraction), but also during a single fraction itself (intra-fraction). A simple scheme of how the dose reconstruction from PG would be involved in the RT-APT workflow is presented in figure 3. During the delivery, the dose distribution reconstructed from PG detected during the irradiation of each spot would be compared, in a short time frame, with the planned dose distribution. In case of mismatches, adaptation measures would be applied. This work is thought to represent a step forward this direction.

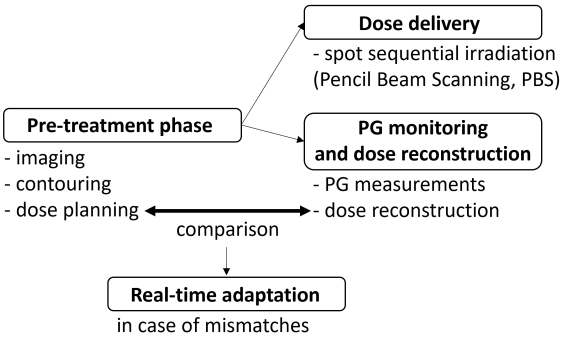


Figure 3: Scheme of RT-APT workflow.

5 Conclusion

PG yields can be exploited to tackle the challenges of dose evaluation in proton therapy. Here, the analytical deconvolution approach, the evolutionary algorithm and the MLEM algorithm, which allow to reconstruct the dose from secondary radiation, were applied successfully to perform dose reconstruction from emitted PG signals coming from simulated 1D mono- and polyenergetic beams in homogenous and heterogenous (slab phantom) targets. Extension to 3D signals and to phantoms with lateral inhomogeneities is in progress. Detected PG data will replace the emitted ones in future, and measurements will be considered for validation. This project is thought to represent a step forward real-time adaptive particle therapy.

This work is performed as part of the RAPTOR project, funded by the EU's Horizon 2020 MSCA, G.A. No. 955956.

References

[1] R. R. Wilson. "Radiological Use of Fast Protons". *Radiology* 47.5 (Nov. 1946), pp. 487–491. DOI: 10.1148/47.5.487.

[2] A. C. Knopf and A. Lomax. "*In vivo* proton range verification: a review". *Phys Med Biol* 58.15 (Aug. 7, 2013), R131–R160. DOI: 10.1088/0031-9155/58/15/R131.

[3] K. Parodi and J. C. Polf. "*In vivo* range verification in particle therapy". *Med Phys* 45.11 (Nov. 2018). DOI: 10.1002/mp.12960.

[4] F. Stichelbaut and Y. Jongen. "Verification of the proton beam position in the patient by the detection of prompt gamma-rays emission". *Talk at the 39th PTCOG meeting (unpublished)* (2003).

[5] C. H. Min, C. H. Kim, M. Youn, et al. "Prompt gamma measurements for locating the dose falloff region in the proton therapy". *Appl Phys Lett* 89.18 (Oct. 30, 2006), p. 183517. DOI: 10.1063/1.2378561.

[6] K. Parodi and T. Bortfeld. "A filtering approach based on Gaussian-powerlaw convolutions for local PET verification of proton radiotherapy". *Phys Med Biol* 51.8 (2006), pp. 1991–2009. DOI: 10.1088/0031-9155/51/8/003.

[7] F. Attanasi, A. Knopf, K. Parodi, et al. "Extension and validation of an analytical model for *in vivo* PET verification of proton therapy - a phantom and clinical study". *Phys Med Biol* 56.16 (2011), pp. 5079–5098. DOI: 10.1088/0031-9155/56/16/001.

[8] K. Frey, J. Bauer, D. Unholtz, et al. "TPS $_{\text{PET}}$ - A TPS-based approach for *in vivo* dose verification with PET in proton therapy". *Phys Med Biol* 59.1 (2014), pp. 1–21. DOI: 10.1088/0031-9155/59/1/1.

[9] T. Hofmann, A. Fochi, K. Parodi, et al. "Prediction of positron emitter distributions for range monitoring in carbon ion therapy: an analytical approach". *Phys Med Biol* 64.10 (2019), p. 105022. DOI: 10.1088/1361-6560/ab17f9.

[10] M. Pinto, K. Kröniger, J. Bauer, et al. "A filtering approach for PET and PG predictions in a proton treatment planning system". *Phys Med Biol* 65.9 (2020), p. 095014. DOI: 10.1088/1361-6560/ab8146.

[11] S. Remmele, J. Hesser, H. Paganetti, et al. "A deconvolution approach for PET-based dose reconstruction in proton radiotherapy". *Phys Med Biol* 56.23 (2011), pp. 7601–7619. DOI: 10.1088/0031-9155/56/23/017.

[12] A. Schumann, M. Priegnitz, S. Schoene, et al. "From prompt gamma distribution to dose: a novel approach combining an evolutionary algorithm and filtering based on Gaussian-powerlaw convolutions". *Phys Med Biol* 61.19 (2016), pp. 6919–6934. DOI: 10.1088/0031-9155/61/19/6919.

[13] T. Hofmann, M. Pinto, A. Mohammadi, et al. "Dose reconstruction from PET images in carbon ion therapy: a deconvolution approach". *Phys Med Biol* 64.2 (2019), p. 025011. DOI: 10.1088/1361-6560/aaf676.

[14] Z.-Y. Yao, Y.-S. Xiao, and J.-Z. Zhao. "Dose reconstruction with Compton camera during proton therapy via subset-driven origin ensemble and double evolutionary algorithm". en. *Nucl Sci Tech* 34.4 (Apr. 2023), p. 59. DOI: 10.1007/s41365-023-01207-1.

[15] T. Masuda, T. Nishio, J. Kataoka, et al. "ML-EM algorithm for dose estimation using PET in proton therapy". *Phys Med Biol* 64.17 (2019), p. 175011. DOI: 10.1088/1361-6560/ab3276.

[16] T. Masuda, T. Nishio, A. Sano, et al. "Extension of the ML-EM algorithm for dose estimation using PET in proton therapy: application to an inhomogeneous target". *Phys Med Biol* 65.18 (2020), p. 185001. DOI: 10.1088/1361-6560/ab98cf.

[17] S. Agostinelli, J. Allison, K. Amako, et al. "Geant4 - a simulation toolkit". *Phys Med Biol* 506.3 (2003), pp. 250–303. DOI: 10.1016/S0168-9002(03)01368-8.

[18] P. J. Green. "Bayesian reconstructions from emission tomography data using a modified EM algorithm". *IEEE T-MI* 9.1 (1990), pp. 84–93. DOI: 10.1109/42.52985.

[19] H. Paganetti, P. Botas, G. C. Sharp, et al. "Adaptive proton therapy". *Phys Med Biol* 66.22 (2021), 22TR01. DOI: 10.1088/1361-6560/ac344f.

[20] F. Albertini, M. Matter, L. Nenoff, et al. "Online daily adaptive proton therapy". *BJR* 93.1107 (2020). DOI: 10.1259/bjr.20190594.

An analytical approach to predict 3D positron emitter distribution in carbon ion therapy

Tianxue Du[1], Julia Bauer[2,3,4], Katia Parodi[1], Marco Pinto[1]

[1]Department of Medical Physics, Ludwig-Maximilians-Universität München, Munich, Germany.
[2]Department of Radiation Oncology, Heidelberg University Hospital, Heidelberg, Germany.
[3]Heidelberg Institute of Radiation Oncology (HIRO), Heidelberg, Germany.
[4]National Center for Tumor diseases (NCT), Heidelberg, Germany.

Abstract Carbon ion therapy provides high accuracy in targeting tumors while sparing healthy tissues, and could highly benefit from possibilities of range verification due to its sensitivity to range uncertainties. Positron emission tomography (PET) is commonly used for range verification, but as PET signals are not directly proportional to dose distributions, comparing them to a predicted positron emitter distribution (PED) is one viable approach. To this end, an analytical approach was previously suggested to predict 1D PED from depth dose profiles. We revised this method by introducing finer modelling functions and considering more positron emitters, and extended it to 3D. In this contribution, we present our proposed 3D PED generation framework based on a pencil beam algorithm and ray-casting method. The method is validated in in-silico studies against full Monte Carlo calculations for several phantoms as well as realistic computed tomography (CT) scans from patient data.

1 Introduction

Carbon ion therapy, benefiting from its characteristic Bragg peak, allows a high accuracy in targeting tumors while sparing healthy tissues. Moreover, carbon ions exhibit higher linear energy transfer (LET) and relative biological effectiveness (RBE) compared to photons and protons, making carbon ion therapy a good treatment option against various cancers [1].

However, range uncertainties can lead to improper dose delivery. Therefore, range verification is highly advantageous, with positron emission tomography (PET) being the most commonly used method [2, 3]. Since the PET signal is not directly proportional to the delivered dose, the measured activity and dose distribution cannot be directly compared to each other [4]. A prediction of the positron emitter distributions (PED) can be obtained with Monte Carlo (MC) simulations, and comparing it to the measurement allows for treatment verification [5]. Despite their accuracy, MC simulations require high computational effort and are time consuming.

To overcome this, alternative analytical methods for PET prediction have been developed. An experimental-based approach predicted PED by linear superposition of positron emitter yields from different reference materials [6], and Helmbrecht et al. [7] extended this approach to 3D PED calculations, though they have limitations in predicting projectile fragment distributions. Another analytical approach, suggested by Hofmann et al. [8], modified a filtering approach originally designed for proton therapy [9] to adapt it for carbon ion therapy. This method, however, was primarily investigated in 1D or quasi-3D scenarios, and exhibited some inaccuracies especially in vicinity of air cavities.

In this work, we addressed the issues found in [8] by introducing finer modelling functions, considering more positron emitters (PE) and suggesting a mapping process for handling longitudinal heterogeneity effects for projectile PED, and extended the computational framework to 3D PED prediction. MC simulations for generation of the initial reference databases are used. A dedicated framework based on a pencil beam algorithm (PBA) to predict 3D PED is proposed, which includes improvements over the 1D PED prediction approach, the parametrization of PED's lateral properties, and methods for handling both lateral and longitudinal heterogeneity effects. In-silico validation studies using different homogeneous and inhomogeneous phantoms, as well as clinical CT data are conducted, with results discussed in section 3.

2 Materials and Methods

2.1 Generation of reference databases by Monte Carlo simulations

The MC method, considered as a golden standard for simulating radiation interactions with matter, is widely used in imaging and radiotherapy. The MC toolkit Geant4 [10] version 10.07.p04 is used in this work to simulate the production of PE by carbon ion irradiation. The physics list QGSP_BIC is selected for the modelling of physical processes.

MC simulations were preformed in a cylindrical phantom of a homogeneous reference material. The so-called reference material consists of six elements (H, C, N, O, P and Ca) which contribute to the main production of PE in human tissue, ans its properties are shown in table 1. In the simulations, mono-energetic and mono-directional carbon ion pencil beams with initial lateral sigma σ_0 of 3 mm hit the surface center perpendicularly, and five isotopes ^{11}C, ^{10}C, ^{15}O, ^{13}N and ^{14}O were considered and scored. Among the five PE considered, ^{11}C and ^{10}C were further categorized into projectile and target contributions based on their kinetic energy. The simulations were performed for 36 initial energies

Material	Ref.	Brain	Bone	Water
Mass density (g·cm^{-3})	1.54	1.04	1.85	1.00
Electron density (10^{20} cm^{-3})	4.71	3.45	5.91	3.34
Mass abundances				
Hydrogen (H)	2%	11%	6%	11%
Carbon (C)	31%	15%	28%	-
Nitrogen (N)	26%	2%	3%	-
Oxygen (O)	38%	71%	41%	89%
Phosphorus (P)	1%	-	7%	-
Calcium (Ca)	3%	-	15%	-

Table 1: Properties of the materials considered for the phantoms and for the reference material (Ref.)

from 80 to 430 MeV/n with increments of 10 MeV/n and for a total 5×10^8 carbon ion primaries. The number of PE and dose were sampled with cylindrical scoring at 0.2 mm radial binning and at 0.5 mm in depth.

2.2 Framework for Calculating 3D PED

A dedicated framework based on a pencil beam algorithm (PBA) was developed to construct 3D PED. This framework required a similar input structure to the PBA[11] used for 3D dose computation, including the laterally integrated depth positron emitter distribution (IDPED), lateral distribution, and material and beam information.

IDPED was derived using analytical functions combined with a filtering approach on the laterally integrated depth dose (IDD) at corresponding energies in the reference material. The parameters of these analytical functions were predefined using IDD and IDPED data from MC simulations in the reference material. In practical applications, IDD in the reference material is reconstructed from the 3D dose distribution per ion beam spot. Following this, IDPED is predicted by applying the filtering approach to the reconstructed IDD, and then utilizing the analytical functions.

The lateral properties of PED were parameterized using Gaussian distributions. For projectile PE, a single Gaussian distribution was used to account for the lateral off-axis contributions. For target PE, a double Gaussian parametrization method was adopted. The resulting sigma parameters were corrected by quadratic subtraction of the initial beam σ_0.

Predicting PED in heterogeneous media required adaptations for longitudinal and lateral heterogeneity effects. For the longitudinal heterogeneity, one needs to consider residual ranges across different materials and the contributions of different reaction channels to the total PE density. The depth in heterogeneous media was converted into a reference-material-equivalent depth for prediction in reference space. As the production of PE is influenced by the nuclear density of target nuclei and the characteristics of the tissue, a local factor $g_i(z)$[9] is employed to adjust the weight of different production channels for target PED. For projectile PED, a mapping process was conceptualized, associating each voxel

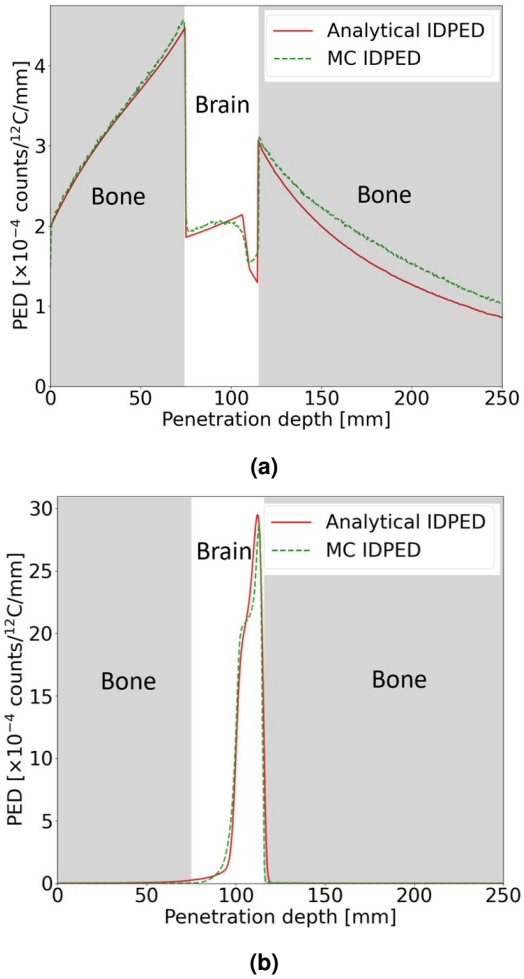
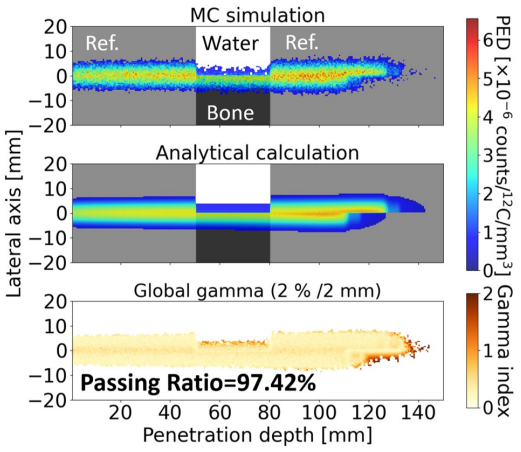


Figure 1: IDPED for (a) $^{11}C_{target}$ and (b) $^{11}C_{proj}$ of a ^{12}C ion beam of 300 MeV/n in a heterogeneous slab phantom by analytical method and Geant4 MC simulation. Slabs were created with brain and compact bone materials shown in table 1.

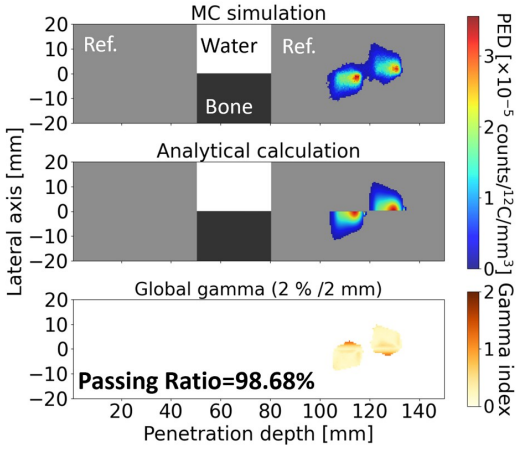
where projectile PE eventually stop (stopping voxels) with the voxels that originally produce them (producing voxels). The material effect on PE's production at each producing voxel was counted by the same $g_i(z)$ factor, then the produced PE at producing voxels was arranged into stopping voxels based on the established mapping relationship. For the lateral heterogeneity, the ray-casting method [12] is used.

2.3 In-silico Validation Studies

Validation studies were conducted using MC simulations and our analytical approach's calculations with several phantom configurations, including slab like inhomogeneous phantoms and patient based on CT data from the Heidelberg Ion-Beam Therapy Center (HIT) [13]. Mono-energetic and mono-directional ion pencil beams with initial sigma σ_0 of 3 mm was adopted for the cases with the slab phantoms, and several spots extracted from actual patient treatment plans from HIT were simulated.



(a)



(b)

Figure 2: 2D PED for (a) $^{11}C_{target}$ and (b) $^{11}C_{proj}$ of a ^{12}C ion beam of 300 MeV/n with FWHM of 7.1 mm in a laterally heterogeneous phantom by analytical method and MC simulation.

3 Results

Figure 1 shows the slab phantom arranged in a bone-brain-bone composition. We used the two materials with significant elemental differences to validate our method in addressing the impact of longitudinal heterogeneity on the distribution of target and projectile PE. Figure 1(a) and (b) displays the IDPED results for $^{11}C_{target}$ and $^{11}C_{proj}$ by analytical method and MC simulation. These profiles' trends and heights in different regions are in reasonable agreement with the simulation results. To quantify their differences, we calculated their normalized root-mean-square error (NRMSE) as 0.0534 and 0.0313, respectively, which were normalized by the difference between maximum and minimum. The brain region contains the distal fall-off positions of both the $^{11}C_{target}$ and $^{11}C_{proj}$ profiles, where the analytical results and simulation results show sufficiently close falling positions. The noticeable difference in Figure 1(a) is in the tail part, mainly due to deviations in the amplitude of the tail modelling function when initially constructing the IDPED in the reference space compared to the simulation results.

Figure 2 presents the two-dimensional results for $^{11}C_{target}$ and $^{11}C_{proj}$ in the presence of lateral heterogeneity. Slabs composed of water and bone are inserted vertically in the middle of the phantom, one on top of the other. The discrepancies between analytical and simulation results were evaluated through global gamma analysis. For the 2%/2 mm criteria, the passing ratios were 97.42% and 98.68%, respectively. The satisfactory passing rates demonstrate the effectiveness of using the ray-casting method to address longitudinal heterogeneity. However, there are noticeable differences in the longitudinal distribution in Figure 2(b): the analytical result shows clear boundaries due to lateral heterogeneity, while the simulation results do not. This is because particles undergo scattering, causing blurring at the boundaries, whereas the ray-casting method can only consider heterogeneity in a straight-line propagation manner, making the boundaries clearer [12]. However, for heavy ions, their scattering is relatively minor, so the errors introduced by the ray-casting method are acceptable.

Figure 3 shows the total distribution of all the five PE produced by one ion pencil beam of 192.13 MeV/n extracted from an actual treatment plan from HIT [13] calculated by MC simulation and our analytical approach. The ion beam's delivery was implemented using the phase space of [14]. The discrepancies between analytical and simulation results were evaluated through global gamma analysis. For the 2%/2 mm criteria, the passing rate is 99.91%. For the 1%/1 mm criteria, the passing rate is 91.75%.

4 Discussion and Conclusion

This work improves upon and extends [8] to three-dimensional PED calculation. Hofmann et al. [8] primarily explored the possibility of predicting 1D PED by combining a filtering method [9] with analytical modelling. Our work first improves upon this method, considering a more comprehensive range of energies and types of positron emitters, along with proposing more detailed one-dimensional analytical modelling functions. Additionally, we extended the improved one-dimensional method to three dimensions based on the pencil beam algorithm. The lateral properties of PED are obtained through Gaussian fitting. This ensures that the lateral properties of PED considered both the primary and secondary fluence.

The feasibility of our approach was demonstrated by our in-silico studies. In terms of clinical interest, we also demonstrated the predicted results of a beam spot's PED using patient data, which verified the feasibility of our method under more practical conditions.

Our next step will involve MC simulations and analytical predictions based on complete treatment plans, coupled with measured PET data, to facilitate a more comprehensive validation of our method.

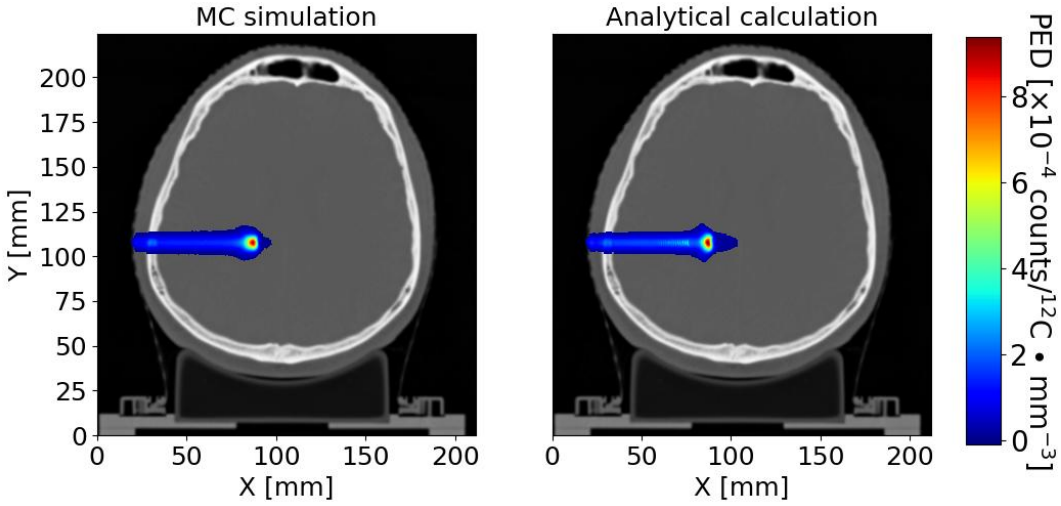


Figure 3: 2D distribution for all the five PE (^{11}C, ^{10}C, ^{15}O, ^{13}N and ^{14}O) for a ^{12}C ion beam of 192.13 MeV/n extracted from an actual treatment plan from HIT, as calculated by the analytical method and MC simulation.

5 Acknowledgments

We would like to thank Thomas Tessonnier from HIT for support on several questions on the phase space. This work is financially supported by the China Scholarship Council (CSC).

References

[1] T. D. Malouff, A. Mahajan, S. Krishnan, et al. “Carbon ion therapy: a modern review of an emerging technology”. *Frontiers in oncology* 10 (2020), p. 82.

[2] K. Parodi, T. Bortfeld, and T. Haberer. “Comparison between in-beam and offline positron emission tomography imaging of proton and carbon ion therapeutic irradiation at synchrotron-and cyclotron-based facilities”. *International Journal of Radiation Oncology* Biology* Physics* 71.3 (2008), pp. 945–956.

[3] F. Pennazio, G. Battistoni, M. G. Bisogni, et al. “Carbon ions beam therapy monitoring with the INSIDE in-beam PET”. *Physics in Medicine & Biology* 63.14 (2018), p. 145018.

[4] K. Parodi. “PET monitoring of hadrontherapy”. *Nuclear Medicine Review* 15.C (2012), pp. 37–42.

[5] A. C. Kraan. “Range verification methods in particle therapy: underlying physics and Monte Carlo modeling”. *Frontiers in oncology* 5 (2015), p. 150.

[6] M. Priegnitz, F. Fiedler, D. Kunath, et al. “An Experiment-Based Approach for Predicting Positron Emitter Distributions Produced During Therapeutic Ion Irradiation”. *IEEE Transactions on Nuclear Science* 59.1 (2012), pp. 77–87.

[7] S. Helmbrecht, M. Priegnitz, W. Enghardt, et al. “Application of a Yield Approach for the Prediction of Positron Emitter Distributions Produced During Therapeutic Carbon-Ion Beam Irradiation”. *IEEE Transactions on Nuclear Science* 63.1 (2016), pp. 61–69.

[8] T. Hofmann, A. Fochi, K. Parodi, et al. “Prediction of positron emitter distributions for range monitoring in carbon ion therapy: an analytical approach”. *Physics in Medicine & Biology* 64.10 (2019), p. 105022.

[9] K. Parodi and T. Bortfeld. “A filtering approach based on Gaussian-powerlaw convolutions for local PET verification of proton radiotherapy”. *Phys. Med. Biol.* 51.8 (2006), pp. 7089–7106.

[10] S. Agostinelli, J. Allison, K. a. Amako, et al. “GEANT4—a simulation toolkit”. *Nuclear instruments and methods in physics research section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 506.3 (2003), pp. 250–303.

[11] M. Soukup, M. Fippel, and M. Alber. “A pencil beam algorithm for intensity modulated proton therapy derived from Monte Carlo simulations”. *Physics in Medicine & Biology* 50.21 (2005), p. 5089.

[12] B. Schaffner, E. Pedroni, and A. Lomax. “Dose calculation models for proton treatment planning using a dynamic beam delivery system: an attempt to include density heterogeneity effects in the analytical dose calculation”. *Physics in Medicine & Biology* 44.1 (1999), p. 27.

[13] J. Bauer, F. Sommerer, A. Mairani, et al. “Integration and evaluation of automated Monte Carlo simulations in the clinical practice of scanned proton and carbon ion beam therapy”. *Physics in Medicine Biology* 59.16 (July 2014), p. 4635.

[14] T. Tessonnier, T. Marcelos, A. Mairani, et al. “Phase Space Generation for Proton and Carbon Ion Beams for External Users’ Applications at the Heidelberg Ion Therapy Center”. *Frontiers in Oncology* 5 (2016).

Simultaneous optimization of multiple plans within one treatment course for temporally feathered radiation therapy

Jenny Bertholet[1], Julius Arnold[1,2], Chengchen Zhu[1], Gian Guyer[1], Silvan Mueller[1], Marco FM Stampanoni[3], Peter Manser[1], and Michael K Fix[1]

[1]Division of Medical Radiation Physics and Department of Radiation Oncology, Inselspital, Bern University Hospital, and University of Bern, Bern, Switzerland, [2]Department of Physics, ETH Zürich, Zürich, Switzerland, [3]Institut of Biomedical Engineering, ETH Zürich and PSI, Villigen, Switzerland

Abstract Temporally feathered radiation therapy (TFRT) is a new technique where several iso-curative sub-plans deliver an alternance of high and low doses to the same number of feathered organs-at-risks (OARs) to exploit the dynamic recovery process of healthy tissues. The goal is that each OAR receives, in alternance, one high dose and four low doses where enhanced repair can take place to offset the sub-lethal damage caused during the high dose fraction. The aim of this work is to enable simultaneous optimization of multiple sub-plans within one treatment course for TFRT planning. A clear high to low dose pattern was achieved for an academic case according to different TFRT planning objective variations (soft, medium, and hard). Feathering was also achieved for a clinically motivated head and neck case with high to low dose patterns and/or reduction in total mean dose for the ipsilateral parotid and submandibular glands, pharynx, oral cavity, and larynx.

1 Introduction

In radiation therapy, the same treatment plan is generally delivered five days per week for several weeks with spatial distribution of the dose being a crucial aspect of treatment plan quality. Highly conformal dose distributions can be obtained using inverse planning for intensity-modulated treatment techniques. Nevertheless, we may be reaching a limit in our capability to shape the spatial dose distribution. New approaches have been proposed to optimize the temporal dimension through which radiation therapy is delivered: different dose distributions are delivered on different treatment days. For example, in temporally feathered radiation therapy (TFRT), five sub-plans are created to deliver a homogeneous dose to the target volume for each sub-plan (iso-curative) with an alternance of high and low doses to five feathered organs-at-risks (OARs) to exploit the dynamic recovery process of healthy tissues [1]. The goal is that each feathered OAR receives each week one high dose and four low doses where enhanced repair can take place to offset the sub-lethal damage caused during the high dose fraction. A potential additional benefit of TFRT is that, for the same number of beams/arcs delivered per fraction, the total physical dose in the composite TFRT plan may be lower than in a standard non-feathered plan because there are more degrees-of-freedom to solve the optimization problem when five different plans are optimized instead of one. However, currently, all sub-plans have to be created independently and then combined for evaluation which is cumbersome and may result in sub-optimal solutions.

Here, we implement simultaneous optimization of multiple plans within one treatment course in an in-house optimizer. We simultaneously optimize five TFRT sub-plans for one academic and one clinically motivated head and neck case.

2 Materials and Methods

Our in-house optimizer uses Monte Carlo calculated beamlet doses and direct aperture optimization (DAO) to optimize the apertures shape and monitor unit (MU)-weights for a variety of treatment techniques, deliverable on standard linacs [2,3]. The optimization's goal is to minimize the cost-function consisting of upper and lower dose-volume objectives, generalized equivalent uniform dose objectives (gEUD) [4], and normal tissue objectives.

2.1. Multi-optimization workflow

In this work, the in-house optimizer was modified by including a MultiManager class that simultaneously handles multiple sub-plans, each with their set of apertures, and dose distribution for a given number of fractions. Each sub-plan has a cost-function consisting of objectives on the sub-plan dose distribution and "total objectives" evaluated on the composite dose distribution (the sum of all sub-plans dose distributions).

As illustrated in figure 1, in each DAO iteration, all sub-plans are cloned (clones 1-5). The aperture with the steepest gradient on the cost-function is added and the resulting sub-plan cost-function is updated for each clone in parallel (F_i', i = 1-5). The total cost-function, F_{tot}, is calculated based on the updated cost-function for the cloned sub-plan and the current cost-function for the other sub-plans. The clone that offers the best total objective function is selected and the corresponding sub-plan is updated; the other clones are deleted.

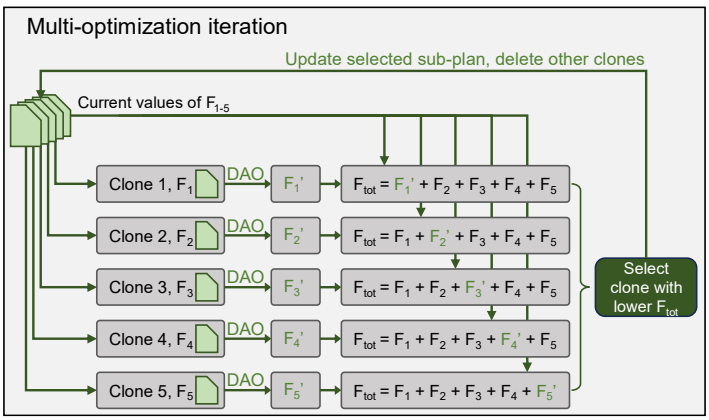
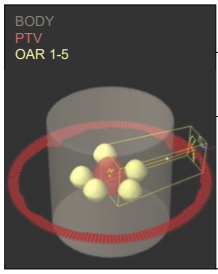


Figure 1: In each multi-optimization iteration, apertures are added/modified in the sub-plan that offers the best total cost-function F_{tot}. F_{1-5} consists of objectives on the sub-plan dose distribution as well as on the total dose distribution.

2.2. TFRT planning

An academic case (figure 2) was designed to test the multi-optimization workflow for TFRT planning. The phantom is cylindrical with a planning target volume (PTV) surrounded by five spheroid OARs, arranged asymmetrically.



Structure	Dimensions [cm]	Center of mass position [cm]			Distance to PTV [cm]
		x	y	z	
Body	25.0 Ø	0.0	0.0	0.0	
PTV	7.0 Ø axial, 8.6 SI	1.6	-0.9	0.0	
OAR1	5.0 Ø	-2.0	-5.6	-1.2	0.2
OAR2	5.0 Ø	5.0	-5.7	0.4	-0.1 (overlapping)
OAR3	5.0 Ø	7.5	0.7	-0.3	0.1
OAR4	5.0 Ø	1.6	5.0	-0.7	0.0 (abutting)
OAR5	5.0 Ø	-3.9	0.6	1.3	0.0 (abutting)

Figure 2: Academic case description. SI: superior-inferior.

A standard volumetric modulated arc therapy (VMAT) plan was created as reference. Three variations (soft, medium, hard) of TFRT plans were created using VMAT, with the multi-optimizer. Individual sub-plan objectives and total objectives were used as described in table 1. The three TFRT variations have increasing difference in high to low mean dose objectives (gEUD with t = 1). Note that the average of one high and four low dose objectives is the same in all three variations and corresponds to the objectives of the standard plan and the total objectives for the TFRT optimization.

A standard and TFRT plan were also created for a clinically motivated head and neck case on an anthropomorphic phantom (locoregionally advanced oropharyngeal carcinoma with unilateral elective nodal volume [5]) prescribed 50 Gy to 95% of the PTV in 25 x 2 Gy.

Table 1. Mean dose objectives (in % of prescribed dose) to the OARs for the standard plan and three TFRT variations.

	Standard	TFRT variations		
		soft	medium	hard
Objectives for dose from individual sub-plan				
High dose	-	36	44	52
Low dose	-	26	24	22
Objectives for total dose distribution				
All OARs	28	28	28	28

The feathered OARs were the ipsilateral parotid gland, ipsilateral submandibular gland, pharynx, oral cavity and larynx. The average of the high/low mean dose objectives in the subplans equaled the total objectives for the TFRT optimization and the standard plan (table 2).

Table 2. Mean dose objectives (in % of prescribed dose) to the feathered OARs for TFRT and standard plan for the clinically motivated case. PG: parotid gland, SMG: submandibular gland, OC: oral cavity, IL: ipsilateral

	PG R	SMG R	Pharynx	OC	Larynx
High dose	62.0	84.6	90.0	66.0	84.6
Low dose	42.0	64.6	70.0	46.0	64.6
Standard / total TFRT	46.0	68.6	74.0	50.0	68.8

Standard and TFRT plans are compared based on OAR dose in the sub-plans as well as on the total dose distribution.

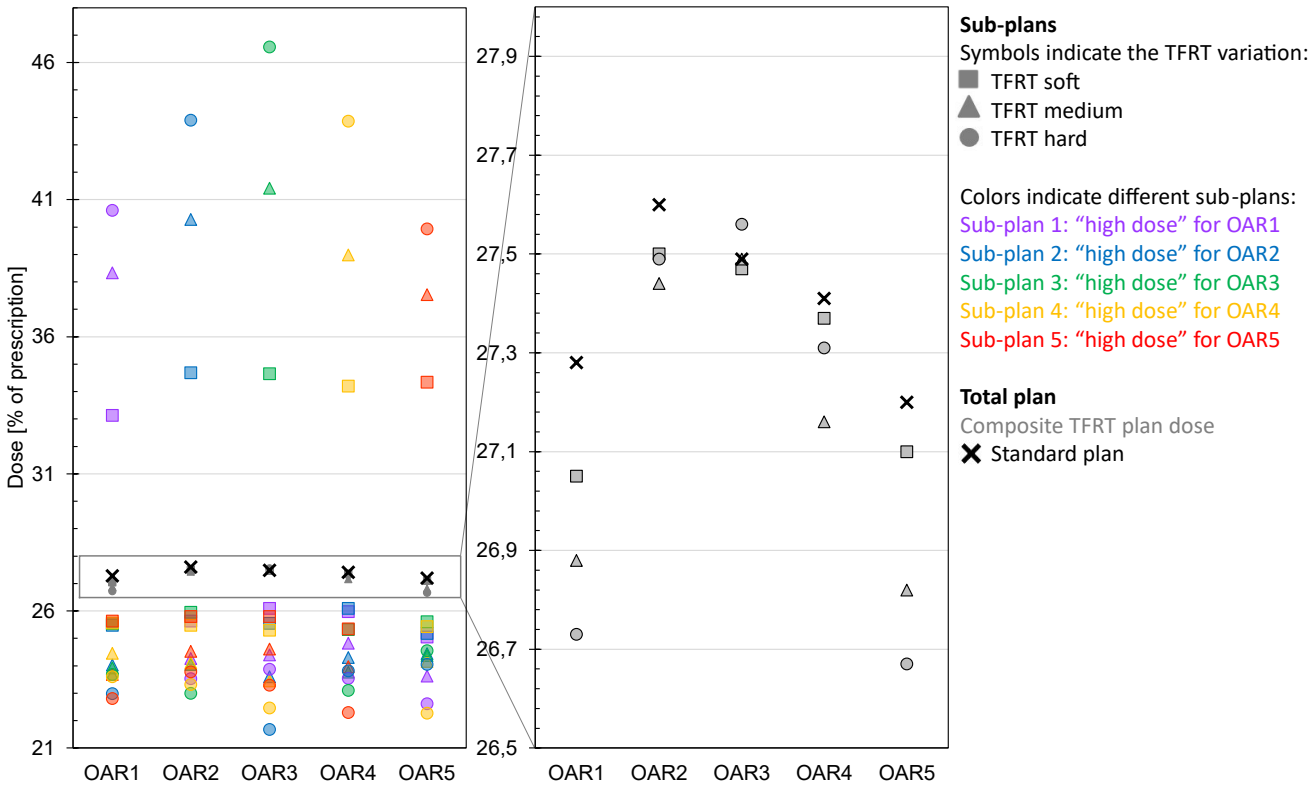


Figure 3: Dose received by the five OARs of the academic case in the three different TFRT variations (different symbols). Each color shows the dose received by all OARs for one sub-plan with high dose to a specific OAR. The right panel shows a zoom on the values for the reference standard plan and the composite TFRT plan.

3 Results

All sub-plans were iso-curative for the academic case. Figure 3 shows the dose received by each OAR in the TFRT sub-plans (colors) and composite plans (grey) as well as in the standard plan (crosses). The high to low dose alternance is visible in the color pattern. The greater differences achieved between high and low doses from the "soft" to the "medium" and to the "hard" TFRT variations are also visible in the symbol pattern. The dose for the composite TFRT plan (grey symbols) is lower than for the standard plan for OARs 1, 2, 4, and 5. For OAR3, the doses are similar in all composite TFRT and in the standard plans.

Figure 4 shows the DVHs for the feathered OARs and the PTV for each TFRT sub-plan (scaled to full prescription),

Mean dose (% of prescribed dose) to the feathered OARs.

	TFRT 1	2	3	4	5	tot	standard
IL PG	51.5	49.1	52.4	49.9	57.3	51.5	53.2
IL SMG	70.7	70.9	71.1	79.4	69.4	71.5	71.7
Pharynx	78.7	76.9	80.6	77.7	78.7	77.8	79.2
OC	56.1	60.3	56.1	54.7	53.4	55.6	55.2
Larynx	74.8	68.7	69.5	69.4	71.3	70.1	70.9

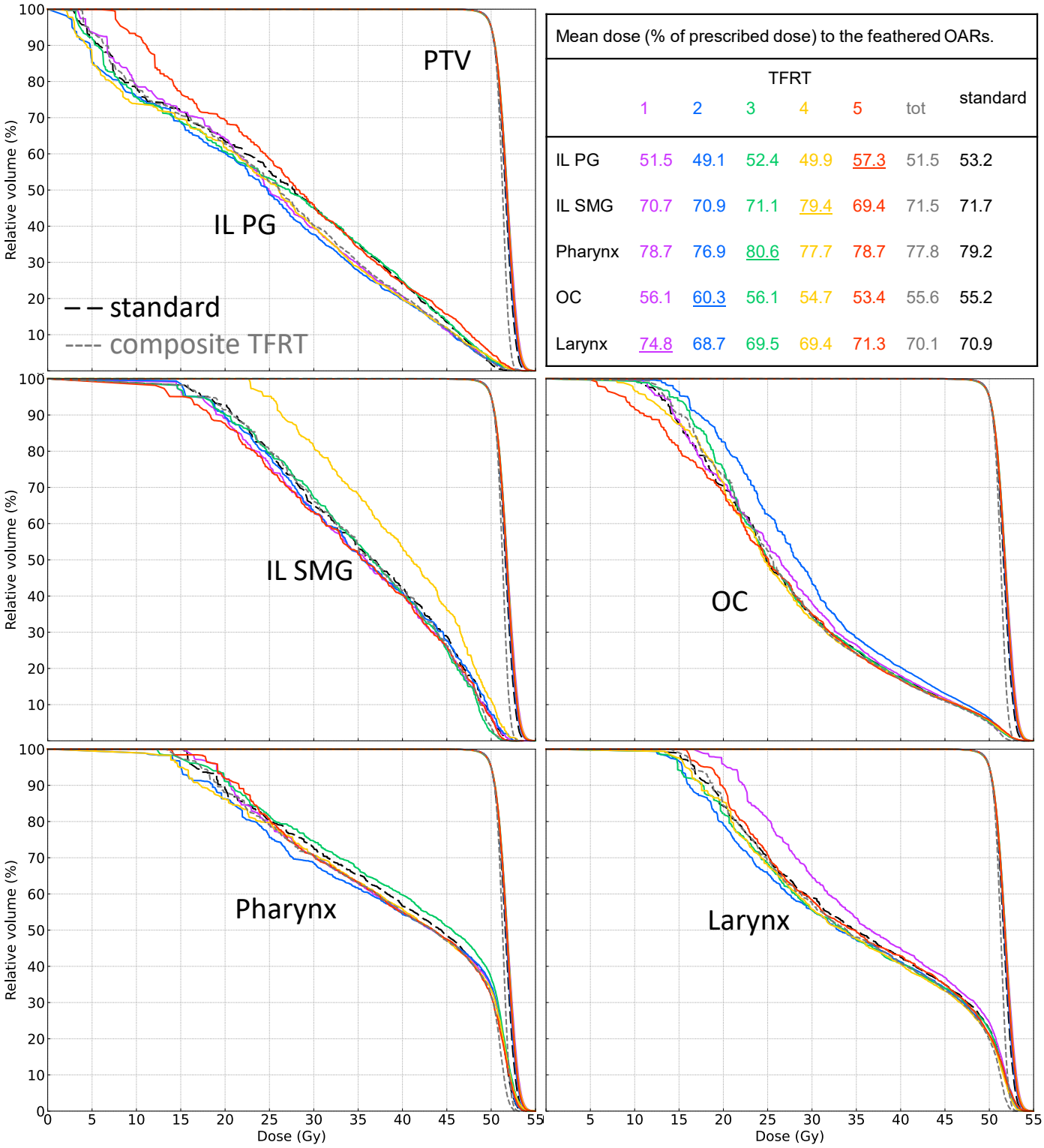


Figure 4: DVHs for the feathered OARs in the clinically motivated plan. The achieved mean dose to the feathered OARs in each sub-plan, in the total TFRT plan and in the standard plan are reported in the table. PG: parotid gland, SMG: submandibular gland, OC: oral cavity, IL: ipsilateral. Each color represents a sub-plan, the thin dashed grey line represents the composite TFRT plan and the black dashed line represents the standard plan.

the composite TFRT plan, and the standard plan for the clinically motivated case. The mean doses to the OARs are indicated in the table with the OAR receiving the high dose in each sub-plan underlined.

The TFRT sub-plans are iso-curative but have slightly less homogeneous PTV coverage than the standard plan, however the composite TFRT plan is the most homogeneous.

The high and low dose pattern is evident for the submandibular gland, oral cavity, and larynx. For these three OARs, the mean dose in the composite TFRT plan is close to that of the standard plan with 35.9 Gy (standard) and 35.7 Gy (TFRT) for the submandibular gland, 27.6 Gy (standard) and 27.8 Gy (TFRT) for the oral cavity, and 35.5 Gy (standard) and 35.1 Gy (TFRT) for the larynx.

The high and low dose pattern is less marked for the parotid gland and the pharynx although still noticeable. However, the mean dose in the composite TFRT is lower than in the standard plan with 26.6 Gy (standard) and 25.8 Gy (TFRT) for the parotid gland, and 39.6 Gy (standard) and 38.9 Gy (TFRT) for the pharynx.

4 Discussion

We have successfully implemented DAO-based simultaneous optimization of multiple plans within one treatment course in our in-house optimizer. The different plans are optimized using standard upper, lower, generalized equivalent uniform dose, and normal tissue objectives, but these can be applied (and combined) to the individual sub-plan dose distribution and to the total dose distribution. In each DAO iteration, an aperture is added to the sub-plan where most benefit over the overall objective function is expected.

We have characterized the high to low dose pattern and the resulting composite TFRT dose for an academic case with a PTV surrounded by five OARs. The high to low dose alternation was well correlated with the differences in high and low dose objectives (soft, medium, and hard TFRT variations) applied to the sub-plans. This alternation is expected to have a biological benefit on healthy tissue repair [1].

In addition, the mean dose in the composite TFRT plan was lower than in the standard plan for four out of the five feathered OARs despite the fact that they had the same total dose objectives. This can be explained by the greater number of degrees-of-freedom for the optimization: for TFRT, five times one VMAT arcs are optimized and the pool of possible apertures is therefore five times greater than for the standard plan with one VMAT arc. This advantage might vanish if the standard plan used 5 arcs or more, however, delivery time would be increased.

The TFRT optimization method was also successfully applied to a clinically motivated case, despite a more complex geometry with substantial PTV/OAR overlap.

Although five OARs and five sub-plans were considered, the multi-optimizer can handle an arbitrary number of sub-plans and number of fractions per sub-plan. Another potential application is to simultaneously optimize the phases of a sequential boost plan.

Future developments will include the design of objectives that consider the difference in mean dose to the OARs between two sub-plans to further optimize the high to low dose alternance.

5 Conclusion

We have successfully implemented simultaneous optimization of multiple plans in our in-house optimizer and demonstrated the use of our method for TFRT planning. We achieved clear high to low dose alternance which are expected to reduce overall OAR toxicity by offsetting the damage of the high dose fractions during the low dose fractions.

Acknowledgements

This work was supported by Varian, a Siemens Healthineers Company and SNSF grant 200021_185366. Dose calculations and optimizations were performed on UBELIX (www.id.unibe.ch/hpc), the high-performance cluster at the University of Bern.

References

[1] López Alfonso JC, Parsai S, Joshi N, Godley A, Shah C, Koyfman SA, et al. Temporally feathered intensity-modulated radiation therapy: A planning technique to reduce normal tissue toxicity. Med Phys 2018;45:3466–74. https://doi.org/10.1002/mp.12988.

[2] Mueller S, Guyer G, Risse T, Tessarini S, Aebersold DM, Stampanoni MFM, et al. A hybrid column generation and simulated annealing algorithm for direct aperture optimization. Phys Med Biol 2022;67:075003. https://doi.org/10.1088/1361-6560/ac58db.

[3] Guyer G, Mueller S, Koechli C, Frei D, Volken W, Bertholet J, et al. Enabling non-isocentric dynamic trajectory radiotherapy by integration of dynamic table translations. Phys Med Biol 2022;67:175003. https://doi.org/10.1088/1361-6560/ac840d.

[4] Niemierko A. A generalized concept of equivalent uniform dose (EUD). Med Phys 1999;26:1100.

[5] Bertholet J, Mackeprang P-H, Mueller S, Guyer G, Loebner HA, Wyss Y, et al. Organ-at-risk sparing with dynamic trajectory radiotherapy for head and neck cancer: comparison with volumetric arc therapy on a publicly available library of cases. Radiation Oncology 2022;17:122. https://doi.org/10.1186/s13014-022-02092-5.

The relationship between treatment plan complexity and dosimetry quality: An analysis of three stereotactic treatment planning challenges

Nicholas Hardcastle[1,2,3], Mathieu Gaudreault[1,2], Alisha Moore[4], Benjamin E Nelms[5], Jordi Saez[6], Victor Hernandez[7]

[1]Physical Sciences, Peter MacCallum Cancer Centre, Melbourne, Australia
[2]Sir Peter MacCallum Department of Oncology, University of Melbourne, Melbourne, Australia
[3]Centre for Medical Radiation Physics, University of Wollongong, Wollongong, Australia
[4]Trans Tasman Radiation Oncology Group, Newcastle, Australia
[5]Canis Lupus LLC, Merrimac, USA
[6]Department of Radiation Oncology, Hospital Clínic de Barcelona, 08036 Barcelona, Spain
[7]Department of Medical Physics, Hospital Sant Joan de Reus, IISPV, 43204 Tarragona, Spain

Abstract Radiation therapy treatment planning challenges provide the radiation therapy community with a standardized means to compare treatment plans from different techniques, software and users. Treatment planning challenges thus provide a unique opportunity to evaluate the impact of treatment plan complexity on treatment plan quality. In this study we compared the treatment plan complexity in three stereotactic treatment planning challenges between different treatment planning systems, and determined the relationship between treatment plan dosimetric quality and complexity. There was a large variation in treatment plan complexity between treatment planning systems, MLC widths and between challenges. Although treatment plan dosimetric score was positively correlated with increasing treatment plan complexity, we have demonstrated this was not a necessity to achieve a high dosimetric score. Thus limiting treatment plan complexity should be considered when optimizing treatment plans, to minimize associated increased uncertainties with highly complex treatment plans.

1 Introduction

Radiation therapy treatment planning challenges are often used in the radiation therapy community to evaluate variations in practice, determine the impact of technology or staff training on treatment plan quality and to service as an education process for a new clinical trials with novel treatment planning scenarios[1-3]. In treatment planning challenges, typically a single CT dataset with structures is provided to all participants, along with a planning guideline. Participants aim to create a treatment plan which maximizes the score based on a priori scoring matrix in which points are awarded according to adherence to and exceeding planning objectives.

Planning challenges often provide treatment plans of high dosimetric quality, which deliver the prescription dose to the target volumes in a highly conformal manner, with very low dose to adjacent critical structures. Such plans however may be highly complex, involving not only complicated beam geometry which may be infeasible to deliver in routine clinical practice, but potentially high complexity (small or complicated beam apertures, large variations in dose rate and gantry speed). This however may reflect clinical practice; in the absence of clear guidelines on minimization of treatment plan delivery complexity, multiple rounds of optimization may result in highly conformal dose distributions with the cost of high complexity [4]. Plan complexity is a critical component of treatment plan quality and there is a wide range of complexity metrics that have been used [5-9]. Despite this, there are limited tools in treatment planning systems (TPS) to measure and optimize complexity [10].

To date, there is sparse data describing the range in treatment plan modulation complexity and how it relates to treatment plan dosimetric quality. Treatment planning challenges provide a unique opportunity to assess treatment plan complexity using standardized treatment plan dosimetric quality metrics from a range of users and vendors. In this work, we evaluate the relationships between treatment plan complexity metrics and dosimetric quality in three stereotactic treatment planning challenges.

2 Materials and Methods

Planning Challenge Data

Three stereotactic treatment planning challenges were selected for analysis covering multi-metastasis stereotactic radiosurgery (SRS), pancreas stereotactic ablative body radiotherapy (SABR) and vertebral SABR. The target and organ at risk (OAR) structures for each challenge, and representative dose distributions are shown in Figure 1. The planning challenges were performed by the Trans Tasman Radiation Oncology Group (TROG) based on TROG clinical trial protocols, and were hosted by Elekta ProKnow Systems [1, 11]. Participation in the planning challenge involved downloading a CT image and associated structure set in DICOM format from the ProKnow website, creating a treatment plan to meet the protocol, and uploading the plan and dose grid in DICOM format. For each challenge, a scoring matrix was determined prior to the start of the plan challenge and was available to all participants at the time of planning. The scoring matrix provided a plan dosimetry score out of 150 based on meeting and exceeding trial dose objectives for target and OAR structures. The DICOM RTPlan files were made available by ProKnow Systems for analysis.

Complexity metric extraction

Only VMAT plans were considered in this study, which constituted 88-90% of plans in each planning challenge. The following complexity metrics were calculated for each VMAT treatment plan submitted to the planning challenges. Complexity metrics were calculated in an in-house code developed in Matlab [9].

- The Edge Metric (EM) [12], which measures the ratio of the X (MLC leaf widths) and Y perimeters to that of the area of the MLC-defined field. In this calculation we used parameters C1 = 0 and C2 = 1.
- The Modulation Index Total (MItotal) [13], which takes into account MLC complexity, gantry speed variability and dose rate variability.

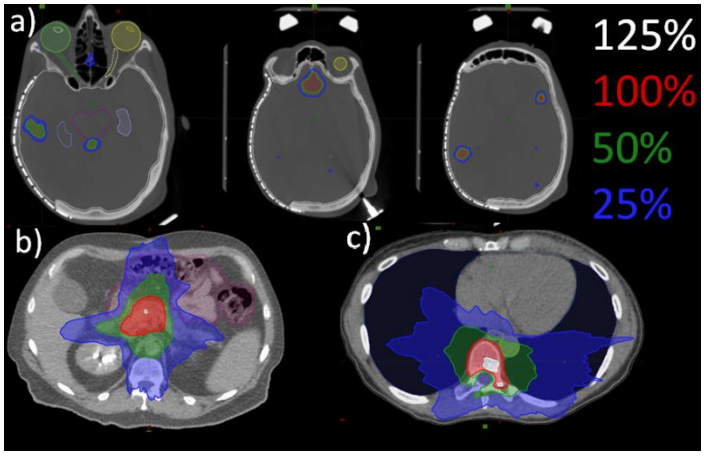


Figure 1: Target volume, critical organs and dosimetry for (a) SRS plan challenge (b) pancreas SABR challenge and (c) vertebral SABR challenge, with example dose distributions in percentage of prescription dose

The treatment plan complexity metric values were compared between different TPS using the two-sample Kolmogorov-Smirnov test. The relationship between treatment plan dosimetric quality, as defined by the plan dosimetry score in each challenge, and the treatment plan complexity, was assessed using the Spearman correlation coefficient r. A p value < 0.05 was used for statistical significance.

3 Results

Figure 2 shows the EM and MItotal for each of the plans in each planning challenge, organized by TPS and MLC width. The p value for the comparison between the distributions from each TPS for all MLCs and 5 mm MLCs only are shown in Table 1. The EM was higher (more complex) for the Eclipse TPS for all three planning challenges; this was driven in part by the high proportion of 2.5 mm MLC widths in the Eclipse treatment plans, however even when only 5 mm MLCs were considered this held for both the Pancreas and Vertebra challenges. Conversely MItotal was higher with Monaco TPS compared with the other TPS for the Pancreas and Vertebra challenges, but only higher than Eclipse for the SRS challenge. This was driven mainly by increased gantry speed and dose rate modulation in the Monaco treatment plans.

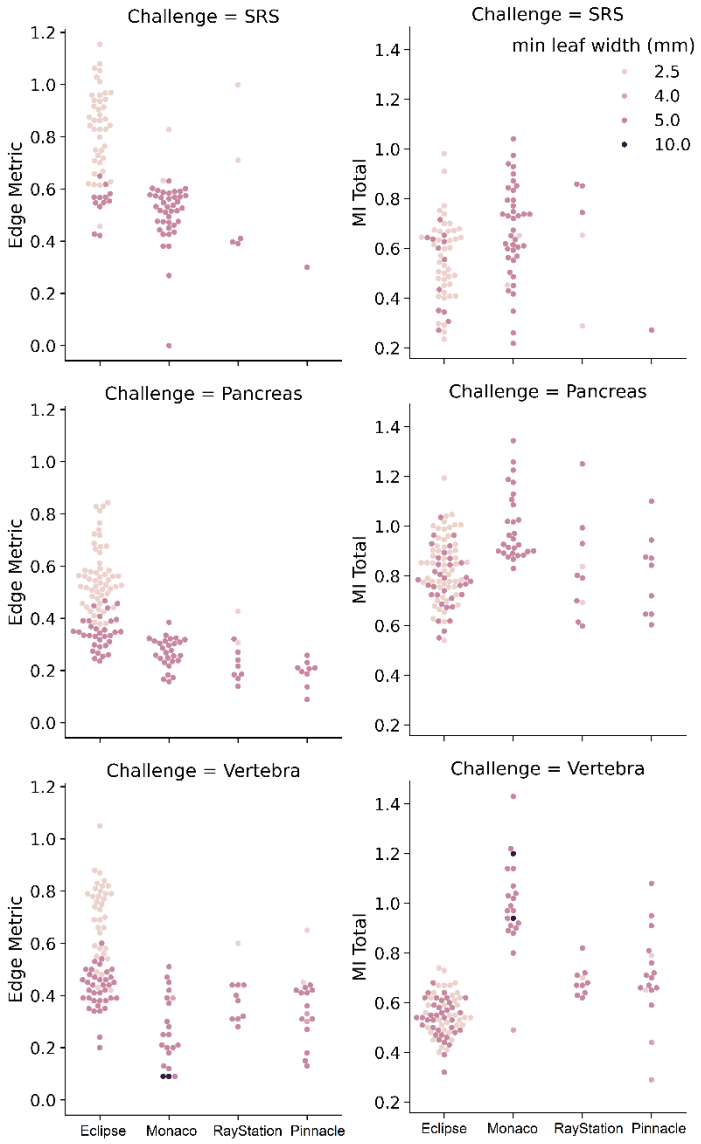


Figure 2: Edge metric and MItotal for each of the plan challenges, categorized by treatment planning system and colored by MLC leaf width.

Figure 3 shows the total plan score plotted as a function of EM and MItotal, organized by TPS. There was a positive correlation between the total plan score and the EM for all three plan challenges (r = 0.48, p < 0.005 for the SRS challenge, r = 0.38, p < 0.005 for the Pancreas challenge and r = 0.59, p < 0.005 for the Vertebra challenge). It should be noted however that although a high EM typically resulted in a higher total plan score, this was not a necessity; there were high scoring plans in all three challenges that did not have high EM. There was no correlation between the total plan score and MItotal (r = -0.12, p = 213 for the SRS

challenge, r = -0.13, r = -0.05, p = 0.561 for the Pancreas challenge and p = 0.164 for the Vertebra challenge).

Table 1: p values from the two-sample Kolmogorov-Smirnov test comparing complexity distributions between treatment planning systems. Those in bold are < 0.05.

		SRS		Pancreas		Vertebra	
Metric	Comparison	All MLCs	5 mm MLCs	All MLCs	5 mm MLCs	All MLCs	5 mm MLC
Edge Metric	Eclipse vs Monaco:	**0.0000**	0.1468	**0.0000**	**0.0000**	**0.0000**	**0.0001**
	Eclipse vs RayStation:	**0.0431**	**0.0044**	**0.0000**	**0.0009**	**0.0026**	0.0888
	Eclipse vs Pinnacle3:	**0.0370**	0.1538	**0.0000**	**0.0000**	**0.0000**	**0.0355**
	Monaco vs RayStation:	0.1411	**0.0057**	0.2449	0.0811	**0.0112**	**0.0413**
	Monaco vs Pinnacle3:	0.1395	0.1463	**0.0008**	**0.0008**	0.0790	0.2203
	RayStation vs Pinnacle3:	0.3333	0.5000	0.2942	0.8167	0.7898	0.5962
MItotal	Eclipse vs Monaco:	**0.0014**	**0.0126**	**0.0000**	**0.0000**	**0.0000**	**0.0000**
	Eclipse vs RayStation:	0.0774	**0.0044**	0.6608	0.5851	**0.0000**	**0.0000**
	Eclipse vs Pinnacle3:	0.1481	0.3077	0.5931	0.7608	**0.0000**	**0.0000**
	Monaco vs RayStation:	0.7264	0.1144	**0.0011**	**0.0084**	**0.0000**	**0.0000**
	Monaco vs Pinnacle3:	0.1395	0.1463	**0.0010**	**0.0010**	**0.0000**	**0.0003**
	RayStation vs Pinnacle3:	0.3333	0.5000	0.8343	0.9895	0.7013	0.6831

4 Discussion

We have evaluated the relationship between treatment plan modulation complexity in three stereotactic treatment plan challenges. We observed that different TPSs control plan complexity in different ways or to different degrees, as demonstrated previously [9]. This includes increased MLC or aperture complexity, compared with increased modulation of the dose rate and gantry speed.

Although aperture complexity was correlated with high plan score in all three plan challenges, this was demonstrated to not be a necessity. Increasing plan complexity may only result in a modest increase in the score, so this dosimetric benefit must be carefully balanced against the pitfalls of increased complexity. These include increased uncertainty in dose calculation and delivery accuracy[13-18].

The critical result demonstrated here is that there were many plans which were obtained with lower plan complexity that exhibited high total scores. This highlights that often a TPS may introduce unnecessary plan complexity in the optimization process. It is important to control treatment plan complexity via any available tool in treatment plan optimization [10, 12]. It should be cautioned however that plan complexity is not prioritized over dosimetry quality; some level of complexity is expected and required to ensure a high-quality treatment plan, particularly when there is challenging spatial relationships between target volumes and adjacent critical structures. It remains a challenge to determine the optimal tradeoff between these two components for routine treatment plan optimization.

As such, it is important to continue to report complexity metrics in treatment planning studies and challenges, and to enable comparison between cohorts to enable determination of ranges of expected complexity for different tumour sites and fractionation regimes. This may be challenging for some metrics due to interpretation of complexity computation, and some level of standardization is required. Further, it is critical to continue to investigate complexity metrics to determine real-world impacts on deliverability/robustness and treatment efficiency to determine which complexity metrics are pertinent to patient outcomes.

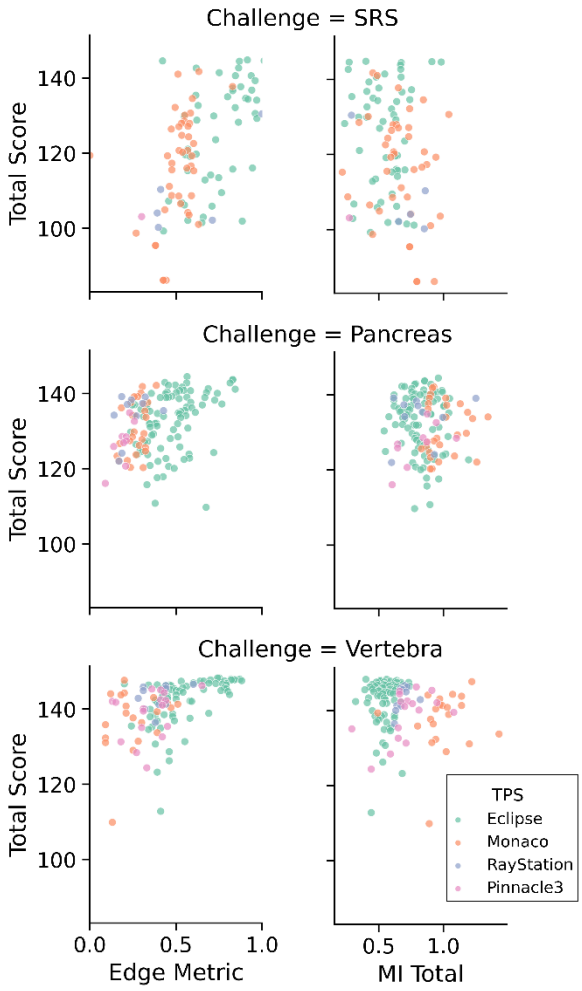


Figure 3: Total plan score as a function of Edge Metric and MItotal, colored by treatment planning system.

5 Conclusion

In three stereotactic treatment planning challenges, the treatment plan complexity varied between TPS. Treatment plan aperture complexity was associated with improved dosimetric plan quality. Treatment plan complexity taking into account aperture, gantry speed and dose rate modulation however was not associated with improved dosimetric plan quality. Further, high treatment plan dosimetric quality could be achieved with treatment plans with low complexity. This suggests that complexity should be limited in treatment planning for stereotactic radiation therapy to minimize the associated increased uncertainty with high complexity, for which appropriate software tools and standardization are warranted.

References

1. Hardcastle, N., F. Bignell, B. Nelms, S. Siva, A. Kneebone, L. Lao, et al., *The challenge of planning vertebral body SBRT: Optimizing target volume coverage.* Med Dosim, 2020. **45**(3): p. 302-307.
2. Moghanaki, D., B. Slotman, A. Swaminath, B. Nelms, and B. Wang, *Assessing the Variability and Quality of Lung Stereotactic Radiation Therapy Treatment Plans Using a Web-Based Crowdsourcing Platform.* Pract Radiat Oncol, 2020. **10**(3): p. e118-e127.
3. Nelms, B.E., G. Robinson, J. Markham, K. Velasco, S. Boyd, S. Narayan, et al., *Variation in external beam treatment plan quality: An inter-institutional study of planners and planning systems.* Pract Radiat Oncol, 2012. **2**(4): p. 296-305.
4. Kaplan, L.P., L. Placidi, A. Back, R. Canters, M. Hussein, A. Vaniqui, et al., *Plan quality assessment in clinical practice: Results of the 2020 ESTRO survey on plan complexity and robustness.* Radiother Oncol, 2022. **173**: p. 254-261.
5. Antoine, M., F. Ralite, C. Soustiel, T. Marsac, P. Sargos, A. Cugny, et al., *Use of metrics to quantify IMRT and VMAT treatment plan complexity: A systematic review and perspectives.* Phys Med, 2019. **64**: p. 98-108.
6. Chiavassa, S., I. Bessieres, M. Edouard, M. Mathot, and A. Moignier, *Complexity metrics for IMRT and VMAT plans: a review of current literature and applications.* Br J Radiol, 2019. **92**(1102): p. 20190270.
7. Hansen, C.R., M. Hussein, U. Bernchou, R. Zukauskaite, and D. Thwaites, *Plan quality in radiotherapy treatment planning - Review of the factors and challenges.* J Med Imaging Radiat Oncol, 2022. **66**(2): p. 267-278.
8. Hernandez, V., C.R. Hansen, L. Widesott, A. Back, R. Canters, M. Fusella, et al., *What is plan quality in radiotherapy? The importance of evaluating dose metrics, complexity, and robustness of treatment plans.* Radiother Oncol, 2020. **153**: p. 26-33.
9. Hernandez, V., J. Saez, M. Pasler, D. Jurado-Bruggeman, and N. Jornet, *Comparison of complexity metrics for multi-institutional evaluations of treatment plans in radiotherapy.* Phys Imaging Radiat Oncol, 2018. **5**: p. 37-43.
10. Kamperis, E., C. Kodona, K. Hatziioannou, and V. Giannouzakos, *Complexity in Radiation Therapy: It's Complicated.* Int J Radiat Oncol Biol Phys, 2020. **106**(1): p. 182-184.
11. Oar, A., M. Lee, H. Le, K. Wilson, C. Aiken, L. Chantrill, et al., *AGITG MASTERPLAN: a randomised phase II study of modified FOLFIRINOX alone or in combination with stereotactic body radiotherapy for patients with high-risk and locally advanced pancreatic cancer.* BMC Cancer, 2021. **21**(1): p. 936.
12. Younge, K.C., M.M. Matuszak, J.M. Moran, D.L. McShan, B.A. Fraass, and D.A. Roberts, *Penalization of aperture complexity in inversely planned volumetric modulated arc therapy.* Med Phys, 2012. **39**(11): p. 7160-70.
13. Park, J.M., S.Y. Park, H. Kim, J.H. Kim, J. Carlson, and S.J. Ye, *Modulation indices for volumetric modulated arc therapy.* Phys Med Biol, 2014. **59**(23): p. 7315-40.
14. May, L., N. Hardcastle, V. Hernandez, J. Saez, A. Rosenfeld, and J. Poder, *Multi-institutional investigation into the robustness of intra-cranial multi-target stereotactic radiosurgery plans to delivery errors.* Med Phys, 2023.
15. Crowe, S.B., T. Kairn, J. Kenny, R.T. Knight, B. Hill, C.M. Langton, et al., *Treatment plan complexity metrics for predicting IMRT pre-treatment quality assurance results.* Australas Phys Eng Sci Med, 2014. **37**(3): p. 475-82.
16. Kairn, T., S.B. Crowe, J. Kenny, R.T. Knight, and J.V. Trapp, *Predicting the likelihood of QA failure using treatment plan accuracy metrics.* Journal of Physics: Conference Series, 2014. **489**(1): p. 012051.
17. Younge, K.C., D. Roberts, L.A. Janes, C. Anderson, J.M. Moran, and M.M. Matuszak, *Predicting deliverability of volumetric-modulated arc therapy (VMAT) plans using aperture complexity analysis.* J Appl Clin Med Phys, 2016. **17**(4): p. 124-131.
18. Park, J.M., J.I. Kim, and S.Y. Park, *Modulation indices and plan delivery accuracy of volumetric modulated arc therapy.* J Appl Clin Med Phys, 2019. **20**(6): p. 12-22.

Addressing the Gap in Oncologist Training: Generating Clinically Realistic, Suboptimal Dose Distributions through Geometry Aware Convolutions

Skylar S. Gay[1,2], Mary P. Gronberg[3], Raymond Mumme[1], Christine Chung[1], Meena Khan[1], Chelsea Pinnix[4], Sanjay Shete[1], Brent Parker[1,5], Tucker J. Netherton[1], Carlos E. Cardenas[5], and Laurence E. Court[1]

[1]Department of Radiation Physics, The University of Texas MD Anderson Cancer Center, Houston, USA
[2]The University of Texas MD Anderson Cancer Center UTHealth Houston Graduate School of Biomedical Sciences, Houston, USA
[3]Department of Radiation Oncology, UT Southwestern Medical Center, Dallas, USA
[4]Department of Radiation Oncology, The University of Texas MD Anderson Cancer Center, Houston, USA
[5]Division of Physics and Engineering, The University of Texas Medical Branch at Galveston, Galveston, USA
[5]Department of Radiation Oncology, The University of Alabama at Birmingham, Birmingham, USA

Abstract Radiation oncologists are responsible for determining the clinical acceptability of radiotherapy treatment plans. Their training typically revolves around routine clinical practice, observing the approval of patient plans. This approach depends on a sufficiently rich patient population (which is difficult for smaller clinics and less common cancer types) and is example-starved as the trainee typically only sees a single plan example for each patient. To address this, we have developed techniques to rapidly create clinically realistic plans that are suboptimal in a controllable way. These plans can then be used to train oncologists to understand what dose distributions are possible, and how to spot plans that could be improved. Here we first use a deep learning-based dose prediction to create a high-quality treatment plan (trained using only high-quality plans). We then increase dose by controllable amounts around selectable organs-at-risk. This approach is shown to generate clinically realistic, but suboptimal plans, as scored by experienced dosimetrists.

1 Introduction

Radiotherapy plan quality and radiation oncologist experience directly correlate to patient outcome and quality of life[1–4]. Therefore, it is vital that the quality of the final approved plan be high – that is, the prescription to the target be met while minimizing dose delivered to healthy tissue. Unfortunately, the literature reveals that suboptimal plans are approved frequently[5–8], even by experienced radiation oncologists. A contributing factor is the training environment during residency in which plan review is taught. Residents report the absence of examples across diverse treatment sites; the substantial clinical time pressures and high-stakes environment which make effective education difficult; and discontinuities introduced between identifying suboptimalities and making requests for edits and receiving the updated plan.

It is important to address these challenges by providing a wide range of example plans that need review and improvement in an environment that is low-stakes and without the pressure to have plans approved quickly for treatment. Simply curating lower-quality plans would be infeasible, as clinics typically have limited numbers available and inadvertent example memorization by the trainees would become common. Instead, a method of generating clinically realistic but suboptimal plans is needed. To date, however, most research has focused on generating high-quality plans[9–18]. Therefore, the current study develops a tool to create suboptimal dose distributions. These will be clinically realistic and suitable for integration into the training curriculum, providing the ability for real-time education in a low-stakes environment with virtually limitless examples.

2 Materials and Methods

In this study, suboptimal plans are modeled as potentially ideal dose distributions to which error has been added. This approach allows starting with consistent, high-quality dose distributions, and then creating areas of suboptimality around select OARs.

High-quality plans are required as the standard and what residents should be learning to achieve during plan review training. In this study, volume-modulated arc therapy (VMAT) dose distributions were predicted using the 3-Dimensional Dense Dilated U-Net model[10]. Two sets of models were used, one for head and neck (HN) radiotherapy[11] and the other for gynecological (GYN) radiotherapy[12]. Each set of models was previously trained using three-fold cross-validation on clinical data. This training data was carefully selected to be high-quality based upon both DVH metrics as well as upon physician review and consisted of planning computed tomography (CT) images, the clinical dose map from the treatment planning system, planning target volume (PTV) contours of varying prescriptions, and organ at risk (OAR) contours.

As the resultant dose distributions will be used for training, it is important that error be added to the high-quality dose predictions in a controllable manner that results in dose distributions free from obvious manipulation and appearing to have originated in the clinic. Dose changes should be localized to the OARs in which suboptimality is desired without substantial excess change elsewhere, but also without introducing visual errors such as introducing unrealistic dose gradients or other artifacts around the regions. To accomplish this, a convolutional technique was developed which adjusts changes based upon geometry –

that is, large changes in dose are only made near the desired OARs.

To start, a geometric feature λ_{OAR} was derived for each group of OARs. It describes proximity to the region in which suboptimality is desired through a combination of the Euclidean distances outside the OAR boundary and a directionally insensitive beam's eye view (BEV) angle. The geometric features were normalized to the range [0, 1], where 0 indicates least physical and angular distance from the OAR and 1 indicates greatest. Then a sigmoid-like function was applied with parameters controlling the full-width half-max (FWHM) and steepness of the sigmoid curve β, to provide additional control over the geometric space in which the dose was increased.

Equation 1

$$f(\lambda, FWHM, \beta) = \frac{\left(\lambda^{-\frac{\log(2)}{\log(FWHM)}}\right)^{\beta}}{\left(\lambda^{-\frac{\log(2)}{\log(FWHM)}}\right)^{\beta} + \left(1 - \lambda^{-\frac{\log(2)}{\log(FWHM)}}\right)^{\beta}}$$

As the target region contains the highest dose, a similar geometric feature λ_{PTV} was derived and normalized to the range [0, 0.98]. This feature indicated proximity for the PTV for which to make very small perturbations to the dose. This allowed changes to be made to the dose delivered to target, but not unrealistically large relative to the dose increase around each OAR. The final geometric feature was the 3-dimensional element-wise minimum of the OAR feature and the target feature.

Equation 23

$$\lambda = min\left(\lambda_{OAR,ijk}, \lambda_{PTV,ijk}\ \forall i,j,k\ in\ (Z,X,Y)\right)$$

Next, a kernel was designed to take this geometric information into account. There are two parameters for this kernel, geometric feature λ which varies at every step in the convolution, and the constant AUC_m, which is always 1 or greater and determines the maximum area under the curve of the kernel and therefore the resultant dose scaling.

Equation 4

$$g(\lambda, AUC_m) = s(\lambda, AUC_m) \cdot \frac{1}{\sum k} k(\lambda)$$

This may be further divided into two components. First, it is important that the overall kernel area under the curve (AUC) changes with response to geometry, so that greatest dose scaling is applied only near the OARs of interest. The linear function $s(\lambda, AUC_m) = (1 - AUC_m)\lambda + AUC_m$ controls the overall AUC of the kernel, such that it is equal to 1 as λ is largest (far from the OAR) and equal to AUC_m as λ is smallest (closes to the OAR and along the BEV angle).

The second component, the inner kernel k, is self-normalizing to AUC=1 and changes its receptive field based upon geometry. At the limits of the geometric feature, k is the identity kernel to preserve dose as-is ($\lambda = 1$), or the mean kernel to incorporate the most information from neighboring values ($\lambda = 0$).

Equation 5

$$k(\lambda) = \begin{bmatrix} 1-\lambda & 1-\lambda & 1-\lambda \\ 1-\lambda & 1 & 1-\lambda \\ 1-\lambda & 1-\lambda & 1-\lambda \end{bmatrix}$$

Finally, suboptimalities were introduced with greatest relative increase near the selected OARs through convolving the high-quality dose distribution with the geometry aware kernel.

Equation 6

$$D' = D * g(\lambda, AUCm)$$

To test this process, two datasets were used. First, a set of twenty-five HN VMAT radiotherapy cases were curated from our institution. These cases included a planning CT image, the clinical dose map from the treatment planning system, PTV contours of varying prescriptions, and OAR contours required to generate the high-quality dose predictions as described by Gronberg et al.[11] To assess translatability of our approach, an additional fourteen gynecological (GYN) VMAT cases were curated from our institution. These cases included a planning CT image, the clinical dose map from the treatment planning system, a single PTV contour prescribed to 45Gy, and OAR contours required to generate the high-quality dose predictions as described by Gronberg et al.[12]

Five HN OAR groupings were selected in this study (Table 1). The generation parameters listed in Table 1 were selected based upon preliminary experimentation to provide realistic-appearing dose distributions.

Group	*OARs*	*AUC_m*	*FWHM/β*
1	Esophagus Larynx	1.1-1.7	0.15/4
2	Brainstem	1.1-1.7	0.15/4
3	Optic Nerves Lenses	1.1-1.7	0.15/4
4	Parotids	1.1-1.7	0.15/4
5	Cochleae	1.1-1.7	0.15/4

Table 1: Groups of HN OARs for which the dose was increased in this study, with the generation parameters.

Changes to the plan dose was assessed as relative difference in structure mean dose D_{mean} between the edited plans and the predicted high-quality plans. For each OAR in Table 1,

the relative change was calculated as $\Delta D_{mean,OAR} = \frac{D'_{mean,OAR} - D_{mean,OAR}}{D_{mean,OAR}}$. A paired one-sided T-test was applied to assess significance in change to D_{mean} for each OAR, with significance taken at the $p<0.01$ level.

An experienced dosimetrist also reviewed a subset of the dose distributions, all generated with AUC_m=1.4, to assess clinical realism. To avoid bias, the dosimetrist was blinded to plan origin. A random mix of 10 clinical dose distributions and 10 suboptimal distributions was generated. The dosimetrist reviewed each dose distribution individually and answered if they thought the doses were (1) generated by a human with a treatment planning system, (2) generated by a deep learning dose prediction model, or (3) unable to determine or ambiguous.

3 Results

Dose was increased around the selected OARs in each group in a controllable way, as shown in Figure 2. This increase was statistically significant. Increasing kernel AUC_m increased mean dose as shown in Figure 1, where the relatively small error bars indicate strong control of change in dose. Figures 1 and 3 also demonstrate the localization of changes; mean dose is increased most around the OARs of interest in each group. The smaller increase in dose to other OARs are appropriate for clinical realism; while not changed as substantially as for the OARs of interest, they still exhibit realistic dose increases. While Figures 1 & 3 show a subset of groups and OARs for clarity, the general trend remains across all groups.

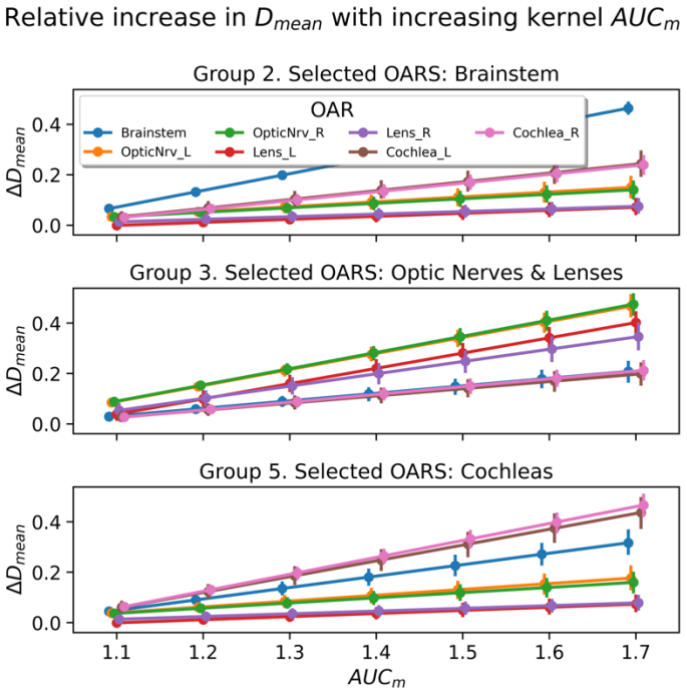


Figure 1: Relative increase in mean dose for selected OAR groups as the maximum kernel AUC increases.

The rate at which the dosimetrist was unable to determine if the dose distribution originated from a clinical treatment plan or a generated prediction was 80% for the suboptimal dose distributions, comparable to 70% for the clinical dose distributions. This indicates that the suboptimal dose distributions do not appear substantially different to plans that might be encountered in the clinic – that is, the generated plans are suboptimal but still clinically realistic.

The GYN dose distributions were edited using the same parameters as in Table 1, but with three groups of OARS: Bladder, Rectum, and the Femoral Heads together. D_{mean} was again increased for each OAR with increasing AUC_m (Table 2). This demonstrates the translatability of our approach across treatment sites.

AUC_m	*Bladder*	*Rectum*	*L. Femur Head*	*R. Femur Head*
1.1	0.02	0.01	0.07	0.07
1.2	0.05	0.03	0.13	0.14
1.3	0.07	0.04	0.20	0.20
1.4	0.09	0.06	0.26	0.27
1.5	0.12	0.07	0.33	0.34
1.6	0.14	0.09	0.39	0.40
1.7	0.16	0.10	0.46	0.47

Table 2: Relative increase in D_{mean} for the GYN cases.

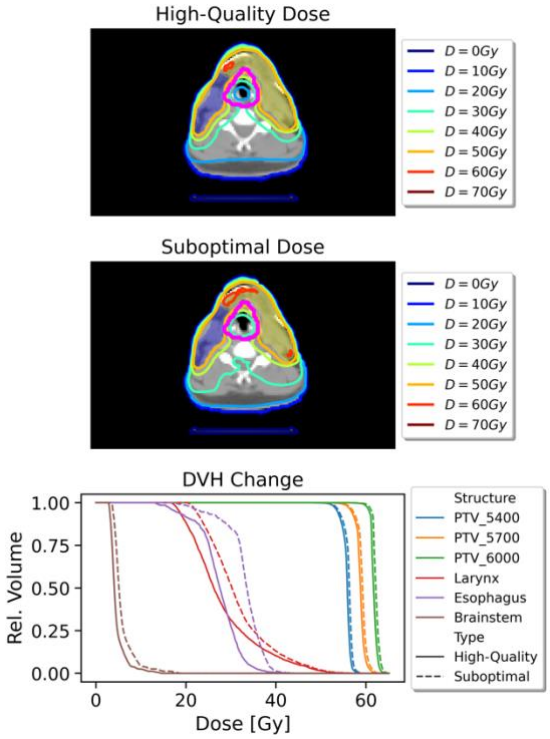


Figure 2: Comparison of the suboptimal dose distribution with the initial high-quality dose distribution. This example came from Group 1 where the larynx (magenta) and esophagus (only seen in the DVH)) were the OARs for which the changes are focused. Blue and yellow regions: low- and mid-risk PTV, respectively.

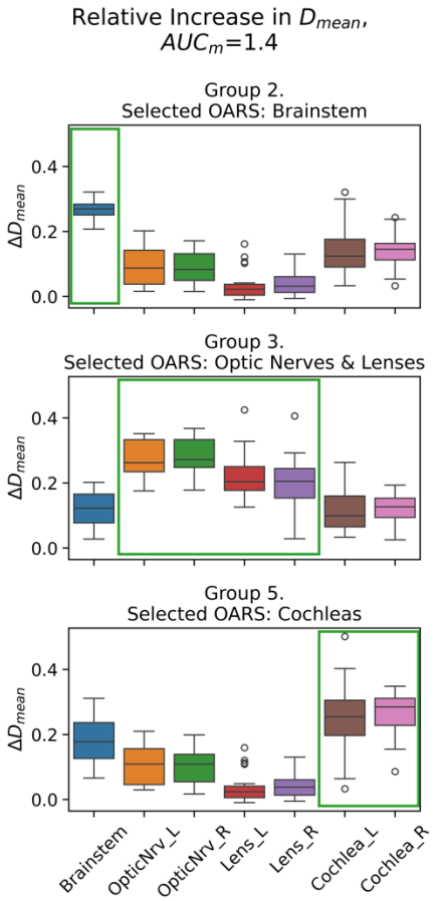


Figure 3: Distribution of relative increase in mean dose for selected OAR groups, with a single kernel AUC_m of 1.4. Green boxes indicate the selected OARs for each group.

4 Discussion

In this study, a tool for generating suboptimal but clinically realistic dose distributions was successfully developed. To our knowledge, this is the first such approach and opens the door to a range of applications for medical education and training.

The next step of this project will be to integrate this tool into the treatment planning system. Using scripting, an interactive model will be designed that presents trainees with different scenarios and asks them to identify where the treatment plan quality can be improved.

5 Conclusion

In this work, we develop a tool for generating artificial dose distributions that are clinically realistic. These dose distributions are suboptimal in predictable and controllable locations.

References

1. Peters, L. J. *et al.* Critical Impact of Radiotherapy Protocol Compliance and Quality in the Treatment of Advanced Head and Neck Cancer: Results From TROG 02.02. *JCO* **28**, 2996–3001 (2010).
2. Chien, C.-R. *et al.* High case volume of radiation oncologists is associated with better survival of nasopharyngeal carcinoma patients treated with radiotherapy: a multifactorial cohort analysis. *Clinical Otolaryngology* **36**, 558–565 (2011).
3. Wuthrick, E. J. *et al.* Institutional Clinical Trial Accrual Volume and Survival of Patients With Head and Neck Cancer. *JCO* **33**, 156–164 (2015).
4. Boero, I. J. *et al.* Importance of Radiation Oncologist Experience Among Patients With Head-and-Neck Cancer Treated With Intensity-Modulated Radiation Therapy. *JCO* **34**, 684–690 (2016).
5. Wu, B. *et al.* Patient geometry-driven information retrieval for IMRT treatment plan quality control. *Medical Physics* **36**, 5497–5505 (2009).
6. Moore, K. L., Brame, R. S., Low, D. A. & Mutic, S. Experience-Based Quality Control of Clinical Intensity-Modulated Radiotherapy Planning. *International Journal of Radiation Oncology*Biology*Physics* **81**, 545–551 (2011).
7. Moore, K. L. *et al.* Quantifying Unnecessary Normal Tissue Complication Risks due to Suboptimal Planning: A Secondary Study of RTOG 0126. *International Journal of Radiation Oncology*Biology*Physics* **92**, 228–235 (2015).
8. Talcott, W. J. *et al.* A Blinded, Prospective Study of Error Detection During Physician Chart Rounds in Radiation Oncology. *Practical Radiation Oncology* **10**, 312–320 (2020).
9. Babier, A. *et al.* OpenKBP: The open-access knowledge-based planning grand challenge and dataset. *Medical Physics* **48**, 5549–5561 (2021).
10. Gronberg, M. P. *et al.* Technical Note: Dose prediction for head and neck radiotherapy using a three-dimensional dense dilated U-net architecture. *Medical Physics* **48**, 5567–5573 (2021).
11. Gronberg, M. P. *et al.* Deep Learning–Based Dose Prediction for Automated, Individualized Quality Assurance of Head and Neck Radiation Therapy Plans. *Practical Radiation Oncology* (2023) doi:10.1016/j.prro.2022.12.003.
12. Gronberg, M. P. *et al.* Deep learning–based dose prediction to improve the plan quality of volumetric modulated arc therapy for gynecologic cancers. *Medical Physics* **50**, 6639–6648 (2023).
13. Barragán-Montero, A. M. *et al.* Three-dimensional dose prediction for lung IMRT patients with deep neural networks: robust learning from heterogeneous beam configurations. *Medical Physics* **46**, 3679–3691 (2019).
14. Chen, X., Men, K., Li, Y., Yi, J. & Dai, J. A feasibility study on an automated method to generate patient-specific dose distributions for radiotherapy using deep learning. *Medical Physics* **46**, 56–64 (2019).
15. Fan, J. *et al.* Automatic treatment planning based on three-dimensional dose distribution predicted from deep learning technique. *Medical Physics* **46**, 370–381 (2019).
16. Kearney, V., Chan, J. W., Haaf, S., Descovich, M. & Solberg, T. D. DoseNet: A volumetric dose prediction algorithm using 3D fully-convolutional neural networks. *Physics in Medicine and Biology* **63**, (2018).
17. Nguyen, D. *et al.* 3D radiotherapy dose prediction on head and neck cancer patients with a hierarchically densely connected U-net deep learning architecture. *Physics in Medicine and Biology* **64**, (2019).
18. Zhang, J. *et al.* Voxel-Level Radiotherapy Dose Prediction Using Densely Connected Network with Dilated Convolutions. in *Artificial Intelligence in Radiation Therapy* (eds. Nguyen, D., Xing, L. & Jiang, S.) 70–77 (Springer International Publishing, Cham, 2019). doi:10.1007/978-3-030-32486-5_9.

Feasibility of training federated deep learning oropharyngeal primary tumour segmentation models without sharing gradient information

Leroy Volmer[1], Ananya Choudhury[2], Aiara Lobo Gomes[2], Andre Dekker[1,2], and Leonard Wee[1,2]

[1]Department of Radiotherapy (MAASTRO), Maastricht University Medical Centre+, Maastricht, The Netherlands
[2]Clinical Data Science, School of Oncology and Reproduction, Maastricht University, Maastricht, The Netherlands

Abstract Federated deep learning is a way of training a deep learning model on vast amounts of privacy-sensitive data without having to exchange the data itself. A vulnerability exposed during federated model training relies on intercepting the gradients during training. In this work, we show that it is feasible to train a model for segmenting an oropharyngeal tumour in a simulated federated manner without any sharing of gradients between institutions. Further, we show that the federally trained model is functionally equivalent to a centrally trained one. In future, we expect geometric similarity performance could be significantly improved by combining multi-modality images as inputs, such as PET and CT.

1 Introduction

Oropharyngeal cancer is one member of a family of rare malignancies that arise from mutations in cells lining the upper aerodigestive tract. Over 50,000 cases are diagnosed in Europe each year, with roughly one in five of these occurring in persons aged under 55 years.
Accurate segmentations of the oropharyngeal gross tumour volume (GTV) are needed to plan and deliver a radiotherapy treatment. This is a time-consuming task being done routinely by experienced radiation oncologists, however deep learning neural networks have recently demonstrated some promising results for automatically segmenting the GTV [1].
Federated learning [2] is a paradigm of working within the limitations of sharing vast amounts of privacy-sensitive medical data to train models, including deep learning neural networks, on rare cancer subjects. An external attack surface has been identified, during a federated model training process, that suggests a method of breaching data privacy that relies on gradient information interception [3]. Being able to train a deep learning model federally without exchanging any gradient information would reduce risk of privacy breaches. Inversion of a model remains a theoretical possibility, but this will be true for any model where the weights are accessible, irrespective of whether it was trained federally or centrally.
In this work, we aim to prove the feasibility of training a deep learning model for oropharyngeal GTV segmentation in a federated learning simulation, but without transmitting any model gradients between cooperating parties. Furthermore, we evaluate the performance of the models in terms of geometric accuracy, and confirm that a federally trained model is functionally equivalent to a centrally trained one.

2 Materials and Methods

Four public datasets - HN1 [4], Montreal [5], OPC [6] and MDA [7] were taken from The Cancer Imaging Archive [8] for training and validation. We only included primary oropharyngeal cancers, resulting in 1223 volumetric CTs with associated primary GTV delineation by radiation oncologists, of which we randomly selected a holdout validation cohort of 20% from each aforementioned dataset. One private institutional dataset named HN3 (n=154) was used entirely as independent test, obtained with ethics authorization from MAASTRO Clinic in The Netherlands. HN3 was demographically the same population as HN1.
CT series and GTV masks were isotropically resampled to 1mm cubes. Hounsfield units were clipped between -1024 and 1024, and the CT intensity rescaled within the range (-1,1). Cubic 144^3 patches were repeatedly sampled for training, such that 75% contained GTV and 25% did not include any part of GTV. To increase robustness, data augmentation was randomly applied to each sampled cube, consisting of cranio-caudal mirror, translation (0/30mm), rotation (-30/30 degrees), shear (-0.1/0.1), gamma intensity scale (0.8/1.2) and Gaussian blur (0/3mm).
The deep learning architecture was based on a squeeze-and-excitation normalization model that has been used for HNC GTV segmentation [1,9]. Succinctly, the model is a 3D U-Net architecture with ResNet elements, with squeeze and excitation normalization blocks following each convolutional block that allows the network to find the ideal normalization for Hounsfield units during training.
Additionally, we included attention gates at each spatial resolution level following every skip connection, aiming to accelerate the localization of the GTV and enhance neuron activations in important areas [10]. Extra up-sampling paths were added to allow low-resolution features to have an impact further along in the network. Lastly, network attention was visualized using Grad-CAM++ [11].
ReLU activation functions were applied throughout, and an Adam optimizer with betas set to 0.9 and 0.99. Weight decay was 1e-4. A batch of 1 was used to reduce GPU memory usage. A compound function of Dice loss and Focal loss was used during training since the GTV represented only a small part of the input CT [12]. False negative predictions are thus penalized less than false positive predictions, therefore we added a weight to over-penalize false negative predictions. Dropout rate was 25% to reduce over-fitting. The learning rate followed a cosine annealing plus warm restart schedule, starting at amplitude 1e-4 decaying to 1e-7 every

10 epochs. Learning rate schedule was the same for all training runs. Starting weights were set with He initialization through a constant seed number for all training runs [13].
Federated deep learning was simulated by placing each dataset in a separate hard disk partition. Four partial models were trained on the data, one in each partition, then only the weights were merged into a single global model using the synchronous federated averaging (FedAvg) [14] algorithm. Averaged weights were used as starting state for the next training epoch. During each epoch of training, each partial model iteratively ran through all of the training samples inside its own partition. There was no exchange of gradient information whatsoever between any of the partial models and between partial and globally averaged model.
We monitored training and convergence dynamics while globally averaging every 1, 5 and 10 epochs, up to a total of 100 epochs. We compared the three "federally-trained" models, "Avg1", "Avg5" and "Avg10", respectively, with a fully centrally trained model with all the data in a single pool for 100 epochs. Only a fixed learning rate was used during the training cycle with FedAvg, to remove the effect of decaying learning rates in the epochs between averaging. For inference, a sliding window method was applied with overlap factor 0.5. No post-inference cleanup processing was applied. The geometric accuracy metric used was Dice Similarity Coefficient (DSC). Statistical significance of DSC between groups was calculated using a two-sided Mann-Whitney-Wilcoxon test including Bonferroni correction.

3 Results

Holdout subjects in HN1 (n=18), OPC (n=101), MDA (n=80) and Montreal (n=48) were used for validation of the trained models. HN3 (n=154) was used as independent test set to evaluate segmentation performance. Median DSC results are summarized in Table 1.

	HN1	*OPC*	*MDA*	*Montreal*	*HN3*
Central	**0.68** (0.20-0.77)	**0.69** (0.52-0.76)	**0.69** (0.42-0.72)	**0.61** (0.42-0.72)	**0.65** (0.41-0.74)
Avg1	**0.73** (0.16-0.77)	0.65 (0.55-0.73)	0.67 (0.45-0.78)	**0.63** (0.46-0.75)	**0.62** (0.33-0.76)
Avg5	0.72 (0.60-0.74)	0.66 (0.54-0.74)	0.65 (0.49-0.79)	0.52 (0.41-0.72)	0.55 (0.30-0.73)
Avg10	0.72 (0.59-0.79)	**0.68** (0.54-0.76)	**0.69** (0.50-0.77)	0.62 (0.46-0.73)	0.59 (0.30-0.74)

Table 1: Median Dice similarity coefficients (DSC) of predicted versus reference GTV segmentations, as a function of averaging every 1, 5 or 10 epochs. Interquartile range in DSC given in parentheses under median values.

Statistically significant differences (assuming $p<0.05$) in DSC were not detected, neither within datasets across all methods of training, nor when comparing centralized versus any federated averaging method. Among the federated averaging training methods, "Avg1" gave maximum DSC for 3 out of 5 datasets. "Avg10" was marginally better for OPC and MDA. Overall, "Avg1" and the centralized model were functionally equivalent in performance across all datasets.
However, the difference in performance as a function of total size of GTV was statistically significant. Figure 1 shows the DSC of each subject in the validation and test dataset as a dot, and grouped according to GTV $\geq 10cm^3$ versus GTV $< 10cm^3$. Model performance in the smaller GTVs were poorer in terms of median DSC and interquartile spread, and has the overall effect of dragging the median DSC results downwards.

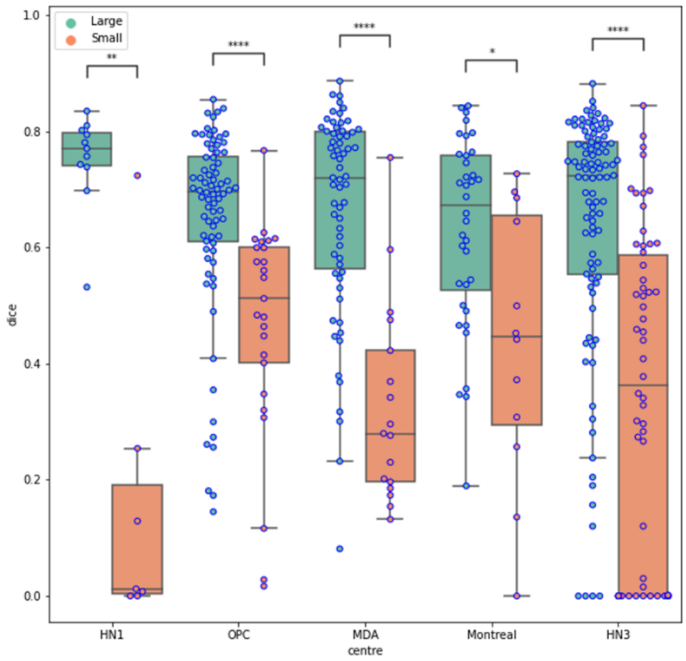


Figure 1: Boxplot showing distribution of Dice similarity coefficient for each subject in the validation datasets as a function of GTV size ($\geq 10cm^3$ in green, $<10cm^3$ in orange) when using the Avg1 trained model. Differences in DSC due to size of GTV was statistically significant.

Examples of automated segmentations produced by model Avg1 in validation subjects are provided in Figure 2. Overall, the model accurately locates the majority of the middle of the GTV, but the extreme outer boundaries of the tumour were not in agreement with the radiation oncologist delineation. Grad-CAM activation maps were partially helpful, such that the vast majority of subjects showed activations near the GTV, but the GTV was not consistently contained within the strongly activation area.

4 Discussion

This study shows that it would be possible to train a GTV segmentation model for oropharyngeal carcinoma in a federated fashion, where we would not need to exchange gradients between institutions. This would mean that there is no opportunity to eavesdrop on gradients, thus removing a potential threat to privacy breach in federated deep learning. Furthermore, with synchronized federated averaging (FedAvg) of partial model weights into a global model every epoch, we showed that we could obtain a federated model that was functionally equivalent to training on all the data centrally. In terms of geometric performance, it is clear that GTVs smaller than 10cm^3 in total volume poses a significant challenge for automated segmentation. Future work would explore whether splitting the deep learning model for large and small GTVs, or considering an ensemble of models, could be feasible options.

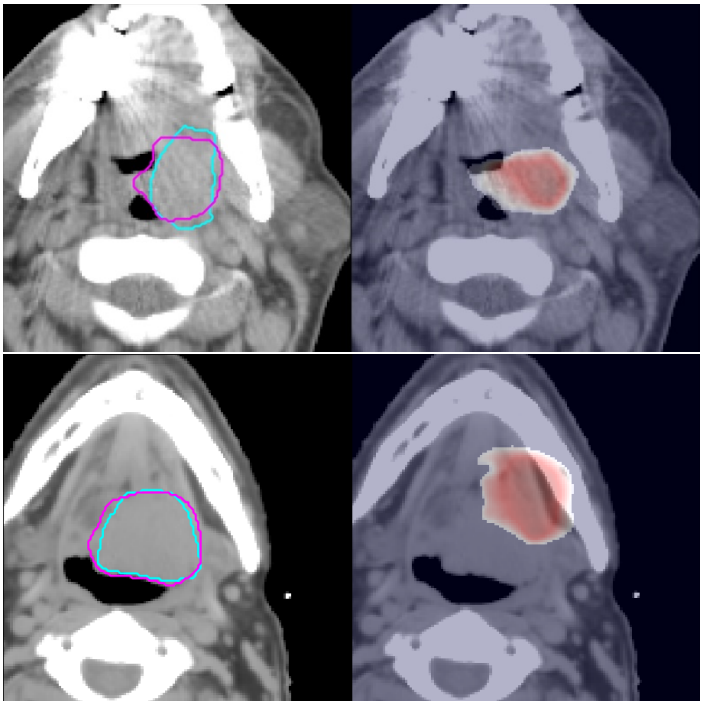


Figure 2: Two examples comparing predicted and reference GTV segmentations. On the left, purple contours are reference delineations of GTV by radiation oncologist and blue contours are inferred by model Avg1. On the right, the red hotspots indicate the highly activating areas using Grad-CAM++. In the background of all images are the original CT image intensities.

There remain a number of issues in a real-world federated deep learning training run that we do not presently have space to address in this report. Inter-observer delineation differences and contouring quality disparities persist between radiotherapy clinics. Our approach for multi-institutional collaboration suggests we can train a deep learning model towards an average global consensus, but we cannot address the underlying data quality issues by federated learning alone.

Secondly, we have only used CT images in this work, which is by far the most common imaging modality in radiotherapy, but other studies [15] show that combining multi-imaging inputs, such as PET, CT and MRI, would should lead to significantly better geometric agreement with radiation oncologists' delineations of GTV.

Third, we surmise that while inversion from updated model weights could be possible for fully-connected neural networks trained on very small sample sizes [16], this kind of model inversion would be significantly harder to execute for a realistic training run with a 3D U-Net on a large sample size, at least in comparison to intercepted training gradients.

In a real-world training run, we may recommend the following as safeguards – (i) that partial model weights from each institution should consist of learning on a reasonably large number of subjects (e.g., 100 subjects or greater per institution), and (ii) that global model weights for each averaging epoch should be aggregated on a dedicated trusted third-party device and then the partial weights should be deleted from that device after running FedAvg. In this way, neither the training institutions nor the researcher that initiated the deep learning run would have access to any partial model weights, only the globally-averaged weights on the total training sample would be available to the training run participants.

Lastly, obviating the need to exchange gradients between training institutions defeats one well-known attack surface, but other security risks persist. In a real world situation, it might still be possible to steal access credentials, execute man-in-the-middle attacks or attempt data re-engineering from a final fully-trained model. Additional privacy-enhancing techniques such as differential privacy and homomorphic encryption may be added to boost security, but one ought to be wary of privacy enhancement technology alone as a single line of protection against potential privacy breaches.

5 Conclusion

It is feasible to train a 3D U-Net for primary oropharyngeal GTV segmentation in federated learning without transmitting model gradients between institutions. The FedAvg algorithm for aggregating multiple partial models into a global model does not significantly impact model performance. The Avg1 model was functionally equivalent in terms of DSC with a deep learning model trained centrally on all patient data at the same time in a single pool.

References

[1] V Oreiller, V Andrearczyk, M Jreige, S Boughdad, H Elhalawani, J Castelli, et al. Head and neck tumor segmentation in PET/CT: The HECKTOR challenge. Med Image Anal. 2022 Apr 1;77:102336.

[2] S Gupta, S Kumar, K Chang, C Lu, P Singh, J Kalpathy-Cramer. Collaborative Privacy-preserving Approaches for Distributed Deep Learning Using Multi-Institutional Data. RadioGraphics 43 (2023). DOI : 10.1148/rg.220107.
[3] Liu, P., Xu, X. & Wang, W. Threats, attacks and defenses to federated learning: issues, taxonomy and perspectives. Cybersecurity 5, 4 (2022). DOI : 10.1186/s42400-021-00105-6.
[4] Wee L, Dekker A. et al. Data from HEAD-NECK-RADIOMICS-HN1 [Internet]. The Cancer Imaging Archive; 2019. https://wiki.cancerimagingarchive.net/x/iBglAw.
[5] Vallières M, Kay-Rivest E, Perrin L, Liem X, Furstoss C, Khaouam N, et al. Data from Head-Neck-PET-CT [Internet]. The Cancer Imaging Archive; 2017. https://wiki.cancerimagingarchive.net/x/24pyAQ.
[6] Kwan JYY, Su J, Huang SH, Ghoraie LS, Xu W, Chan B, et al. Data from Radiomic Biomarkers to Refine Risk Models for Distant Metastasis in Oropharyngeal Carcinoma [Internet]. The Cancer Imaging Archive; 2019. https://wiki.cancerimagingarchive.net/x/XAQGAg.
[7] Grossberg A, Elhalawani H, Mohamed A, Mulder S, Williams B, White AL, et al. HNSCC [Internet]. The Cancer Imaging Archive; 2020. https://wiki.cancerimagingarchive.net/x/sIN5Ag.
[8] Clark K, Vendt B, Smith K, Freymann J, Kirby J, Koppel P, et al. The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository. J Digit Imaging. 2013 Dec 1;26(6):1045–57.
[9] Iantsen A, Jaouen V, Visvikis D, Hatt M. Squeeze-and-Excitation Normalization for Brain Tumor Segmentation. In: Crimi A, Bakas S, editors. Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. Cham: Springer International Publishing; 2021. p. 366–73. (Lecture Notes in Computer Science).
[10] Oktay O, Schlemper J, Folgoc LL, Lee M, Heinrich M, Misawa K, et al. Attention U-Net: Learning Where to Look for the Pancreas [Internet]. arXiv; 2018. http://arxiv.org/abs/1804.03999.
[11] Chattopadhay A, Sarkar A, Howlader P, Balasubramanian VN. Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). 2018. p. 839–47.
[12] Lin TY, Goyal P, Girshick R, He K, Dollár P. Focal Loss for Dense Object Detection. IEEE Trans Pattern Anal Mach Intell. 2020 Feb;42(2):318–27.
[13] K. He, X. Zhang, S. Ren and J. Sun, "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification," 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 2015, pp. 1026-1034, DOI: 10.1109/ICCV.2015.123.
[14] McMahan Brendan, Moore Eider, Ramage Daniel, Hampson Seth, and Arcas Blaise Aguera y. 2017. Communication-efficient learning of deep networks from decentralized data. In Proceedings of the Artificial Intelligence and Statistics. PMLR, 1273–1282.
[15] Ren J, Eriksen JG, Nijkamp J, Korreman SS. Comparing different CT, PET and MRI multi-modality image combinations for deep learning-based head and neck tumor segmentation. Acta Oncol. 2021 Nov;60(11):1399-1406. doi: 10.1080/0284186X.2021.1949034.
[16] D Enthoven and Z Al-Ars, Reconstructing Private Training Samples from Weight Updates in Federated Learning. [Internet] arXiv ; 2022. https://arxiv.org/pdf/2101.00159.

Follow the Leader—a General Template-Guided Deep Learning Framework for Adaptive Radiotherapy Target Delineation

Mahdieh Kazemimoghadam[1], Qingying Wang[1], Mingli Chen[1], Xuejun Gu[1, 2*] and Weiguo Lu[1*]

[1] Department of Radiation Oncology, The University of Texas Southwestern Medical Center, Dallas TX, 75390 United States of America
[2] Department of Radiation Oncology, Stanford University, Stanford, CA 94305, United States of America

Abstract
Adaptive Radiotherapy (ART) heavily relies on efficient and accurate target delineation due to the need for real-time adjustments. However, manual contouring hinders ART integration with its variability, labor-intensive, and time-consuming nature. This study introduces a novel template-guided deep learning model for target segmentation. The model built upon a 3D nnUNet featuring three input channels: primary CT image, template CT image, and template mask (GTV, CTV, or PTV). The structure allows "Follow-the-Leader" nature of learning, i.e. the output to dynamically adjust based on the chosen template (the leader), providing segmentations tailored to the template. Additionally, it incorporates stylistic cues from the template leading to accurate segmentation. The approach was tested on a dataset of breast cancer CT images. Patient data (n=114) were randomly split into training (84), validation (15), and test (15). The model achieved Dice similarity coefficients (GTV: 0.74±0.06, CTV: 0.86±0.04, PTV: 0.87±0.03) showcasing its potential for improving ART workflow.

1 Introduction

Online Adaptive Radiotherapy (ART) represents a sophisticated radiotherapy approach wherein the daily treatment plan is adjusted to accommodate the patient's evolving anatomy, such as a shrinking tumor and variations in body weight [1]. These benefits come with the drawback of extended workflows, primarily due to the necessity for online re-contouring. Efficiency is crucial in ART given the real-time nature of this treatment, requiring the patient to be immobilized in the treatment position while awaiting the initiation of treatment [2], [3]. However, the time-consuming task of delineating targets and organs-at-risk (OARs) for all fractions has emerged as a significant bottleneck, impeding the widespread adoption of online ART.

In ART, the GTVs (Gross Tumor Volumes), CTVs (Clinical Target Volumes), and PTVs (Planning Target Volumes) on the treatment day are either manually delineated or auto-deformed from planning CT and validated by radiation oncologists. The definition of these target areas relies on the clinician's expertise, experience, and a comprehensive understanding of the patient's diagnostic imaging, surgical reports, and other relevant data [4], [5]. The quality, efficiency, and repeatability of manual contouring can vary among different radiation oncologists, and the time spent on contouring a patient is influenced by their proficiency. For instance, in clinical practice, manual target delineation for breast cancer cases could range from 30 minutes to several hours.

Atlas-based automatic segmentation algorithms have been extensively employed for target segmentation [6]. Nonetheless, these algorithms encounter challenges when dealing with targets characterized by indistinct boundaries or intricate shapes [7] and frequently yield unsatisfactory results. Consequently, there is a demand for accurate auto-segmentation tools to address the challenges.

Our proposed deep learning (DL) model presents an innovative approach by leveraging template-guided segmentation. The proposed model is based on 3D nnUNet [8] by incorporating three input channels: the input CT image, a template CT image, and a template mask. The template mask encompasses either GTV, CTV, or PTV. The model leverages the knowledge from the template to guide the segmentation in dual aspects. 1) The utilization of templates enables the model to dynamically adapt to various targets, generating precise segmentation based on the type of template selected. In other words, the choice of the template determines the output, allowing the model to generate contours for GTV, CTV, or PTV. For instance, if the GTV template is chosen, the output corresponds to GTV; if the CTV template is selected, the output is CTV, and similarly for PTV. This versatility empowers the model to seamlessly tailor its segmentation output to the specific target of interest, offering a flexible and personalized approach to adaptive radiotherapy. 2) In adaptive radiation therapy, a sequence of images is required to be contoured throughout the treatment course. Typically, some prior CT and contour information is usually available from the initial treatment planning and a few early fractions [9]. Our proposed DL approach is designed in a way to leverage the insights from existing contours as templates to effectively segment the new image. This strategic use of prior information enhances the adaptability of the model, allowing it to learn and apply segmentation patterns in the template to ensure accurate and consistent results throughout the treatment course.

2 Materials and Methods

In this study, we proposed and developed a template-based auto-segmentation scheme for ART. Initially, three 3D nnUNet models were trained exclusively on data corresponding to their specific targets—GTV, CTV, or PTV, each operating with a single-channel setup (i.e., CT image). These three models are used as baseline models for comparison. Further, three distinct 3-channel 'target-specific' models were trained, each dedicated to GTV, CTV, and PTV. These models incorporated CT images, template CT images, and target-specific template masks. The applicability of target-specific models is restricted to the segmentation of their designated targets, namely GTV, CTV, or PTV, lacking versatility for other target types.

Table 1. The table summarizes input requirements and training data for the baseline models, target-specific models, and the proposed model

Models	Input(s)	Training data
Baseline models	• CT image	The three models are trained separately for each target type (GTV, CTV, or PTV) using their specific template.
Target-specific models	• CT image • Template CT image • Specific template mask, limited to GTV, CTV, or PTV, depending on the model	The three models are trained separately for each target type (GTV, CTV, or PTV) using their specific template.
Proposed model	• CT image • Template CT image • Template mask, which can flexibly be GTV, CTV, or PTV depending on the chosen target	The model is trained on pooled template masks of GTV, CTV, and PTV for comprehensive learning and adaptability.

Recognizing the limited adaptability of target-specific models for diverse targets, we introduced the proposed model (Table 1, Figure 1), which trained on pooled of GTV, CTV, and PTV data, offering a more versatile solution. This comprehensive approach aims to overcome the constraints posed by exclusive target-specific models and provides flexibility in choosing templates, addressing the need for a more adaptable and generalized model for adaptive radiotherapy. Overall, seven models as listed in Table 1 are trained in this study for comprehensive models evaluation and comparison.

2.1 Adaptive Output Based on Template Selection

During training, the proposed model learns to adapt its output based on the selected template. The three input channels—CT images, template CT images, and template target masks (GTV, CTV, or PTV)—facilitate joint learning. The model grasps the relationships between these inputs, enabling it to dynamically adjust its segmentation output according to the specific target delineated by the chosen template. This adaptability is achieved through continuous refinement of network weights via supervised learning, ensuring precise target segmentation aligned with the chosen template.

2.2. Stylistic Cue Utilization

Simultaneously, the model gains insights into the stylistic cues present in the template CT and template target mask. This information is incorporated into the learning process, allowing the model to generalize stylistic characteristics and apply them to generate segmentation masks for the input image.

Throughout training, the model refines its weights through continuous updates, minimizing the loss function by comparing predictions with manually contoured masks. Here, the loss function is the sum of cross-entropy and Dice loss (*Eq. 1*). The trained model, with optimized parameters, is then ready for deployment during the testing stage, where it utilizes template guidance to dynamically adapt and generate predicted segmentations.

$$L(G,P) = -\frac{1}{N}\sum_{k=1}^{K}\sum_{n=1}^{N}\left(g_{n,k}\log p_{n,k} + \frac{2\,g_{n,k}\,p_{n,k}}{g_{n,k}^2 + p_{n,k}^2}\right);\ \ g_{n,k} \in G \quad p_{n,k} \in P \qquad \textit{Eq.1}$$

In *Eq. 1*, G denotes the ground truth label, and P is the predicted probabilities for class k and voxel n in the image.

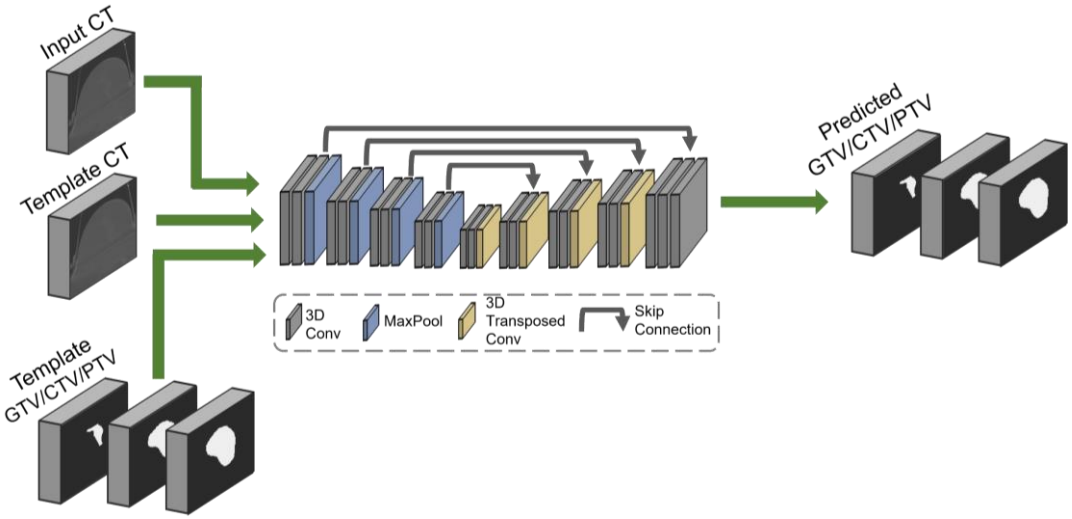


Figure 1. The proposed template-guided model structure. The 3D nnUNet is the model backbone. The model is trained using three input channels including, the CT images of the patient, a template CT image, and the template target mask. The utilization of the input template (GTV, CTV, or PTV) enables the model to dynamically adapt to various targets and stylistic cues provided by the template to predict the output segmentation masks.

2.3 Experimental Data

We utilized our in-house retrospective dataset of 114 post-operative breast SBRT patients in this study. Prior to each fraction, head-first-prone CT scans were acquired on either Airo® Mobile Intraoperative CT (Brainlab AG, Munich, Germany) or Philips Brilliance Big Bore CT (Philips Healthcare, Amsterdam, Netherlands), both with 120 kVp and scanning lengths covering the whole chest region. CT images had slice thickness of 1 mm. To mitigate potential bias arising from high correlation across fractional images for each patient, we deliberately partitioned the data based on patients rather than individual images. The test dataset, comprising 15 patients, remained entirely untouched during experiments, ensuring no overlap with the training set. Out of the total 114 patients, 84 were randomly assigned to the training set, 15 to the validation set, and the remaining 15 to the test. GTV, CTV, and PTV delineation on the CT images was conducted through a meticulous process by expert radiation oncologist from the department of radiation oncology at UT Southwestern Medical Center who had similar levels of expertise. Within each patient's dataset, our training approach for the proposed model involved exhaustive exploration of all possible combinations of fractions for GTV, CTV, and PTV. Given the five fractions per patient, our training dataset for each patient encompassed 20 instances, and in total 5040 (20 instances × 84 patients × 3 targets) training samples. This comprehensive sampling facilitates robust learning.

2.4 Implementation Details

The experiments utilized the PyTorch framework (version 1.8) and Python environment (version 3.8) on a Windows 10 64x system, equipped with an Intel Xeon processor CPU (64 GB RAM) and an NVIDIA GeForce 2080 Ti GPU (12 GB memory). For uniformity, all CT images and corresponding GTV, CTV, and PTV masks were resampled to voxel size of 1.5 mm × 1.5 mm × 1.5 mm. Subsequently, images and masks underwent center cropping to 128 × 128 × 128, focusing on the breast region.

2.5 Network Evaluation

In simulating adaptive radiation therapy, our testing approach replicates the scenario where initial treatment planning and early fractions provide prior CT and contour information. Hence, the input CT image was sourced from fractions 2, 3, 4, and 5, while maintaining fraction 1 as the template. The Dice similarity coefficient (DSC) was used to quantitatively assess the segmentation performance of the model (*Eq.2*). DSC was computed to measure the degree of overlap between the model's output and the ground truth GTV, CTV, and PTV masks. The proposed network was evaluated against the baseline models and the target-specific models.

$$DSC = \frac{2\times|A\cap B|}{|A|+|B|} \qquad Eq.2$$

3 Results

In Table 2, the baseline models provide a reference point, indicating the performance achieved without any specific target adaptation. The target-specific models as described in Table 1, are designed for a particular target type during training, and the proposed model is trained on combined data for all target volumes (GTV, CTV, and PTV). According to Table 2, notable improvements were observed across all target segmentation categories (i.e., GTV, CTV, PTV), for the target-specific models relative to the baseline. The target-specific models demonstrated enhanced performance, with respective DSC of 0.75 (GTV), 0.85 (CTV), and 0.87 (PTV). These values signify a substantial increase in accuracy compared to the baseline model, which yielded coefficients of 0.69 (GTV), 0.81 (CTV), and 0.82 (PTV). The narrower standard deviations associated with the proposed model's results (0.05 for GTV, 0.04 for CTV, and 0.03 for PTV) further highlight its consistency and robustness. The Proposed model maintained competitive performance across GTV, CTV, and PTV with DSC values of 0.74, 0.86, and 0.87 respectively, demonstrating its adaptability and effectiveness in handling various target types.

Table 2. Comparing Dice similarity scores for the proposed model, target-specific model, and the baseline. Values represent mean (standard deviation) across patients.

	GTV	CTV	PTV
Baseline models	0.69 (0.08)	0.81 (0.07)	0.82 (0.05)
Target-specific models	0.75 (0.05)	0.85 (0.04)	0.87 (0.03)
Proposed model	0.74 (0.06)	0.86 (0.04)	0.87 (0.03)

Visual inspection of the segmentation results provides insights into the performance of the proposed model compared to the baseline (Figure 2). To generate the predicted targets in Figure 2, the template—GTV, CTV, and PTV—were individually fed into the proposed model along with the input CT image and the CT template. The proposed model demonstrates a unique capability when aligning with the template; as presented in Figure 2, the model adapts its output based on the selected template (i.e., GTV, CTV, or PTV). Across all targets, the baseline model exhibits a lower accuracy in generating masks compared to the proposed model, signifying the superior segmentation capabilities of the latter. Using the baseline model, significant under-segmentation of the predicted GTV, CTV, and PTV was observed (Figure 2). However, the proposed model utilized the stylistic cues embedded in the template mask and generated segmentation masks that align with the visual characteristics of the specific template leading to improved performance. The results showcases the adaptability of the proposed model to follow the chosen template's characteristics, which has potential for tackling real-time adjustment challenges in adaptive radiotherapy.

4 Discussion

Adaptive Radiotherapy depends significantly on precise and efficient delineation of targets volumes which is

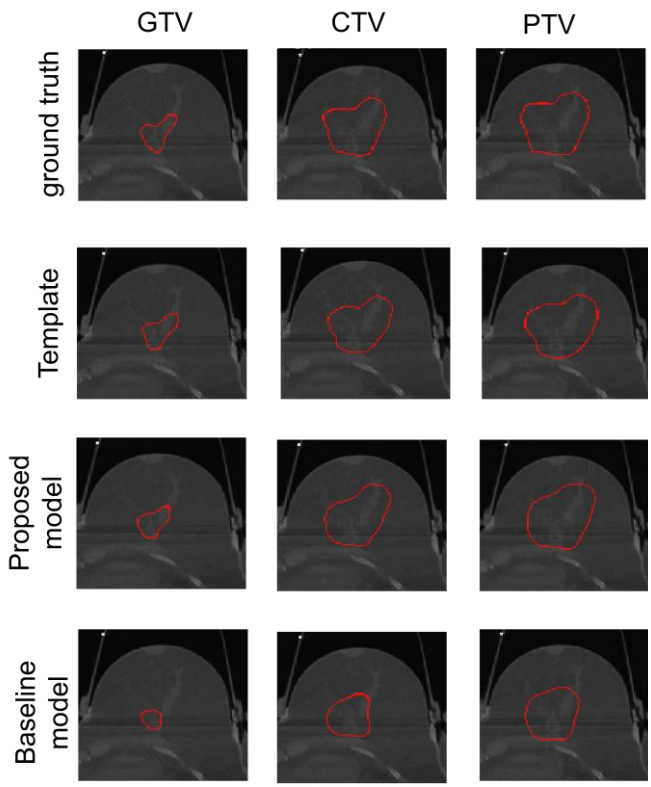


Figure 2. Visual comparison of segmentation results. Each row illustrates the ground truth mask, template, output from the proposed model, and output from the baseline model for GTV, CTV, and PTV, respectively.

essential for prompt adjustments in real-time [10]. This study introduces a solution for automatic target segmentation in ART by integrating DL and template guidance.

The model's dual functionality emerges as a key innovation. The adaptability of the model to generate precise segmentation based on the selected template (GTV, CTV, or PTV) provides a tailored and flexible approach. This adaptability ensures that the model can dynamically adjust to various targets, enhancing its versatility and applicability in different clinical scenarios. The utilization of stylistic cues from the template represents a novel aspect of the model. By capturing and generalizing stylistic intricacies, the model enhances its ability to generate accurate segmentation masks for evolving anatomies. The enhanced performance of the proposed model is attributed to the utilization of stylistic cues embedded in the template mask. By incorporating the template mask during training, the model learns to capture distinctive stylistic features associated with the chosen target (GTV, CTV, or PTV). During inference, when the model encounters a new image, it leverages these learned stylistic cues to generate segmentation masks that align with the visual characteristics of the specific target. This feature holds promise for addressing challenges related to real-time adjustments in adaptive radiotherapy. The ability to leverage prior information from template expedites the contouring process, reducing the time and effort traditionally associated with manual delineation. This efficiency is crucial for the practical implementation of ART in clinical settings.

While the model is demonstrated on breast cancer data, its underlying structure and principles can be expanded to apply to diverse adaptive radiotherapy scenarios. The model's generalizability to different anatomical sites positions it as a versatile tool for broader clinical adoption.

It is essential to consider potential limitations. The model's performance could be influenced by variations in imaging protocols, patient characteristics, and anatomical complexities. Addressing these factors in future studies will contribute to a more comprehensive understanding of the model's robustness.

5 Conclusion

The template-guided deep learning model with an inclusive training scheme presented in this paper showcases a novel paradigm for ART target segmentation. Adapting output based on chosen templates and employing stylistic cues presents a promising solution to the challenges of dynamic anatomical changes in ART workflow.

References

[1] S. Lim-Reinders, B. M. Keller, S. Al-Ward, A. Sahgal, and A. Kim, “Online Adaptive Radiation Therapy,” *Int. J. Radiat. Oncol. Biol. Phys.*, vol. 99, no. 4, pp. 994–1003, 2017, doi: 10.1016/j.ijrobp.2017.04.023.

[2] O. L. Green, L. E. Henke, and G. D. Hugo, “Practical Clinical Workflows for Online and Offline Adaptive Radiation Therapy,” *Semin. Radiat. Oncol.*, vol. 29, no. 3, pp. 219–227, 2019, doi: 10.1016/j.semradonc.2019.02.004.

[3] J. Lamb *et al.*, “Online Adaptive Radiation Therapy: Implementation of a New Process of Care,” *Cureus*, vol. i, no. 8, 2017, doi: 10.7759/cureus.1618.

[4] S. K. Vinod, M. Min, M. G. Jameson, and L. C. Holloway, “A review of interventions to reduce inter-observer variability in volume delineation in radiation oncology,” *J. Med. Imaging Radiat. Oncol.*, vol. 60, no. 3, pp. 393–406, 2016, doi: 10.1111/1754-9485.12462.

[5] D. Jin, D. Guo, J. Ge, X. Ye, and L. Lu, “Towards automated organs at risk and target volumes contouring: Defining precision radiation therapy in the modern era,” *J. Natl. Cancer Cent.*, vol. 2, no. 4, pp. 306–313, 2022, doi: 10.1016/j.jncc.2022.09.003.

[6] C. E. Cardenas, J. Yang, B. M. Anderson, L. E. Court, and K. B. Brock, “Advances in Auto-Segmentation,” *Semin. Radiat. Oncol.*, vol. 29, no. 3, pp. 185–197, 2019, doi: 10.1016/j.semradonc.2019.02.001.

[7] J. E. Iglesias and M. R. Sabuncu, “Multi-atlas segmentation of biomedical images: A survey,” *Med. Image Anal.*, vol. 24, no. 1, pp. 205–219, 2015, doi: 10.1016/j.media.2015.06.012.

[8] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, “nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation,” *Nat. Methods*, vol. 18, no. 2, pp. 203–211, 2021, doi: 10.1038/s41592-020-01008-z.

[9] R. de Jong, J. Visser, N. van Wieringen, J. Wiersma, D. Geijsen, and A. Bel, “Feasibility of Conebeam CT-based online adaptive radiotherapy for neoadjuvant treatment of rectal cancer,” *Radiat. Oncol.*, vol. 16, no. 1, pp. 1–11, 2021, doi: 10.1186/s13014-021-01866-7.

[10] Y. Fu *et al.*, “Pelvic multi-organ segmentation on cone-beam CT for prostate adaptive radiotherapy,” *Med. Phys.*, vol. 47, no. 8, pp. 3415–3422, 2020, doi: 10.1002/mp.14196.

Clarifying the dosimetric impact of autosegmentation inaccuracies

Joëlle E. van Aalst[1,2], Tomas Janssen[3], Roel J. H. M. Steenbakkers[2], Ilse G. van Bruggen[2], Jelmer M. Wolterink[4], Stefan Both[2], Johannes A. Langendijk[2], Charlotte L. Brouwer[2]

[1]Technical Medicine, University of Twente, Enschede, The Netherlands
[2]Department of Radiation Oncology, University Medical Centre Groningen, Groningen, The Netherlands
[3]Department of Radiation Oncology, Netherlands Cancer Institute, Amsterdam, The Netherlands
[4]Department of Applied Mathematics, Technical Medical Centre, University of Twente, Enschede, The Netherlands

Abstract Deep learning-based autosegmentation is conventionally assessed using geometric metrics. However, for clinically relevant quality evaluation, the dosimetric impact and the impact on normal tissue complication probability (NTCP) must be considered. Two distinct dosimetric effects can be distinguished. First, segmentation influences the optimized treatment plan (dosimetric impact on planning, DIP) and second, segmentation influences the dose evaluated in an organ (dosimetric impact on evaluation, DIE). We provide a comprehensive autosegmentation evaluation, considering geometric metrics, the DIP and DIE and the impact on NTCP of 68 head and neck cancer patients on 11 OARs. Mean Dice scores (0.50-0.90) and mean Hausdorff distances (0.74-1.26 mm) of autosegmentation did not correlate with the dosimetric measures. No significant mean dose effect exceeding 2 Gy was observed. The DIE showed larger variations than the DIP. An NTCP difference of ≥1.0 percentage point was found with the DIE for xerostomia. Furthermore, the DIE can be approximated using retrospective clinical dose distributions, reducing the workload of autosegmentation evaluation.

1 Introduction

Deep learning-based autosegmentation has gained prevalent attention by having the potential to improve accuracy, reproducibility and efficiency in segmentation in clinical practice [1]. Many different models are currently being developed, both in research groups and by commercial parties. For a proper comparison of these models and their safe introduction into clinical practice, it is important to understand the clinical impact of errors due to autosegmentation inaccuracies

Typically, autosegmentaion models are evaluated using a geometric analysis on metrics like the Sørensen-Dice similarity coefficient (DSC) or Hausdorff distance (HD). However, it has been shown that geometric autosegmentation discrepancies do not always correspond to clinical impact. [2] Even when clinical contours (CCs) and deep learning contours (DLC) are geometrically comparable, differences can be found when comparing the dose delivered to organs at risk (OARs) between the CC- and DLC-based treatment plans [3]. It is important to note, however, that when evaluating the dosimetric impact of autosegmentation errors, two distinct measures should be distinguished: the dosimetric impact on planning (DIP) and the dosimetric impact on evaluation (DIE). [4] These metrics can be interpreted as the impact of not correcting the DLC on treatment plan optimization and the impact on clinical evaluation of the dose parameters, respectively. While the DIP has been demonstrated to be small in previous studies [3,5], the DIE is, so far, not commonly addressed. The DIP and DIE introduce the task of making new treatment plans for autosegmentation evaluation. The proper assessment of the dosimetric impact on evaluation, however, is crucial since it influences clinical decision making. For example, in the model-based approach [6] to select patients for individualized treatment strategies (i.e. proton therapy or adaptive therapy), the accuracy of dose evaluation parameters resulting from an autosegmentation model is thus clinically relevant.

The aim of this study was to go beyond the use of only geometric metrics for autosegmentation evaluation and perform a comprehensive clinical evaluation of autosegmentation of head and neck DLC OARs. To this end, we compare DLC with CC and evaluate the geometric errors, the dosimetric impact on treatment planning, the dosimetric impact on plan evaluation and NTCP, while clarifying the clinical value of the different measures.

2 Materials and Methods

A total of 68 oropharyngeal cancer patients from a single center treated with primary radiotherapy (70 Gy in 35 fractions) between 2018 and 2022 were included. For all patients, DLC sets of OARs in the head and neck region were available (WorkflowBox 2.0, Mirada Medical Ltd., Oxford, United Kingdom) as well as manually adjusted CC sets. IMPT dose plans for both contour sets were automatically created in RayStation Development 10B (RaySearch Laboratories, Stockholm, Sweden) with a U-Net deep learning model.

A geometric comparison between the DLC and CC was performed for 11 head and neck OARs using the DSC and HD. For all OARs, the mean dose difference was reported as the dosimetric measure of interest. The dose differences were determined using the DIP and the DIE. The DIP was defined as the difference between plans optimized using the CC and DLC, evaluated on the CC. While the DIE was defined as the difference between evaluation on the CC and DLC using the plan optimized on the DLC, see Figure 1. To assess the DIE, an optimized dose distribution on the DLC is required. In a typical retrospective study this might not be available, therefore we also study to which extent the dosimetric impact approximation, the DIA, which is equivalent, but based on the optimized dose using the CC (see Figure 1), can be used as a substitute for the DIE.

NTCP-values for xerostomia and dysphagia, based on the mean dose on the parotid glands and mean dose on the oral cavity and superior pharyngeal constrictor muscle respectively, were chosen as clinically relevant metrics for

autosegmentation evaluation. A difference in mean dose of ≥2.0 Gy or NTCP-value of ≥1.0 percentage point (pp) was defined as clinically relevant by the head and neck radiation oncologist team. Statistical significance was assessed with a Wilcoxon signed-rank test. Correlation between geometric and dosimetric measurements was evaluated using scatter plots and the Spearman's Rank correlation coefficient.

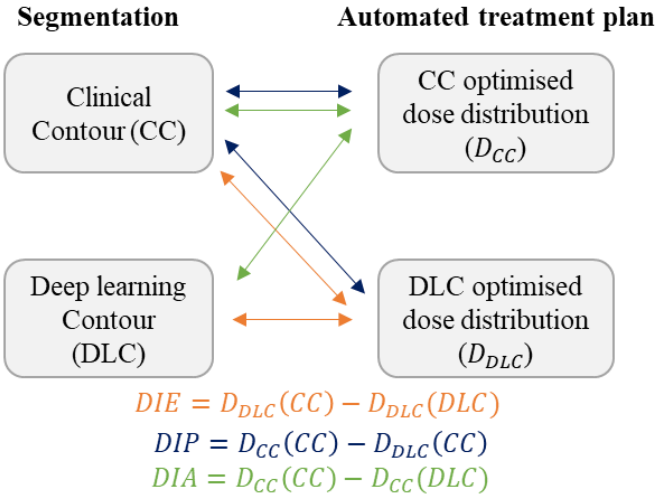


Figure 1: Schematic overview of DIP and DIE evaluation. DIP = dosimetric impact on planning. DIE = dosimetric impact on evaluation. DIA = dosimetric impact approximation.
D_x(y) = Dose optimized using segmentation x, evaluated on segmentation y

3 Results

Geometric evaluation

Geometric evaluation was performed using the DSC and HD. The DSC scores ranged from 0.50-0.90, see Table 1. The HD ranged from 0.74-1.26 mm.

Dosimetric impact on planning vs dosimetric impact on evaluation

The mean dose differences based on the DIP, DIE and DIA can be found in Table 1. A significant difference in mean dose with the DIP analysis was found in 7 out of 11 OARs. However, in none of the OAR structures the mean dose difference was more than 2.0 Gy. For the DIE and DIA analysis, the mean dose differences were generally higher than with the DIP, however, never larger than 2.0 Gy. The OARs portray a predominantly negative dose difference value (i.e. overestimation of dose deposition on DLCs compared to CCs).

Impact on NTCP

The NTCP values for dysphagia and xerostomia using the DIP, DIE, and DIA mean doses can be found in Table 1. A significant difference in NTCP-value ≥1.0pp was found for xerostomia grade II when using the DIE and DIA analysis. The other mean NTCP-value differences were smaller than 1.0pp.

A)

	DIP $D_{CC}(CC) - D_{DLC}(DLC)$		DIE $D_{DLC}(CC) - D_{DLC}(DLC)$		DIA $D_{CC}(CC) - D_{CC}(DLC)$	
	NTCP difference in % ± std	*p-value*	*NTCP difference in % ± std*	*p-value*	*NTCP difference in % ± std*	*p-value*
Xerostomia grade II	-0.1 ± 0.2	p = 0.092	**-1.0 ± 6.1***	**p < 0.001**	**-1.1 ± 1.5**	**p < 0.001**
Xerostomia grade III	-0.0 ± 0.1	p = 0.121	-0.4 ± 1.8	p < 0.001	-0.4 ± 0.5	p < 0.001
Dysphagia grade II	-0.3 ± 0.8	p < 0.001	-0.4 ± 4.4	p = 0.302	-0.7 ± 2.9	p = 0.094
Dysphagia grade III	0.1 ± 0.6	p < 0.001	0.2 ± 1.2	p = 0.041	0.3 ± 1.2	p = 0.011

B)

	DIP $D_{CC}(CC) - D_{DLC}(DLC)$		DIE $D_{DLC}(CC) - D_{DLC}(DLC)$		DIA $D_{CC}(CC) - D_{CC}(DLC)$		Geometric comparison	
	Dose difference in Gy ± std	*p-value*	*Dose difference in Gy ± std*	*p-value*	*Dose difference in Gy ± std*	*p-value*	*DSC ± std*	*HD (mm) ± std*
Crico	-0.7 ± 1.0	p < 0.001	0.9 ± 3.4	p < 0.001	0.8 ± 3.2	p < 0.001	0.60 ± 0.2	0.85 ± 0.4
Cerv. oesophagus	-0.5 ± 1.2	p < 0.001	1.7 ± 3.7	p < 0.001	0.0 ± 3.3	p = 0.774	0.55 ± 0.2	1.26 ± 0.6
Glottic area	-1.5 ± 1.2	p = 0.001	0.9 ± 3.9	p < 0.001	0.7 ± 3.8	p < 0.001	0.50 ± 0.2	0.84 ± 0.5
Oral cavity	-0.1 ± 0.2	p < 0.001	-0.2 ± 1.2	p = 0.156	-0.3 ± 1.2	p = 0.057	0.90 ± 0.0	0.89 ± 0.3
PCM	-0.3 ± 0.9	p < 0.001	-1.0 ± 3.6	p = 0.030	-0.3 ± 5.9	p < 0.001	0.53 ± 0.1	1.07 ± 0.5
Parotid L	0.0 ± 0.2	p = 0.087	-1.6 ± 3.6	p < 0.001	-1.7 ± 3.6	p < 0.001	0.85 ± 0.1	1.21 ± 0.6
Parotid R	0.0 ± 0.2	p = 0.293	-1.5 ± 2.6	p < 0.001	-1.6 ± 2.6	p < 0.001	0.85 ± 0.1	1.19 ± 0.6
Submandibular L	-0.1 ± 0.6	p = 0.382	-0.5 ± 2.2	p = 0.285	-0.6 ± 2.4	p = 0.231	0.77 ± 0.1	0.79 ± 0.4
Submandibular R	-0.1 ± 0.4	p = 0.423	-0.9 ± 2.1	p < 0.001	-0.9 ± 2.1	p < 0.001	0.76 ± 0.1	0.74 ± 0.5
Supraglottic area	-0.7 ± 0.7	p < 0.001	0.4 ± 3.4	p = 0.050	0.3 ± 3.6	p = 0.105	0.70 ± 0.1	0.92 ± 0.5
Thyroid	-0.6 ± 0.5	p < 0.001	0.0 ± 0.8	p = 0.662	-0.2 ± 1.0	p = 0.368	0.77 ± 0.1	1.00 ± 0.5

Table 1: Geometric, dosimetric and predicted clinical impact evaluation between the DLC and CC OARs evaluated with the DIP and DIE analysis. A) Mean NTCP differences for xerostomia and dysphagia in percentage point (pp). B) Mean dose difference and geometric comparison per OAR. D_x(Y) indicates the dose, optimized using x and evaluated on y.
*PCM = pharyngeal constrictor muscle. DSC = Sørensen-Dice similarity coefficient. P-value obtained with a Wilcoxon signed-rank test. * NTCP difference ≥ 1.0 pp and statistically significant*

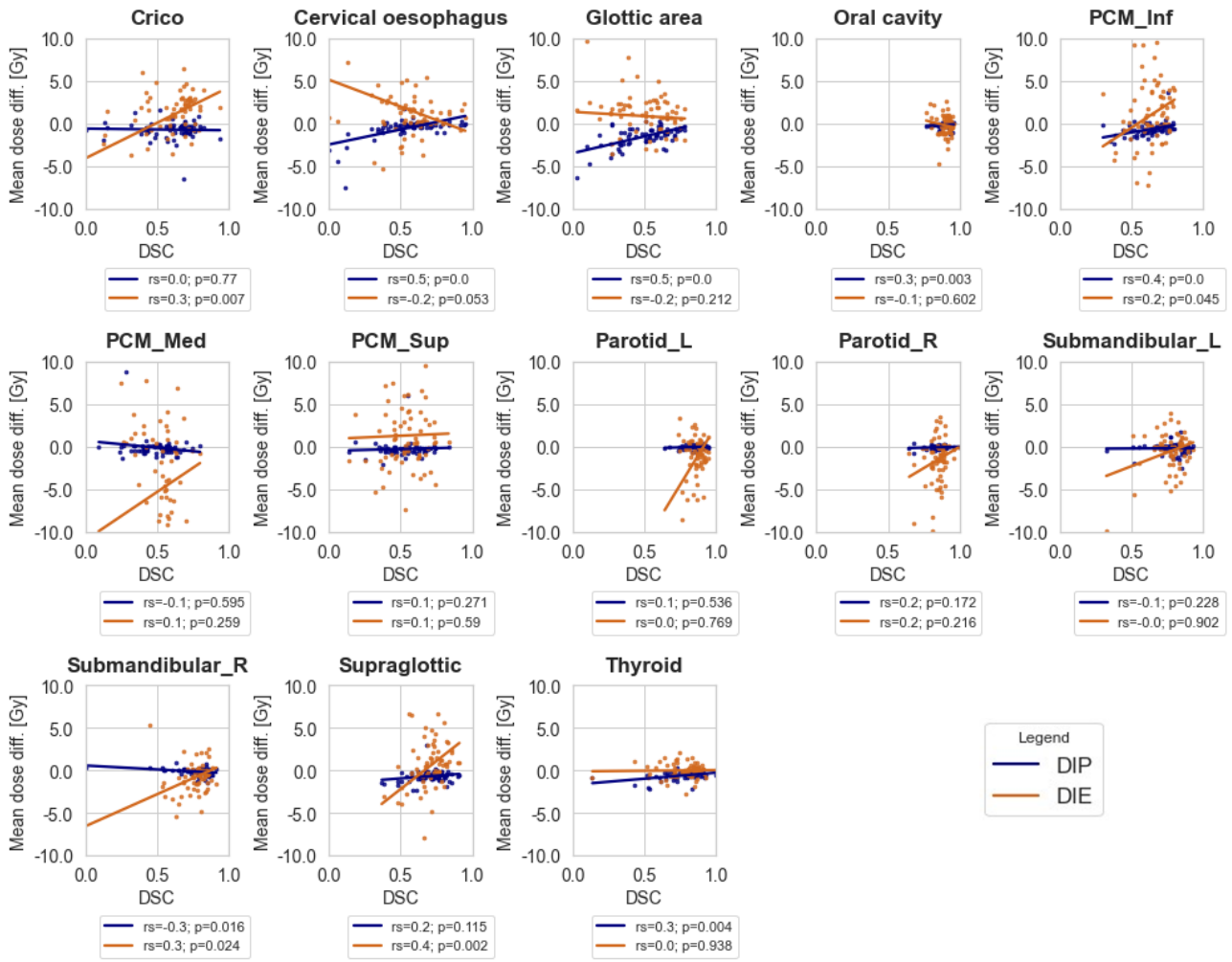


Figure 2: Geometric comparison (DSC) vs mean dose difference (diff.) (CC – DLC) using DIP, DIE and DIA per OAR with Spearman's Rank Correlation coefficient (rs) and corresponding p-value. A correlation coefficient > 0.5 or < - 0.5 with p < 0.05 is seen as a correlation. PCM = pharyngeal constrictor muscle. L = left. R = right. Please note the uneven grid spacing between the

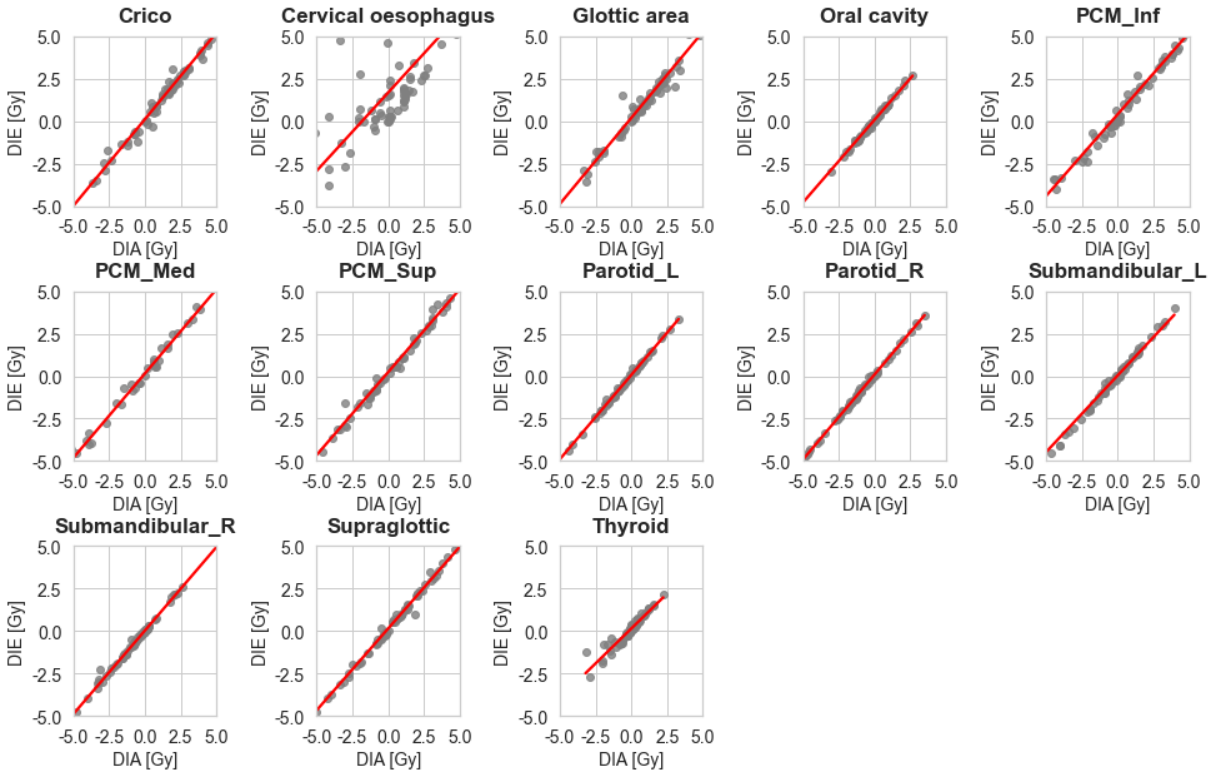


Figure 3: Dosimetric impact on evaluation (DIE) vs dosimetric impact approximation (DIA) mean dose difference (CC-DLC) per OAR.. PCM = pharyngeal constrictor muscle. L = left. R = right.

Measurement comparison
A comparison per OAR between the DIP and DIE relative to the DSC can be found in Figure 2. No significant correlation is found between the mean dose difference and DSC for all OARs.

Segmentation evaluation using retrospective clinical plans
The correlation between the DIE and DIA is shown in scatterplots per OAR in Figure 3. All OARs show a strong correlation.

4 Discussion

The impact of potential autosegmentation inaccuracies on dosimetric measurements in treatment planning (DIP) and treatment plan evaluation (DIE) was limited, with only Grade II xerostomia resulting in a clinically significant difference of 1.0pp. No significant correlation between the DSC and mean dose difference was found (Table 1 and Figure 2). This indicates that the extent to which the DSC is a meaningful metric to evaluate the clinical acceptability of the DLC is limited at best.

Furthermore, analyses of the DIP and DIE show that they are two distinct measures evaluating two different ways in which autosegmentation inaccuracies have a dosimetric impact. On average, the mean dose difference for the impact on plan evaluation (DIE) is larger than the impact on the optimized treatment plan (DIP), the standard deviation for the mean dose difference is larger with the DIE (Table 1) and the correlation with the Dice score is unalike between the DIP and DIE (Figure 2). The dosimetric analysis shows that the autosegmentation quality of this DL model is sufficient for automated treatment planning as the quality is not influenced by it, but that the autosegmentation quality is not satisfactory for treatment plan evaluation purposes. This is also shown in the clinical measurements where an NTCP-value of 1.0pp was seen for grade II xerostomia using the DIE analysis while for the DIP all NTCP values were below 1.0pp.

A possible explanation for the distinct results between the DIP and DIE can be found in the way dose plans are optimized using mean dose criteria. This is possibly less sensitive to the 3D dose distribution: presumably because if an OAR is in the beam's line of sight, the dose fluence coming through that line of sight will be lowered in the optimization. The precise shape of the OAR has only a limited impact on how the dose fluence is modulated. On the other hand: the impact on evaluation can be quite sensitive e.g. to volume changes of the OAR along dose the steep dose gradients. Thus, conceivably leading to larger mean dose difference when using the DIE compared to the DIP.

We also found that there is a strong correlation between the DIE and DIA for all OARs and that the DIA could thus be used as a substitute for the DIE evaluation. This would alleviate the workload for a retrospective dosimetric evaluation on autosegmentation, as no new treatment plan for the DLC has to be created for this analysis.

5 Conclusion

For all OARs the geometric quality of autosegmentation was found to be uncorrelated to the clinical impact measures. This indicates the importance of a comprehensive analysis beyond sole geometric evaluation for autosegmentation. The mean dose and NTCP differences when evaluating the dosimetric impact on treatment plan optimization (DIP) are lower compared to the impact on treatment plan evaluation (DIE). Indicating that both dosimetric measurements are essential in a clinically relevant evaluation of autosegmentation as they encompass distinct but both clinically important aspects of autosegmentation quality. The DIE might be approximated by using the dose differences between the CC and DLC in the retrospective analysis of clinical plans (DIA), reducing the workload of autosegmentation evaluation.

References

[1] L. J. Isaksson, P. Summers, F. Mastroleo et al. "Automatic Segmentation with Deep Learning in Radiotherapy". Cancers 15.17 (2023), p. 4389. DOI: 10.3390/cancers15174389

[2] M. V. Sherer, D. Lin, S. Elguindi, et al. "Metrics to evaluate the performance of auto-segmentation for radiation treatment planning: A critical review". Radiotherapy and Oncology 160 (2021), pp. 185-191. DOI: 10.1016/j.radonc.2021.05.003

[3] M. Costea, A. Zlate, M. Durand, et al. "Comparison of atlas-based and deep learning methods for organs at risk delineation on head-and-neck CT images using an automated treatment planning system". Radiotherapy and Oncology 177 (2022), pp. 61-70. DOI: 10.1016/j.radonc.2022.10.029

[4] P. J. González, R. Simoões, K. Kiers et al. "Explaining the dosimetric impact of contouring errors in head and neck radiotherapy". Biomedical Physics and Engineering Express 8.5 (2022), p. 055001 DOI: 10.1088/2057-1976/ac7b4c

[5] W. van Rooij, M. Dahele, H. Ribeiro Brandao, et al. "Deep Learning-Based Delineation of Head and Neck Organs at Risk: Geometric and Dosimetric Evaluation". International Journal of Radiation Oncology 104.3 (2019), pp. 677-684. DOI: 10.1016/j.ijrobp.2019.02.040

[6] J. A. Langendijk, F. J. P. Hoebers, Martin A. de Jong, et al. "National Protocol for Model-Based Selection for Proton Therapy in Head and Neck Cancer". Internation Journal of Particle Therapy 8.1 (2021), pp. 354-365. DOI: 10.14338/IJPT-20-00089.1

Team work makes the dream work: An ensemble learning approach for segmentation with Convolutional Neural Networks and Vision Transformers

Anne Andresen[1, 2], Yasmin Lassen-Ramshad[2], Christian Rønn Hansen[4, 5, 6], Slavka Lucakova[1, 3], Nadine Vatterodt[1, 2], and Jesper Kallehauge[1, 2]

[1]Department of Clinical Medicine, Aarhus University, Aarhus, Denmark
[2]Danish Center For Particle Therapy, Aarhus University Hospital, Aarhus, Denmark
[3]Department of Oncology, Aarhus University Hospital, Aarhus, Denmark
[4]Department of Oncology, Odense University Hospital, Odense, Denmark
[5]Laboratory of Radiation Physic, Odense University Hospital, Odense, Denmark

Abstract

Introduction
Recent developments in segmentation focus on enhancing accuracy and efficiency. Convolutional Neural Networks (CNN), especially the UNet architecture, has been applied for segmentation, however lately Vision Transformers ViT models have achieved high accuracy for segmentations. Therefore using a CNN UNet and ViT in ensemble learning could benefit from the different properties, to enhance a models' capabilities of handling variations in organ sizes, while providing high quality segmentations on smaller datasets.

Method and Materials
nnUNet and a ViT model were trained individually to perform segmentation of 8 organs at risk (OAR) in the brain. The dataset included 55 T1 contrast enhanced MRI scans split with 49 used for training and 6 for testing. Ensemble learning was applied to the models from fold 4 to generate the final segmentation. The evaluation compared the predicted segmentations and manual delineations provided by oncologists, using dice score, 1mm normalized surface dice and hausdorff distance 95th percentile.

Results
1mm normalized surface dice medians was in the range of 0.90 to 0.94, the dice in the range of 0.85-0.97 and hausdorff 95th percentile ranged from 1.0 to 2.0 across 8 organs at risk,

Discussion
The results suggest that ensemble learning of different small deep learning models could contribute to increasing accuracy when having a smaller dataset size and class imbalance however additional evaluation on a new dataset is needed.

1 Introduction

With the recent advancement in artificial intelligence, in particular in deep learning, more advanced models are being proposed to obtain highly accurate and clinically acceptable segmentations. Traditionally, this is achieved by increasing a single models' complexity, which creates a reliance on the availability of large datasets as the size of the dataset impacts the performance of the segmentation model.[1] This is a challenge in radiotherapy as data is often limited resulting in a small dataset sizes.

Additionally a challenge called class imbalance also arises, which can induce biases in model performance, when the model struggles to adequately learn the underrepresented class. So a limited dataset size coupled with class imbalance presents obstacles for acquiring a high accuracy and generalizability of segmentation models. To enhance model performance and address these challenges effectively, several techniques could be employed.

Therefore we apply ensemble learning for two distinct, separately trained, deep learning models: a Convolutional Neural Network (CNN) UNet and a Vision Transformer (ViT). We hypothesize that this will reduce the impact of the limited dataset size and class imbalance for multi organ segmentation on MRI scans of patients with brain cancer.

1.1 Related works

Significant advancements have been made in the field of medical image segmentation in recent years, to improve accuracy and efficiency of the models . CNNs have frequently been proposed for segmentation, with UNets as a widely employed architecture. CNN UNets has been successful in delineating organs at risk(OARs) in various areas of the body including in the brain. [2] [3] [4] [5] [6].

More recently ViT models have been used for segmentation. [1]. ViTs have been shown to provide highly accurate segmentations of small structures which have been attributed to their ability to capture global context and relationships within the images. [1] [7] [8]

These notable advancements observed in the individual models, prompted a suggestion to integrate the properties of CNNs and transformers. This could use the diverse model properties for capturing both local and global information, thereby enhancing the models' ability to handle variations in organ sizes [9] [10]. For brain structures such architectures have been employed, to enhance segmentation accuracy of OARs, including for brain structures where DSCs of 0.96 has been achieved for larger organs, however this on significantly larger datasets. [1]. More commonly DSCs of 0.70-0.89 are acquired on smaller datasets. [11] [12]

While advancements in medical image segmentation are evident, challenges such as limited dataset sizes and achieving consistently high segmentation accuracy independent of organ size, remain [13] [14]

2 Materials and Methods

2.1 Data

Patient demographics of the patients included in the study is presented in table 1

Median Age (range)	47.0 (35.5 - 59.5)
Gender	Male: 53 % , Female : 47 %
Tumor type:	glioma , meningioma

Table 1: Patient demographics

The available dataset for this study consists of 55 T1-weighted contrast-enhanced (T1w CE) Magnetic Resonance Imaging (MRI) scans with corresponding manually delineated structures of 8 OARs of patients with brain cancer. The 8 OARs manually delineated include

- Brainstem
- Chiasm
- Hippocampus L/R
- Optic nerves L/R
- Optic tract L/R

49 of the patients were allocated as training data with the remaining 6 patients for test.

2.2 Preprocessing

Initial preprocessing consisted of z-score normalization, to standardize the intensity values across all scans. Subsequently, cropping a central region with boundaries based on the label mask was performed to reduce the computational load.

2.3 Ensemble learning

Ensemble learning was applied to integrate the predictions of two individually trained weak models, to collectively improve quality and generalization via a strong model. This approach allows for fine-tuning of individual models, emphasizing their distinctive strengths of smaller compact models which are well-suited for optimization of the grounds of smaller datasets while enhancing accuracy.
The initial weak model in the ensemble is a UNet architecture using CNN, specifically nnUNet, while the second weak model in the ensemble adopts ViTs trained using an identical dataset to collectively create a strong model which provides final segmentation of the 8 OARs as presented in figure 1.

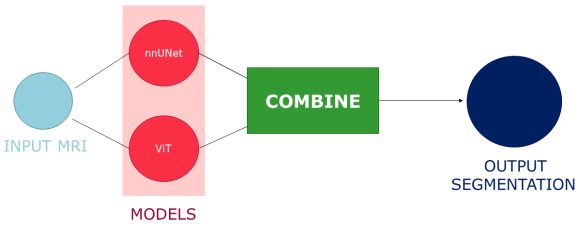


Figure 1: Representation of ensemble learning

2.4 Evaluation

For assessment of the proposed ensemble model, a set of evaluation metrics representing the volumetric and distance accuracy was employed. These metrics encompass the 1mm normalized surface Dice (1NsDSC), Dice score (DSC), and the 95th percentile of the Hausdorff distance (HD95). Using these metrics, we assess the similarity between the predicted segmentations from the ensemble model and the manually delineated OARs provided by an oncologist.

3 Results

3.1 Dice results

The median DSC values ranged from 0.85 to 0.97 across all segmented structures. The lowest median DSC was observed for the right optic tract see figure 2, while the highest obtained median DSC was for the brainstem.

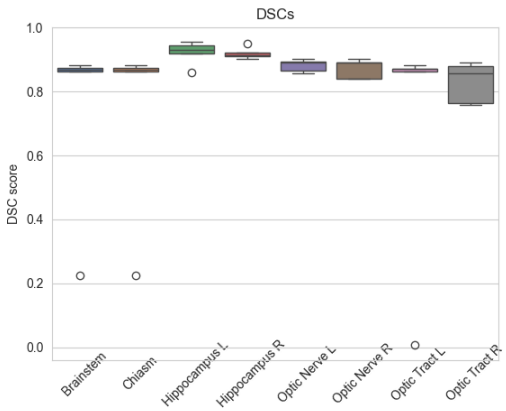


Figure 2: DSCs for the ensemble mode for 8 OARs

3.2 1 mm normalized surface dice results

The 1NsDSC, demonstrated median scores within the range of 0.90 to 0.94, with the lowest medians being for the left hippocampus and the highest being the brainstem

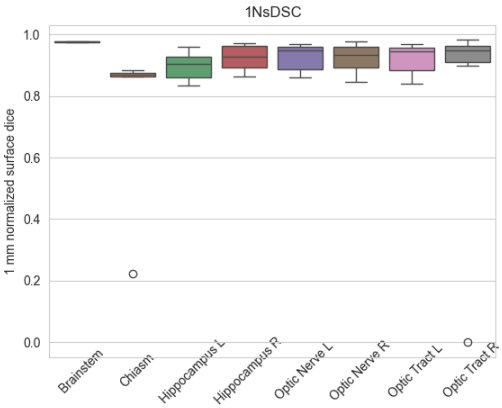


Figure 3: The ensemble models' 1 mm normalized surface dice for 8 OARs

3.3 HD95 distance

The HD95 distance metric, medians ranging from 1.0 to 2.0, with the right hippocampus displaying the lowest HD95 distance and the left optic tract showcasing the highest.

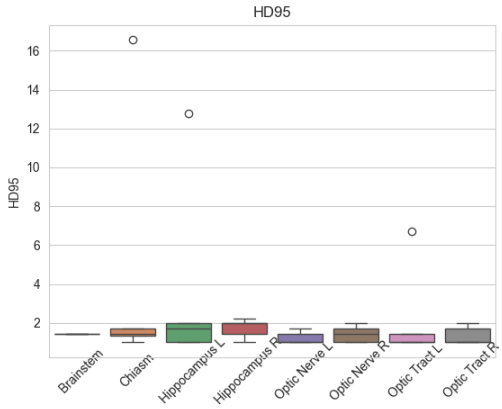


Figure 4: Hausdorff distance 95th percentile for the ensemble model across 8 OARs

4 Discussion

4.1 Results

The ensemble learning approach, combining the strengths of UNet and ViT models suggest a plausible approach when working on smaller datasets. The consistently high DSC and 1NsDSC, coupled low HD95 distances, suggest a high efficacy of the ensemble models' ability to achieve precise segmentations across the diverse set of OARs. Furthermore, these findings indicate a high capacity to localize small anatomical structures in challenging regions, which could indicate that combining various types of deep learning models in such way could contribute to mitigate the class imbalances otherwise present in the dataset.

Assessing the effectiveness of the ensemble approach can be done through a comparative assessment of the ensemble model and the weaker models contributing to the final strong model. Effectiveness is assessed by comparison of the evaluation metrics, specifically the 1NsDSC and HD95, which was used in the evaluation of the ensemble model performance. Examination of the 1NsDSC scores, for the nnUNet, medians lie in the range of 0.77 to 0.89 and 0.85-0.90 for the ViT. Compared to the medians of 0.90-0.94 for the ensemble, it suggests that the ensemble model's ability to define organ boundaries did improve when compared to its individual weak models. For the HD95, the ViT and nnUNet models showed medians in the range of 3.3-1.9 and 4.3-1.0, respectively, whereas the the ensemble model demonstrated distances within the range of 2.0- 1.0. This coupled with the DSC scores results suggest an improvement when utilizing the ensemble to create the final segmentations
The ensemble has notable enhancements in accuracy, as indicated by the evaluation metrics, which suggest a potential of ensemble modeling as a robust strategy for improving medical image segmentation tasks when having small datasets containing class imbalance.

4.2 Model

The decision to employ ensemble learning, combining the strengths of CNN UNet and ViT models, was motivated by their ability distinct features and patterns due to their different architectures. The UNet provided local information, while the ViT ensures a global oversight of the entire image. By using both models, we aimed to exploit the unique capabilities of each model type for more accurate segmentation. Training on the same dataset, the models were provided with a common understanding of the underlying patterns in the dataset which contributed to the consistency of the ensemble model. Meanwhile the likelihood of overfitting can be reduced by distribution of parameters between two models with individualized training sessions.

4.3 Clinical Significance

The consistently high performance across multiple metrics suggests a strong potential for clinical use. Accurate segmentation of OARs is important in treatment planning of patients with brain cancer, and the ensemble models are promising. However, additional clinical qualitative evaluation should be performed.

4.4 Limitations and Future Directions

Result obtained in this study rely on the quality of the training dataset. Future work could explore the impact of dataset variability and the generalization capacity of the models for a larger patient cohort in the test phase. Additionally, the proposed ensemble architecture opens the possibility for additional fine-tuning and optimization on smaller datasets for an enhancement performance on other modalities. Finally, a

clinical assessment to determine the quality of the segmentations in a clinical setting is needed.

5 Conclusion

In conclusion, the ensemble learning approach integrating UNet and ViT models seems promising for the segmentation of OARs in the brain. The obtained results underscore the potential clinical significance of the proposed methodology and contribute to the ongoing discourse on using diverse model architectures for enhanced medical image segmentation. As the field continues to evolve, this work sets the stage for further exploration, refinement, and application in real-world clinical scenarios.

References

[1] J. Ma, Y. He, F. Li, et al. "Segment anything in medical images". *Nature Communications* 15.1 (2024), p. 654.

[2] E. Orasanu, T. Brosch, C. Glide-Hurst, et al. "Organ-at-risk segmentation in brain MRI using model-based segmentation: benefits of deep learning-based boundary detectors". *International Workshop on Shape in Medical Imaging*. Springer. 2018, pp. 291–299.

[3] Z. Akkus, A. Galimzianova, A. Hoogi, et al. "Deep learning for brain MRI segmentation: state of the art and future directions". *Journal of digital imaging* 30 (2017), pp. 449–459.

[4] R. Azad, E. K. Aghdam, A. Rauland, et al. "Medical image segmentation review: The success of u-net". *arXiv preprint arXiv:2211.14830* (2022).

[5] N. Siddique, S. Paheding, C. P. Elkin, et al. "U-net and its variants for medical image segmentation: A review of theory and applications". *Ieee Access* 9 (2021), pp. 82031–82057.

[6] N. Salpea, P. Tzouveli, and D. Kollias. "Medical image segmentation: A review of modern architectures". *European Conference on Computer Vision*. Springer. 2022, pp. 691–708.

[7] A. Dosovitskiy, L. Beyer, A. Kolesnikov, et al. "An image is worth 16x16 words: Transformers for image recognition at scale". *arXiv preprint arXiv:2010.11929* (2020).

[8] K. Han, Y. Wang, H. Chen, et al. "A survey on vision transformer". *IEEE transactions on pattern analysis and machine intelligence* 45.1 (2022), pp. 87–110.

[9] H. Tang, Y. Chen, T. Wang, et al. "HTC-Net: A hybrid CNN-transformer framework for medical image segmentation". *Biomedical Signal Processing and Control* 88 (2024), p. 105605.

[10] R. Ranjbarzadeh, A. Bagherian Kasgari, S. Jafarzadeh Ghoushchi, et al. "Brain tumor segmentation based on deep learning and an attention mechanism using MRI multi-modalities brain images". *Scientific Reports* 11.1 (2021), p. 10930.

[11] D. N. Greve, B. Billot, D. Cordero, et al. "A deep learning toolbox for automatic segmentation of subcortical limbic structures from MRI images". *Neuroimage* 244 (2021), p. 118610.

[12] B. Lee, N. Yamanakkanavar, and J. Y. Choi. "Automatic segmentation of brain MRI using a novel patch-wise U-net deep architecture". *Plos one* 15.8 (2020), e0236493.

[13] D. Zhang, Y. Lin, H. Chen, et al. "Deep learning for medical image segmentation: tricks, challenges and future directions". *arXiv preprint arXiv:2209.10307* (2022).

[14] R. Wang, T. Lei, R. Cui, et al. "Medical image segmentation using deep learning: A survey". *IET Image Processing* 16.5 (2022), pp. 1243–1267.

Data harvesting vs data farming: A study of the importance of variation vs sample size in deep learning-based auto-segmentation for breast cancer patients

Emma Skarsø Buhl[1,2], Else Maae[3], Louise Wichmann Matthiessen[4], Mette Holck Nielsen[5], Maja Maraldo[6], Mette Møller[7], Stine Elleberg[1], Sami Aziz-Jowad Al-Rawi[8], Birgitte Vrou Offersen[9], and Stine Sofia Korreman[1,2]

[1]Danish Center for Particle Therapy, Aarhus University Hospital, Aarhus, Denmark
[2]Department of Clinical Medicine, Aarhus University, Aarhus, Denmark
[3]Department of Oncology, Vejle Hospital, University Hospital of Southern Denmark, Denmark
[4]Department of Oncology, Herlev and Gentofte Hospital, Herlev, Denmark
[5]Department of Oncology, Odense University Hospital, Denmark
[6]Department of Clinical Oncology, Rigshospitalet, Copenhagen University Hospital, Denmark
[7]Department of Oncology, Aalborg University Hospital, Aalborg, Denmark
[8]Department of Clinical Oncology and Palliative Care, Zealand University Hospital, Næstved, Denmark
[9]Department of Experimental Clinical Medicine, Aarhus University Hospital, Aarhus, Denmark

Abstract Large interobserver variations (IOV) are expected when delineating structures on a CT scan, especially small structures with low visibility, like the clinical target volume internal mammary lymph node chain (IMN). A solution to avoiding IOV is deep learning-based auto-segmentation (DL). However, to train a robust and well performing model, a large and high-quality data set is needed. A large data set can become available if using delineations from the daily clinical practice, however these can be prone to large IOV. A data set created specifically for the purpose of training a DL model allows for obtaining a lower IOV, however this also entails a small cohort, as it is time-consuming to create. The aim of this study was to investigate the difference in output, when training a model in three different scenarios: a large clinical delineated data set (with 700/78 patients for training/testing, from the Danish Breast Cancer Group (DBCG) RT Nation Study), a clinical but curated dataset (with 328/36 patients for training/testing, from the DBCG RT Nation Study) and a smaller, but dedicated data set created by delineation experts (with 123/14 patients for training/testing, consensus delineations created by delineation experts). The model performance was estimated based on the performance metrics dice similarity coefficient (DSC), Hausdorff 95th percentile (HD95) and mean surface distance (MSD). Models were tested in test sets from their own cohort, and afterwards also compared in the dedicated data test set. The difference between model output was finally estimated by measuring the mean width and cranial caudal length of the model output for the models. When testing the model output between the clinical models and the dedicated models in their own test set, the two clinical models had a poorer performance, than the dedicated models, but not all metrics showed statistically significance. When testing the models in the dedicated data, the dedicated model showed a slightly better performance, along with fewer segmentation outliers. As a way of taking advantage of the strength from both types of data set, it could be an option to use a large clinical data set as a baseline model, and then finetune with smaller sized cohorts with dedicated delineations.

1 Introduction

Yearly more than 4700 women are diagnosed with breast cancer in Denmark[1]. Most patients are operated and more than 80% of these are treated with post-operative adjuvant radiotherapy[2]. Node positive breast cancer patients are treated with internal mammary lymph node (IMN) irradiation to improve overall survival[3,4]. Constraints to organs at risk (OAR), such as the heart, lung and contralateral breast, often call for a clinical compromise during treatment planning. The radiation treatment plans are optimized based upon manual structure delineations on CT scans. The IMN target is difficult to see on CT scans[5], which may introduce interobserver variation (IOV), despite strict national and international delineation guidelines[5,6]. Recently, DL has been studied as a way of mitigating IOV. For OAR, DL has been shown to work well [7], but for target structures there are still challenges. In order to achieve a good performance with DL models, large data sets of high quality are generally needed. But, achieving both high quantity and high quality in these data sets is a challenge, as delineation is time-consuming.

In the radiotherapy community, the data sets for DL models are often achieved through data farming - a single expert or a small group delineates a dedicated data set, to minimize the IOV. This is only feasible for small data sets due to the limit on clinicians' time, and thus hampers the size of the cohort. This further decreases the generalizability of the model, as the limited size training data set lacks diversity. One solution to obtaining large data sets could be to use the clinical delineations retrospectively. This will not only allow for larger data sets, but also include a variation of patient anatomy, setup scenarios and CT scanners for increased generalizability. Clinical delineations reflect daily high workload and a larger variation in delineation quality is therefore expected compared to dedicated data sets created by a single or group of experts strictly following the same delineation guidelines. The Danish Breast Cancer Group (DBCG) has established the DBCG RT Nation database, containing CT scans and structure sets from more than 7500 Danish high-risk breast cancer patients, who received radiotherapy in 2008-2016[8]. In addition, for a small subset of the DBCG RT Nation data set dedicated edited delineation of target structures has been generated by the DBCG following a target delineation consensus workshop.

The primary aim of this study was to investigate the difference in model performance when training a DL based model for the IMN target structure on clinical delineations vs dedicated delineations.

2 Materials and Methods

The study included left-sided breast cancer patients, treated in 2015-2016 from the DBCG RT Nation study, as this was the period after implementation of national delineation guidelines[5,6]. In total 778 patients were included (data_CLIN). The patients were treated with adjuvant loco-regional radiotherapy at the seven radiotherapy centers in Denmark, generating a variance in scanner and set-up protocols in the data. Depending on tumor location, it is clinical practice to delineate the IMN to intercostal room (IC) 3 or IC 4.

The full CT scans and IMN delineations were used as input in a 3D full-resolution nnUNet [9], and the data set was randomly split into a training set (90%) and test set (10%). Five-fold cross validation was used with default parameters and the number of epochs was varied depending on cohort size, to avoid overfitting. Before training of the two models, the DICOM files were converted into NIfTI files with the Python package dcmrtstruct2nii [10]. Three different models were trained. The first model was trained on the full cohort of 778 clinical IMN delineations (model_CLIN), with no data curation. The model was trained with 800 epochs. For the second model, the data was crudely curated to only encompass IMN delineations extending to IC 3, to lower the variance in the data set occurring due to differences in inclusion of IC 3 and IC 4 (data_CURATED_CLIN). The model was trained with 600 epochs (model_CURATED_CLIN). The third model was trained on a smaller, but dedicated data set (data_DED). The data_DED was created from the data_CLIN. Of the 778 left-sided breast cancer patients, 143 were used to create the data_DED. The patients were chosen to be equally distributed from the seven radiotherapy centers in Denmark and equally distributed between lumpectomy or mastectomy. A national target delineation workshop was held in October 2022 by the DBCG and the Danish Comprehensive Cancer Center, where 21 participants from all RT centres in Denmark discussed the national delineation guidelines [5,6]. Following the workshop days, the participants were asked to correct the 143 clinical delineations (6-7 each), so they followed the national delineation guidelines. The model_DED was trained with 300 epochs. Because the dedicated data set was relatively small, as an extra robustness test, the training and test set split was performed arbitrarily additionally two times with the data_DED. This allowed for an extra two versions of model_DED (model_DED_second_split and model_DED_third_split) and two extra dedicated test sets, to test the performance of model_CLIN and model_CURATED_CLIN. To estimate and compare the IOV in the five training cohorts, the mean width and cranial caudal length were calculated for all delineations in the training cohort. The width was calculated by measuring the longest distance between each coordinate set, and subsequently the mean width was calculated by averaging over all slices. The cranial caudal distance was measured from the most cranial slice to the most caudal slice. The measurements were conducted on the NIfTI masks.

The performance of the five models was first evaluated in their own test-set with the performance metrics: the dice similarity coefficient (DSC), the Hausdorff distance 95th percentile (HD95) and the mean surface distance (MSD). The statistical significance was tested with the Mann-Whitney U-test, assuming statistical significance level at $p<0.05$. As the variation in model training data might be reflected in the testing data as well, the model_CLIN and model_CURATED_CLIN was tested in the three different data_DED test sets, to estimate the performance of all three in a low IOV data set. In addition, the mean width and the cranial caudal length of the three model_DED, model_CLIN and model_CURATED_CLIN segmentations in the model_DED test sets were measured and compared. The statistical significance was tested with the Wilcoxon signed-rank test, assuming statistical significance level at $p<0.05$.

	DSC (mean) range	HD95 [mm] (Mean) range	MSD [mm] (mean) range
Model_CLIN (own test set, n = 78)	0.65 *,+, # (0.46 – 0.80)	11 *, # (2.30 – 125.6)	2.55 *,+,# (0.59 – 19.8)
Model_CLIN (dedicated test set first split , n = 14)	0.71 (0.57– 0.84)	6.54 (1.95–25.8)	1.37 (0.42– 3.49)
Model_CLIN (dedicated test set second split, n = 14)	0.66 (0.54 – 0.84)	10.2 (1.95– 32.0)	2.14 (0.42–7.97)
Model_CLIN (dedicated test set third split, n = 14)	0.69 ¤ (0.58 – 0.81)	8.69 (2.34– 28.4)	1.65 ¤ (0.48–6.10)
Model_CURATED_CLIN (own test set, n = 36)	0.66 # (0.36 – 0.80)	11.8 # (2.16-43.2)	3.00 *,# (0.54 – 11.8)
Model_CURATED_CLIN (dedicated test set first split , n = 14)	0.69 (0.52 – 0.81)	8.60 (2.62– 23.4)	1.20 (0.58–2.33)
Model_CURATED_CLIN (dedicated test set second split, n = 14)	0.64 ¤ (0.51 – 0.78)	13.3 (3.00– 30.0)	1.65 (0.39–6.16)
Model_CURATED_CLIN (dedicated test set third split, , n = 14)	0.67 ¤ (0.53 – 0.77)	9.39 ¤ (2.54– 33.5)	1.05 (0.51–2.22)
Model_DED (dedicated test set first split , n = 14)	0.71 (0.61-0.86)	7.58 (1.95-15.4)	1.41 (0.30-3.31)
Model_DED_second_split (dedicated test set second split, n = 14)	0.71 (0.57-0.83)	5.70 (1.95-14.0)	1.20 (0.54-2.44)
Model_DED_third_split (dedicated test set third split, n = 14)	0.74 (0.62-0.85)	5.30 (2.00-18.4)	0.96 (0.45-2.64)

Table 1. The performance of model_CLIN, model_CURATED_CLIN and model_DED in their respective test set and in the three dedicated test set splits (n=14). The performance metrics are the dice similarity coefficient (DSC), the Hausdorff 95th percentile (HD95) and the mean surface distance (MSD). Statistical significance between clinical models tested in their own test and model_DED is denoted with * for the first split, + for the second split and # for the third split. Statistical significance between clinical models tested in the dedicated test sets and model_DED tested in the corresponding split is denoted ¤. The third test is used as example in both figure 1 and 2.

3 Results

A total of 778 delineations were available from 2015-2016, leaving 700 for training the model_CLIN and 78 for testing the model_CLIN. During curation of the IMN delineations a total of 442 delineations were excluded, leaving 328 for training the model_CLIN_CURATED and 36 for testing the model_CLIN_CURATED. In the data_DED six delineations were excluded, due to the IMN extending to IC 2 or IC 4, leaving 123 for training the model_DED and 14 for testing the model_DED. In the training cohort, the data_CLIN had a mean width = 14.4 mm and a mean cranial caudal length of 108.2mm, the data_CURATED_CLIN had a mean width = 13.5mm and a mean cranial caudal length of 102.1mm and the data_DED (first/second/third) with a mean width = 14.5 mm/14.4mm/14.4mm and mean cranial caudal length of 107.8mm/108.1mm/108.0mm, figure 1.

Statistically significant difference was found for all except the difference in cranial caudal length between data_CLIN and data_DED and cranial caudal length and width between data_CLIN and data_DED_second_split and data_DED_third_split.

In their respective test sets, the two clinical models generally had poorer performance (mean DSC 0.65-0.66, HD95 11-11.8, MSD 2.55-3.00) than the three model_DED (mean DSC 0.71-0.74, HD95 5.3-7.58 MSD 0.96-1.41), however not all metrics showed statistical significance, table 1.

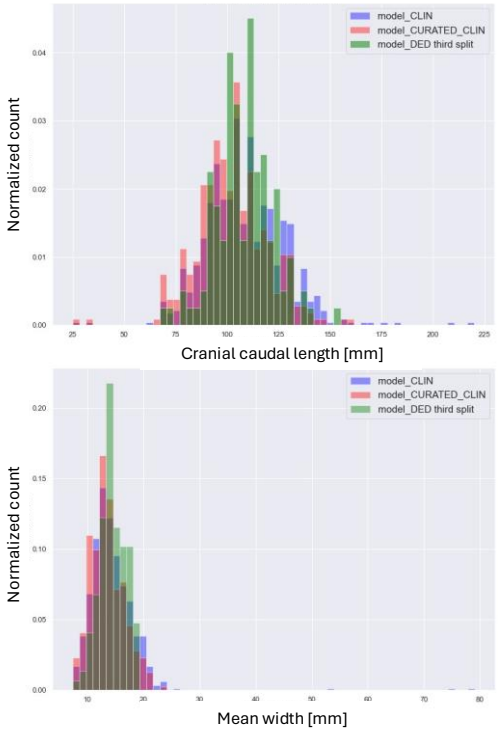


Figure 1: (top) The cranial caudal length of the delineations in the training cohort of the model_CLIN(blue), model_CURATED_CLIN (orange) and the third split of model_DED (green). (bottom) the mean width of the delineations in the training cohortsThe counts are normalized.

The difference between the performance of the two clinical models tested in the three dedicated test sets and the model_DED varied with the different training/test splits, table 1. A large range in both HD95 and MSD was seen in the two clinical models (MSD range 0.42-7.97 and 0.39-6.16) when tested on the dedicated test sets, along with an overall better performance of the three model_DED (MSD range 0.54-2.44), but not all were statistically significant, table 1.

The mean width of the CTVn_IMN with the model_CLIN in the third dedicated test set was 12.9mm and the mean cranial caudal length was 115.7mm and the mean width with the model_CURATED_CLIN was 12.3mm and the cranial caudal length was 104.7mm. The mean width of the CTVn_IMN with the model_DED_third_split in the third dedicated test set split was 137.7mm and mean cranial caudal length was 112.1mm. The difference in cranial caudal length was found statistically significant between model_DED_third_split and model_CURATED_CLIN and the width was found statistically significant between both clinical models and the model_DED_third_split, figure 2 and table 2.

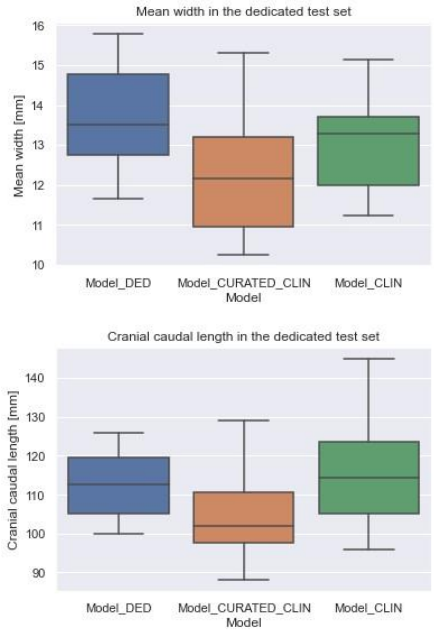


Figure 2: (top) The mean width in model output of the third split with model_DED (blue), model_CURATED_CLIN (orange) and model_CLIN (green) in the dedicated test - set. (Bottom) the cranial caudal length of the segmentations.

	Cranial caudal length [mm] (mean) range	Width [mm] (mean) range
Model_CLIN (dedicated test set first split, n = 14)	119.4 (102.0-150.0)	12.9 (11.0-14.8)
Model_CLIN (dedicated test set second split, n = 14)	113.3 (97.5-132.0)	13.3 * (11.0-16.0)
Model_CLIN (dedicated test set third split, n = 14)	115.7 (96.0-145.0)	12.9 * (11.2-15.1)
Model_CURATED_CLIN (dedicated test set first split, n = 14)	110.0 (86.3-144.0)	12.4 * (10.8-14.7)
Model_CURATED_CLIN (dedicated test set second split, n = 14)	101.3 * (85.0-129.0)	13.3 * (11.1-17.0)
Model_CURATED_CLIN (dedicated test set third split, , n = 14)	104.7 * (88.0-129.0)	12.3 * (10.2-15.3)
Model_DED (dedicated test set first split, n = 14)	115.1 (94.0-130.0)	13.4 (11.6-16.1)
Model_DED_second_split (dedicated test set second split, n = 14)	112.5 (92.5-136.0)	14.2 (12.2-16.7)
Model_DED_third_split (dedicated test set third split, n = 14)	112.1 (100.0-126.0)	13.7 (11.6-15.8)

Table 2. The mean width and cranial caudal length with the model_CLIN, model_CURATED_CLIN and the three model_DED tested in the three dedicated test sets. Statistically significant difference between the clinical models and the respective dedicated model is shown with *.

4 Discussion

In this study, we have investigated the difference in performance when using a clinical cohort, a curated clinical cohort, and a dedicated cohort to train DL-based auto-segmentation models for the left-sided CTVn_IMN. The differences among the three model predictions were minor and dependent on test set choice. However, for geometrical shape metrics, statistically significant differences were seen between the clinical models and the three versions of model_DED. The optimal balance between high-quality delineations and a large cohort is difficult, and both factors contribute to the performance of a DL model. Normal practice in the radiotherapy community is data farming, by creation of a dedicated data set, which limits the cohort size. Studies have investigated how low it is reasonable to go in size of data set without losing the precision of the DL

model. In Henderson et al, it was shown that for OAR in head and neck cancer increasing the data set beyond 250 scans gave very little improvement in model performance[11]. However, in a recent study from our group, we showed that for heart segmentation, smaller training cohorts increased the risk of outlier segmentations, when testing it in a set with larger variety in anatomy[12].
In this present study, there is a difference in model performance metrics among the three models, however there is not a consistent statistically significant difference between the models. It is worth noting that the test sets with dedicated delineations were small (n=14) and therefore does not represent a large variety of patient anatomies. Two additionally splits of the train/test set was performed with data_DED as an extra robustness test. The three different data_DED test sets led to different results of performance for model_CLIN and model_CURATED_CLIN, showcasing the impact of the specific patients included in the test set for such a small number of patients[12]. It would be expected that a larger number of patients in the test set would give a better estimate of the robustness of the models. However, despite the small test sets, there is still evidence of generally poorer performance and larger spread in the range of performance metrics in the clinical models than in the dedicated models, table 1. As a way of taking advantage of the strength from both types of data set, it could be an option to use a large clinical data set as a baseline model, which can then be finetuned with smaller sized cohorts with dedicated delineations. This would give both the variety available for generalizability from the clinic, and the low IOV and high precision delineations from dedicated data generation. If this approach is proven feasible, it might also be a way to tackle the problems arising when the delineation guidelines are updated and there will be a need for an update of the DL model. The dedicated delineations were created by delineation experts from seven institutions and based on editing existing delineations used in the clinic. The choice of correcting clinical delineations instead of delineation from scratch aimed to lower the delineation burden on busy clinicians. However, presenting delineators with an existing delineation imposes a risk of creating a bias as compared to delineating from scratch. The variety of expert delineators also introduce some IOV. Nevertheless, it is expected that the delineations created in the dedicated process, are as precise as possible on a national scale. Only a quantitative comparison has been conducted on these models. For a performance check, a qualitative investigation with expert assessments would be needed, as studies have shown, that the quantitative metrics not always correlate with clinical acceptance[13].

5 Conclusion

Training of a model on a clinical data set, showed on average a good performance, showcasing the benefits of using a large data set. Furthermore, a model trained on a large clinical data set performed better than a model trained on a curated clinical data set. However, the models trained on a dedicated data set showed a slightly improved performance, along with fewer segmentation outliers as compared to the models trained on clinical delineations. A combination of both models may be a way to utilize the best of both worlds.

References

[1] Kræftens Bekæmpelse. Kræft i Danmark 2022. cancer.dk 2022. https://www.cancer.dk/dyn/resources/File/file/9/9699/1665127828/kraeft-i-danmark-2022-bog-singlepages-3udg.pdf (accessed January 26, 2024).

[2] DCCC Stråleterapi. Strålebehandling af brystkræft | DCCC Stråleterapi n.d. https://www.straaleterapi.dk/kraefttyper/brystkraeft/ (accessed November 18, 2022).

[3] Poortmans PM, Weltens C, Fortpied C, Kirkove C, Peignaux-Casasnovas K, Budach V, et al. Internal mammary and medial supraclavicular lymph node chain irradiation in stage I-III breast cancer (EORTC 22922/10925): 15-year results of a randomised, phase 3 trial. Lancet Oncol 2020;21:1602–10. https://doi.org/10.1016/S1470-2045(20)30472-1.

[4] Thorsen LBJ, Overgaard J, Matthiessen LW, Berg M, Stenbygaard L, Pedersen AN, et al. Internal Mammary Node Irradiation in Patients With Node-Positive Early Breast Cancer: Fifteen-Year Results From the Danish Breast Cancer Group Internal Mammary Node Study. J Clin Oncol 2022;40:4198–206. https://doi.org/10.1200/JCO.22.00044.

[5] Nielsen MH, Berg M, Pedersen AN, Andersen K, Glavicic V, Jakobsen EH, et al. Delineation of target volumes and organs at risk in adjuvant radiotherapy of early breast cancer: national guidelines and contouring atlas by the Danish Breast Cancer Cooperative Group. Acta Oncol 2013;52:703–10. https://doi.org/10.3109/0284186X.2013.765064.

[6] Offersen BV, Boersma LJ, Kirkove C, Hol S, Aznar MC, Biete Sola A, et al. ESTRO consensus guideline on target volume delineation for elective radiation therapy of early stage breast cancer. Radiother Oncol 2015;114:3–10. https://doi.org/10.1016/j.radonc.2014.11.030.

[7] Wong J, Fong A, McVicar N, Smith S, Giambattista J, Wells D, et al. Comparing deep learning-based auto-segmentation of organs at risk and clinical target volumes to expert inter-observer variability in radiotherapy planning. Radiotherapy and Oncology 2020;144:152–8. https://doi.org/10.1016/j.radonc.2019.10.019.

[8] Refsgaard L, Skarsø ER, Ravkilde T, Nissen HD, Olsen M, Boye K, et al. End-to-end framework for automated collection of large multicentre radiotherapy datasets demonstrated in a Danish Breast Cancer Group cohort. Physics and Imaging in Radiation Oncology 2023;27:100485. https://doi.org/10.1016/j.phro.2023.100485.

[9] Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat Methods 2021;18:203–11. https://doi.org/10.1038/s41592-020-01008-z.

[10] Phil T, Albrecht T, Gay S, Rasmussen ME. Sikerdebaard/dcmrtstruct2nii: v5 2023. https://doi.org/10.5281/zenodo.7705311.

[11] Henderson EGA, van Herk M, Osorio EMV. The impact of training dataset size and ensemble inference strategies on head and neck auto-segmentation 2023. https://doi.org/10.48550/arXiv.2303.17318.

[12] Skarsø ER, Refsgaard L, Saini A, Sloth Møller D, Lorenzen EL, Maae E, et al. Development of a national deep learning-based auto-segmentation model for the heart on clinical delineations from the DBCG RT nation cohort. Acta Oncologica 2023;0:1–7. https://doi.org/10.1080/0284186X.2023.2252582.

[13] Sherer MV, Lin D, Elguindi S, Duke S, Tan L-T, Cacicedo J, et al. Metrics to evaluate the performance of auto-segmentation for radiation treatment planning: A critical review. Radiother Oncol 2021;160:185–91. https://doi.org/10.1016/j.radonc.2021.05.003.

Development and external validation of a DL-based CTV segmentation model on a large multicentric dataset in the case of whole-breast radiotherapy treatment

Maria Giulia Ubeira-Gabellini*†[1], Gabriele Palazzo†[1], Martina Mori[1], Alessia Tudda[1], Luciano Rivetti[2], Elisabetta Cagni[3], Roberta Castriconi[1], Valeria Landoni[4], Eugenia Moretti[5], Aldo Mazzilli[6], Caterina Oliviero[7], Lorenzo Placidi[8], Giulia Rambaldi Guidasci[9], Cecilia Riani[1], Andrei Fodor[10], Nadia Gisella Di Muzio[10,11], Robert Jeraj[2,12], Antonella del Vecchio[1], and Claudio Fiorino[1]

[1]Medical Physics, IRCCS San Raffaele Scientific Institute, Milan, Italy
[2]University of Ljubljana, Faculty of Mathematics and Physics, Ljubljana, Slovenia
[3]Medical Physics Unit, Department of Advanced Technology, Azienda USL-IRCCS di Reggio Emilia, Reggio Emilia, Italy
[4]Istituto Nazionale dei tumori Regina Elena, Rome, Italy
[5]Department of Medical Physics, University Hospital, Udine, Italy
[6]Medical Physics Dept, University Hospital of Parma AOUP, Italy
[7]University Hospital, "Federico II", Naples, Italy
[8]Fondazione Policlinico Universitario A. Gemelli IRCCS, uosd medical physics and radioprotection, Rome, Italy
[9]UOC di Radioterapia Oncologica, Fatebenefratelli Isola Tiberina – Gemelli Isola, Rome, Italy
[10]Radiotherapy, IRCCS San Raffaele Scientific Institute, Milan, Italy
[11]Vita-Salute San Raffaele University, Milan, Italy
[12]Department of Medical Physics, University of Wisconsin, Madison, USA

Abstract In order to optimize the radiotherapy treatment and reduce toxicities, organs-at-risk (OARs) and clinical target volume (CTV) must be segmented. DL techniques have great potential to accomplish this task taking into account also segmentation variability. Thanks to the availability of a large single-institute data sample together with numerous multi-centric data, a reliable CTV segmentation model is feasible and possible to be validated. A preprocessing, aimed to uniform all the images, followed by a 3D-Unet able to segment both right and left CTV was implemented. The metrics used to evaluate the performance are the Dice similarity coefficient (DSC), the Hausdorff distance (HD) and its 95th percentile variant (HD_95). The segmentation model achieved high performance on the validation set (DSC: 0.901; HD: 20.9 mm; HD_95: 9.6 mm). Moreover, the model predicts contours more extended than clinical ones along the cranio-caudal axis in both directions. The same metrics applied on a large multicentric dataset showed an overall agreement and model transferability for all but one (Institute 9) center considered.

1 Introduction

A largely applied approach in the treatment of breast cancer consists in the sequence of breast-conserving surgery followed by radiotherapy. To optimize the radiotherapy treatment, organs-at-risk (OARs) and clinical target volume (CTV) must be segmented. This is a highly time-consuming task done nowadays manually and prone to large inter and intra-observer variability [1]. Thus, the interest toward automatic tools supporting OARs/CTV segmentation is rapidly increasing, although also if just few studies dealt with CTV breast segmentation using large cohorts [2]. Atlas-based methods were first considered [3], but Deep Learning (DL) based methods (e.g., ResNet, U-net, nnUnet) are preferred to train/test reliable models [4–7], particularly for soft tissue-based regions. This may help in optimizing and reducing the clinical workload in the radiotherapy workflow, also if particular attention must be given to reach high segmentation precision considering the direct influence on treatment planning. Moreover, in order to improve the consistency between manual and automatic segmentation, inter-observer variability has to be taken into account, task not reached by ATLAS based method, but by DL models. Very importantly, the inter-observer (and inter-Institute) variability in breast CTV segmentation is large [8, 9] and, to our knowledge, this issue has been considered just by few works [10] into the development of DL tools for auto-segmentation. Futhermore, the multi-institute external validation of a DL model on the CTV auto-segmentation was not described before.

The aim of this work is to create for the first time an advanced in-house DL 3D model able to automatically segment right and left CTV breast together thanks to the availability of a large patients' cohort treated all in a single institute (611 patients). Moreover, the model has been internally and externally validated on a multicentric cohort of 8 institutes (including Institute 3, the same center on which training was performed) on 978 patients.

2 Materials and Methods

The dataset (3D CT images) consists of 611 patients, who underwent breast-conserving surgery followed by radiotherapy (RT), of which 20% was used for model validation. Patients collected are respectively 332 for right-sided breast cancer (BC) and 279 for left-sided BC. The images were acquired through different scanners and thus have variable 3D size

*Correspondence: ubeira.mariagiulia@hsr.it
†The two authors contributed equally to this research work.

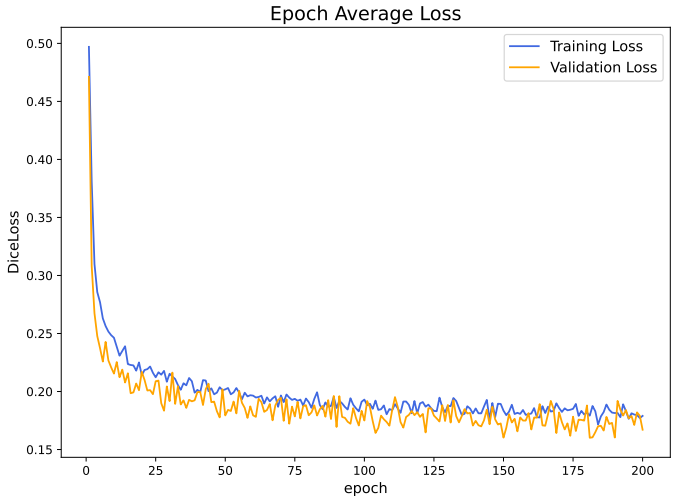


Figure 1: DCS loss over 200 epochs during the training and validation process. The model converges after 178 epochs. The data augmentation was applied to both training and validation for this plot.

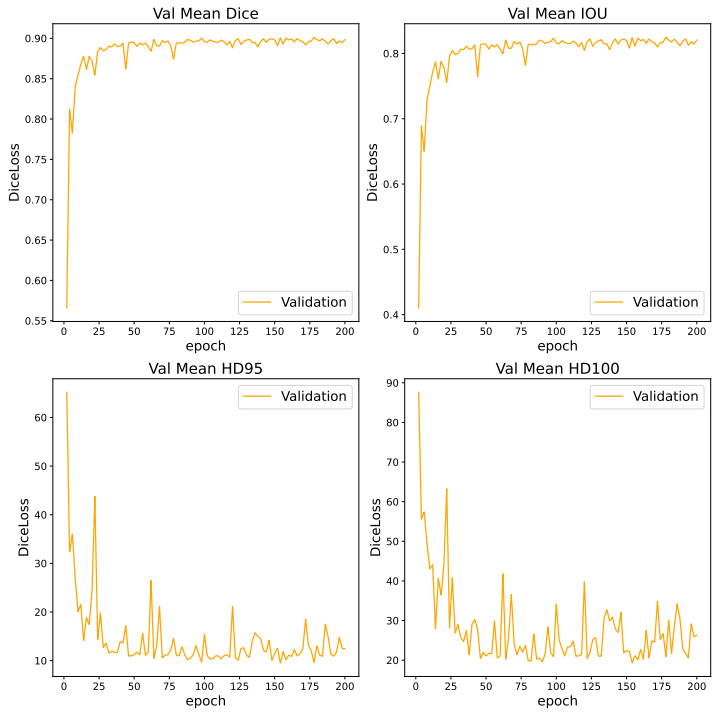


Figure 2: DCS, IOU, HD_95, HD metrics over 200 epochs during the validation process. The model converges after 178 epochs. The data augmentation was not applied to the validation set showed for this plot.

and resolution. CTV delineation was performed by different clinicians in the period 2017-2021. All patients signed an informed consent for treatment and permission for publication of disease-related information following the Helsinki Declaration. The institutional ethics committee approved the retrospective revision of patient outcomes and a prospective study of the quality of life of BCa patients treated with hypofractionation, registered to ClinicalTrials.gov (NCT03077191).

A pre-processing phase was firstly accomplished, aiming to make image format optimal to train a neural network (es. crop, resize, scaling). All images were re-sampled in order to have uniform images of 512*512*256 voxels (voxel space: 1*1*1.5 mm^3). To help in the CTV side identification (right or left) and to reduce the input parameters without reducing the voxel resolution, an automatic cropping of the left/right quadrant was applied, resulting in final images' dimension of 256*256*256 voxels. To do this, the xy coordinates of the body ROI center were computed for each slice along the cranio caudal axis. The median value of these coordinates were used as the lower left/right (depending on the side) corner of the crop rectangle for every slice with the addition of a fixed extra shift of 25 pixels, including more tissue in the image fed to the neural network. Moreover, to make the network properly working and to enhance the glandular and fat tissue intensity difference, a minimum of -200 HU and maximum of 100 HU were chosen and then scaled to be between 0 and 1.

Then, all images (i.e., with left/right BC) were used to train a global model. A 3D Unet, enriched with dropout and residual units, from the MONAI library (`https://zenodo.org/records/8436376`, v1.3). All procedures were implemented in-house and included in a library called libsrdl in mAItre, using python 3.10.9. The best parameters of the Unet were obtained after an hyperparameter optimization done through "optuna" [11]. Other architectures were tested (e.g. DynUnet, SegResNetDS), obtaining performances comparable or lower. Regions with rescaled intensity value of zero were removed to focus on the valid body area. Data augmentation (random flip on x, rotation, translation, scaling, Gaussian sharpening) was performed.

The segmentation performance of the validation dataset was addressed through Dice Similarity Coefficient (DSC), Intersection over Union (IOU), the Hausdorff distance (HD) and its 95th percentile variant (HD_95) metrics. Each 3D segmentation prediction was cropped in the cranio-caudal direction, a posteriori, to match the clinical segmentation, due to their expected large inter-observer variability. Finally, the predicted segmentation was co-registered back to the original image dimension needed in future for clinical purpose.

As external validation dataset, planning CTs from the MIKAPOCo (Multi Institutional Knowledge-based Approach to Plan Optimization for the Community, AIRC Grant IG23150, 248/2021) study were used, which includes 10 centers (eight of which were here analyzed). Within the MIKAPOCo activities already accomplished, plan prediction models were developed aiming to quantify inter-institute variability in plan prediction, suggesting high transferability for both right and left breast scenario [12]. Patients considered are respectively 526 for right-sided BC and 452 for left-sided BC, with a minimum number of patient per center of 54. Of these, 81 patients from

Institute 3 are here used as internal validation set within the MIKAPOCo dataset, not overlapping with the training set.
The preprocessing applied to the external validation dataset is the same applied to the training set and the global model was applied to the images for prediction. From the generated CTV prediction, quantitative features were extracted to assess the contours quality. The metrics chosen to compare between centers are DSC, HD, HD_95 and the difference of the volumes volumes difference (VD) of CTVs segmented by radiation oncologists and the predicted ones.
Considering that the model trained could also seldom segment the controlateral breast, patients with Hausdorff Distance values greater than 50 mm were not considered in this first stage. Follow-up studies will implement a post-processing pipeline where just the main continuum volume predicted will be considered, getting read of both cranio caudal and controlateral misprediction.

3 Results

The best 3D model obtained was able to reach, after 178 epochs, high metrics performance on model validation (DSC: 0.901; HD: 20.9 mm; HD_95: 9.6 mm, see Figure 1, 2), similar to what reached in [9]. Our network, however, is a 3D model and the predictions are smoother between slices, as it should be, well mimicking clinical practice. The predicted segmentation well matches the clinicians' contours on the skin and chest side, while it is more uncertain, as expected, on the lateral and cranio-caudal side (see Fig. 3). The cranio-caudal direction is also uncertain given the different strategies used in that regard by clinicians. In particular, the median absolute difference between predicted and observed segmentation is respectively 4.5 mm (IQR: 1.5-9.0) and 10.5 mm (IQR: 4.5-21.0) in caudal and cranial. They have thus a larger prediction along the cranio-caudal axis in both directions. This variability is also confirmed by running the same model with different random seeds for the dropout layer. Moreover, the predicted segmentation is found fully larger than the clinicians contouring (17 ml).
Finally, we performed prediction on the internal validation dataset (Institute 3) withing the MIKAPOCo set, obtaining on average values similar to model validation (DSC: 0.900, HD: 21.1 mm, HD_95: 9.5 mm; both sides, see black lines in Fig. 4). By using the Mann-Whitney test (with $p < 0.05$), Institute 10 and 7 right, Institute 4 and 9 left are different from Institute 3 for two out of three metrics. Considering that Institute 4 has better performances than Institute 3, it is possible to consider it still compatible. Our model tends to predict CTVs larger than the radiation oncologists' ones, mainly in the cranio caudal side. For what regards the lateral side, it depends on the center behaviour: Institute 4 (right side), 5 and 9 have predictions typically larger than clinical CTV, Institute 2 and 3 are smaller, while the other centers are statistically in line with clinical volumes. Institutes 1, 10 and 7 are in line with clinical volumes. Comparing volume differences to the one of Institute 3, all institutes but 2 and 10 have smaller differences (Mann-Whitney $p < 0.05$).

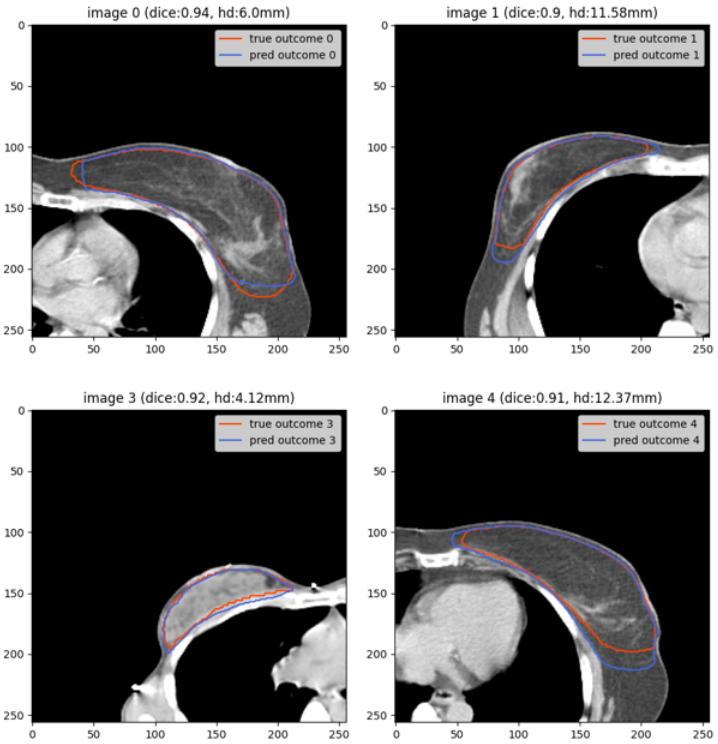


Figure 3: Four randomly chosen patients CT right and left examples taken at the mean cranio-caudal slices from Institute 3. The clinical CTV delineation (red) is superimposed with the observed predicted (blue) segmentation ROI. On top of the images, the DSC and HD_95 value for that specific patient is shown.

4 Discussion

The issue of clinical segmentation done manually with a extensive effort by clinicians is still open, particularly for what regards the CTV segmentation. Thanks to the availability of DL techniques this task started to be feasible automatically. The large dataset both collected in a single institute and multicentric data allowed us to train a reliable 3D neural network and validate it both internally and externally. Our model's performances are in line with the results obtained by previous papers [1, 13]. However, our dataset is larger than the data used by [13] and a 3D model was considered differently from [1]. Moreover, for the first time it is presented a validation done on multiple institutes showing the model consistency. All institutes, indeed, are in line with the training institute, with the exception of Institute 9 that seems to have a different approach in contouring (smaller clinical CTV, [12]). From this, contours of all centers could be included in a next benchmark model, with the exception of Institute 9. Important to note that, for all centers, outliers found with a DSCs below 0.7 have silicone implants, with the exception of only one patient from Institute 1 with a very small CTV (DSC = 0.61).

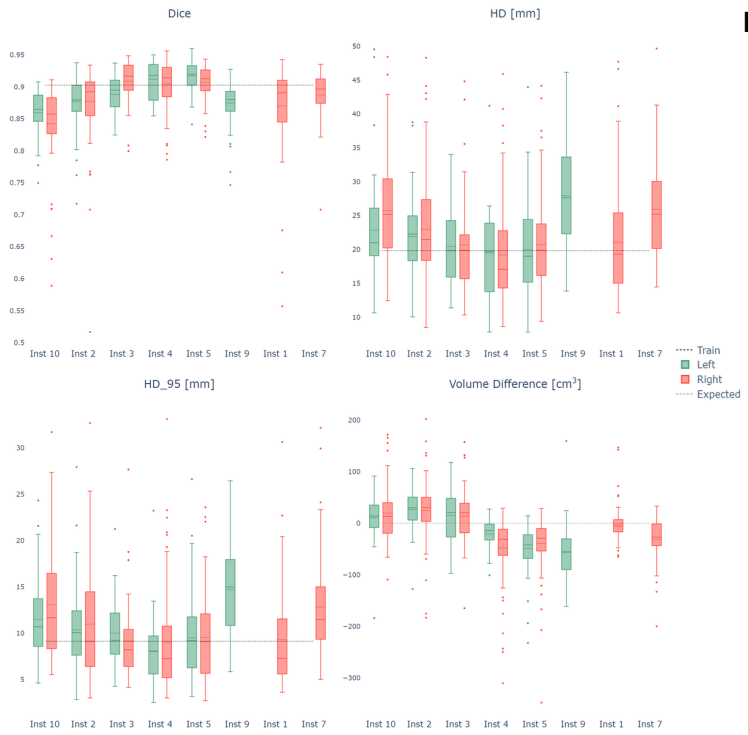


Figure 4: DSC, HD, HD_95 and VD (divided by the side of the CTV) trend across different institutes, being Institute 3 the training one. VD negative means larger CTV prediction. Box plots are shown with median (solid line) and mean values (dashed lines). Outliers are shown with dots. Black line shows the best metric values obtained for validation and gray line the median expected VD (zero).

5 Conclusion

We have constructed a 3D Unet for CTV segmentation trained on 611 patients of a single institute. The training performance was accessed through model validation metrics, obtaining DSC: 0.901 and HD_95: 9.6 mm. Moreover, the model was internally and externally validated. The internal validation confirms the results obtained in the training. The metrics used for comparison showed an overall agreement and model transferability for all centers, with the exception of Institute 9. Institutes 10, 7, 1 showed a larger variability in some metrics, though globally comparable with the internal validation set. Our model tends to predict contours larger than the radiation oncologists' ones in the cranio caudal side, while for the lateral side, it depends on the center: Institute 9 is the most discrepant one, with clinical contours definitely smaller than the other centers. A quality score assigned by radiation onclogists could also assess the impact of the predicted CTVs on the clinical workflow, giving an indication of how much editing is required in order to make the contours clinically acceptable. Considering that the presence of few patients with silicone implants generate most of the outliers present in our results, an increase the model performance is expected with their exclusion.

References

[1] K. Men, T. Zhang, X. Chen, et al. "Fully automatic and robust segmentation of the clinical target volume for radiotherapy of breast cancer using big data and deep learning". *Physica Medica* 50 (June 2018), pp. 13–19. DOI: 10.1016/j.ejmp.2018.05.006.

[2] A. R. Eldesoky, E. S. Yates, T. B. Nyeng, et al. "Internal and external validation of an ESTRO delineation guideline – dependent automated segmentation tool for loco-regional radiation therapy of early breast cancer". *Radiotherapy and Oncology* 121.3 (Dec. 2016), pp. 424–430. DOI: 10.1016/j.radonc.2016.09.005.

[3] M. S. Choi, B. S. Choi, S. Y. Chung, et al. "Clinical evaluation of atlas- and deep learning-based automatic segmentation of multiple organs and clinical target volumes for breast cancer". *Radiotherapy and Oncology* 153 (Dec. 2020), pp. 139–145. DOI: 10.1016/j.radonc.2020.09.045.

[4] S. Y. Chung, J. S. Chang, M. S. Choi, et al. "Clinical feasibility of deep learning-based auto-segmentation of target volumes and organs-at-risk in breast cancer patients after breast-conserving surgery". *Radiation Oncology* 16.1 (Feb. 2021). DOI: 10.1186/s13014-021-01771-z.

[5] Z. Liu, F. Liu, W. Chen, et al. "Automatic Segmentation of Clinical Target Volume and Organs-at-Risk for Breast Conservative Radiotherapy Using a Convolutional Neural Network". *Cancer Management and Research* Volume 13 (Nov. 2021), pp. 8209–8217. DOI: 10.2147/cmar.s330249.

[6] S. S. Almberg, C. Lervåg, J. Frengen, et al. "Training, validation, and clinical implementation of a deep-learning segmentation model for radiotherapy of loco-regional breast cancer". *Radiotherapy and Oncology* 173 (Aug. 2022), pp. 62–68. DOI: 10.1016/j.radonc.2022.05.018.

[7] L. R. Bell, J. A. Dowling, E. M. Pogson, et al. "Atlas-based segmentation technique incorporating inter-observer delineation uncertainty for whole breast". *Journal of Physics: Conference Series* 777 (Jan. 2017), p. 012002. DOI: 10.1088/1742-6596/777/1/012002.

[8] G. Wang, W. Li, M. Aertsen, et al. "Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks". *Neurocomputing* 338 (Apr. 2019), pp. 34–45. DOI: 10.1016/j.neucom.2019.01.103.

[9] P. M. Poortmans, S. Takanen, G. N. Marta, et al. "Winter is over: The use of Artificial Intelligence to individualise radiation therapy for breast cancer". *The Breast* 49 (Feb. 2020), pp. 194–200. DOI: 10.1016/j.breast.2019.11.011.

[10] Y. Zhong, Y. Guo, Y. Fang, et al. "Geometric and dosimetric evaluation of deep learning based auto-segmentation for clinical target volume on breast cancer". *Journal of Applied Clinical Medical Physics* 24.7 (Mar. 2023). DOI: 10.1002/acm2.13951.

[11] T. Akiba, S. Sano, T. Yanase, et al. "Optuna: A Next-generation Hyperparameter Optimization Framework". *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2019. DOI: 10.1145/3292500.3330701.

[12] A. Tudda, R. Castriconi, G. Benecchi, et al. "Knowledge-based multi-institution plan prediction of whole breast irradiation with tangential fields". *Radiotherapy and Oncology* 175 (Oct. 2022), pp. 10–16. DOI: 10.1016/j.radonc.2022.07.012.

[13] H. Baroudi, C. I. Huy Minh Nguyen, S. Maroongroge, et al. "Automated contouring and statistical process control for plan quality in a breast clinical trial". *Physics and Imaging in Radiation Oncology* 28 (Oct. 2023), p. 100486. DOI: 10.1016/j.phro.2023.100486.

Leveraging Network Uncertainty to Identify Regions in Rectal Cancer CTV Auto-Segmentations Likely Requiring Manual Edits

Federica Carmen Maruccio[1], Rita Simões[1], Charlotte Brouwer[2], Jan-Jakob Sonke[1], Tomas Janssen[1]

[1]Department of Radiation Oncology, The Netherlands Cancer Institute, Amsterdam, the Netherlands
[2]University of Groningen, University Medical Center Groningen, Department of Radiation Oncology, Groningen, The Netherlands

Abstract

While Deep Learning (DL) auto-segmentation has the potential to improve segmentation efficiency, manual verification and adjustment of the predictions are still required. To ensure an efficient segmentation workflow, auto-segmentation should be supplemented with appropriate quality assurance. In this work, we consider DL auto-segmentation of the mesorectum clinical target volume (CTV). To identify regions of the prediction likely requiring correction, uncertainty maps based on Monte Carlo dropout were generated and compared to segmentation performance. Interpreting network uncertainty is often complicated due to various sources of uncertainty interacting non-trivially both globally and on a local level. Therefore, we extended existing approaches in two aspects. First, network predictions were compared with both independent manual segmentations and manual corrections of the prediction. We hypothesized that the corrections only edit clinically relevant errors, reducing uncertainty due to less inter-observer variation. Second, we compared network uncertainty with network performance both globally and locally. On a global level, we investigated the correlation between mean network uncertainty and global network performance metrics. To study the correlation between two local metrics, we compared the width of the local uncertainty envelope with the length of the local editing vector. Using the uncertainty envelope width as a binary classifier, we achieved a median AUC of 0.8 in identifying regions that were manually corrected with edits of 2 mm or larger. For small edits, the median AUC to classify inconsistencies compared to the independent segmentations was significantly lower at 0.68. Assuming only truly relevant errors were manually adjusted, this demonstrates the capability of MC dropout uncertainty maps to capture relevant regions requiring corrections.

1 Introduction

Online Adaptive Radiation Therapy (OART) adjusts the dose distribution of each fraction to the anatomy of the day to account for anatomical changes throughout the treatment. This workflow potentially enables the use of smaller treatment margins and lower doses for healthy tissue. However, OART can lead to very long treatment times per fraction. The primary reason for this lies in the segmentation process, which needs to be performed for each treatment fraction and demands the presence of a clinician during treatment. As segmentation represents the primary bottleneck in the widespread clinical introduction of OART, automation of the workflow becomes crucial [1].

Deep learning (DL) for automatic segmentation represents a promising technique to improve both the consistency and speed of segmentation. However, these segmentation models are not perfect yet, therefore the need for manual adjustments reduces the time saved in clinical practice.

For an efficient segmentation workflow under time pressure, auto-segmentation should be supplemented with appropriate quality assurance. Uncertainty estimates can help to quantify the level of confidence in the DL model's prediction. These estimates can potentially guide the correction process of the DL-based contours by identifying the areas that need adjustments. Nevertheless, the interpretation of the predictive uncertainty is complex since it incorporates various sources of uncertainty interacting non-trivially, such as model-specific uncertainty and data-inherent uncertainty [2]. Moreover, probabilistic models suffer from miscalibration which can limit their use [3].

Typically, clinically used ground truth segmentations are employed for DL predictions and hence for uncertainty maps. Inter-observer variation in segmentation, however, is known to be present and therefore not all mismatches between a prediction and a clinical ground truth will be relevant. Moreover, the actual key question is whether the uncertainty maps can highlight the regions of an auto-segmentation that would be manually adjusted. Since the adjustments are biased by the prediction, we hypothesize that fewer sources of uncertainty are present in the edits and only truly relevant errors are addressed in the corrections.

To validate our hypothesis, we analyzed the correlation between uncertainty estimates and manual corrections performed on automatically predicted contours, as well as inconsistencies with manual contours performed independently of the DL model in clinical practice. We performed our analysis both at global and local levels. To the best of our knowledge, this study represents the first attempt to perform such analysis at a local level using corrected contours, aiming to progress towards the integration of uncertainty estimates into the OART workflow.

2 Materials and Methods

2.1 Data and Model

We included 16 patients treated in our institute on the Unity MR-Linac [Elekta AB, Stockholm, Sweden] for rectal cancer. The patients were treated with 5 x 5 Gy with an adapt-to-shape procedure. The dataset includes in total 79 MRI scans with a voxel size of 1.2 x 0.57 x 0.57 mm^3. *Clinical* contours of the mesorectum CTV were delineated online by a clinician before treatment, without access to the DL network.

We employed a nnU-Net framework [4], previously trained in our institute to segment the mesorectum CTV on MRI scans of a separate training set [5]. The model works on a 2D level and achieves a median Dice score of 0.89. *Predicted* contours were generated using this model, and a clinician manually *corrected* the predictions. The

predictions were found to be clinically acceptable only after correction.

2.2 Uncertainty Maps

The Monte Carlo (MC) dropout technique [6] was applied to generate uncertainty maps at inference, using a dropout rate of 0.2. At test-time, $T = 25$ random samples of the softmax output of the network were generated. The probability p_n for each voxel n to belong to the CTV was then computed as the average of the T samples using Equation 1.

$$p_n = \frac{1}{T}\sum_{t=1}^{T} p_{n,t} \qquad 1$$

Probability maps were converted into uncertainty maps by using Equation 2,

$$H_n = -[p_n\, log(p_n) + (1-p_n)\, log(1-p_n)]\frac{1}{log\, 2} \qquad 2$$

where the normalized entropy $H_n \in [0,1]$ is employed as a measure of uncertainty.

2.3 Global Analysis

We first investigated the global relationship between the uncertainty estimates and the segmentation performance of the network. The mean uncertainty ($\bar{H}$) was computed to define the global level of uncertainty of each CTV contour as in Equation 3:

$$\bar{H} = \frac{1}{N}\sum_{n=1}^{N} H_n \qquad 3$$

where N represents the number of voxels in the volume of interest.

To evaluate the segmentation performance, the *predicted* contours were compared with the *clinical* contours and the manual *corrections* using the following global metrics: Dice similarity coefficient (DSC), 95th percentile Hausdorff distance (95th HD), Surface Dice similarity score (SDSC) at 3 mm and added path length (APL).

The correlation between H_n and each performance metric was examined for both the *clinical* and *corrected* contours and quantified with the Pearson correlation coefficient *R*.

2.4 Local Analysis

A local interpretation of the correlation between uncertainty and contours was based on the hypothesis that the uncertain area is larger where the errors, and consequently edits, are bigger. The local analysis was performed per slice (2D) since the model used was 2D and clinicians tend to adjust contours per slice.

A threshold T was set to extract an uncertainty envelope from the uncertainty map. Specifically, pixels with an uncertainty value above the specified threshold were included in the envelope creating a mask:

$$\begin{cases} if\ H_n \geq T \rightarrow 1 \\ if\ H_n < T \rightarrow 0 \end{cases}$$

Subsequently, the width of the uncertainty envelope was calculated using the Bidirectional Local Distance (BLD) at every point of the *predicted* contour [7].

While the envelope width was used as a measure of local uncertainty, for the delineations we chose the length of the editing vector. When comparing the *predicted* contours with the *clinical* contours, this length correlates to the local inconsistency between the two contours. Conversely, when comparing *predictions* with *corrections*, the length can be interpreted as the local edit. The BLD was again employed to compute the length of the editing vector in both cases.

At this stage, for each predicted point of the CTV, three values were available: the width of the envelope, the length of the editing vector for the *clinical* contour, and the length of the editing vector for the *corrected* contour. A threshold, denoted as λ, was applied to the editing vector's length to convert the length values into binary form. In other words, choosing λ = 2 mm implies that we study only all errors larger than 2 mm. Subsequently, each envelope width value served as a binary classifier for the length.

The performance of the classifier was evaluated per slice using the Area Under the ROC Curve (AUC). Then, the AUC values of the slices were averaged per scan. The analysis was repeated for different combinations of T and λ, considering both *clinical* and *corrected* contours. Finally, we studied the calibration of the model by selecting the scan-optimal T, i.e. the uncertainty threshold that optimizes the AUC for a specific scan. This is crucial as distinct thresholds may be optimal for different scans [3].

3 Results

Figure 1 shows the outcomes of the global analysis. Each subplot displays the relationship between $\bar{H}$ and one of the four selected performance metrics, with *corrections* represented by blue dots and *clinical* contours represented by orange dots. Major differences in terms of segmentation performance emerge between *corrections* and *clinical* contours. Notably, the APL metric shows the highest correlation (R_{Corr}=0.76, R_{Clin}=0.81), while the DSC exhibits the smallest Pearson coefficient (R_{Corr}=-0.32, R_{Clin}=-0.37). However, no substantial differences in correlation between *corrections* and *clinical* contours were identified.

Figure 2 illustrates the AUC-based performance of the uncertainty estimates employed as a binary classifier for segmentation edits, utilizing three different thresholds (2 mm, 6 mm, and 12 mm). Each subplot, accordingly, shows the performance of the classifier for edits or inconsistencies larger than the specified threshold. Each boxplot illustrates the distribution of the AUC values across the 79 scans for a specific combination of T and λ. The blue boxplots pertain to edits based on the *corrected* contours, while the orange boxplots relate to errors associated with the *clinical* contour. No editing thresholds λ smaller than 2 mm and higher than

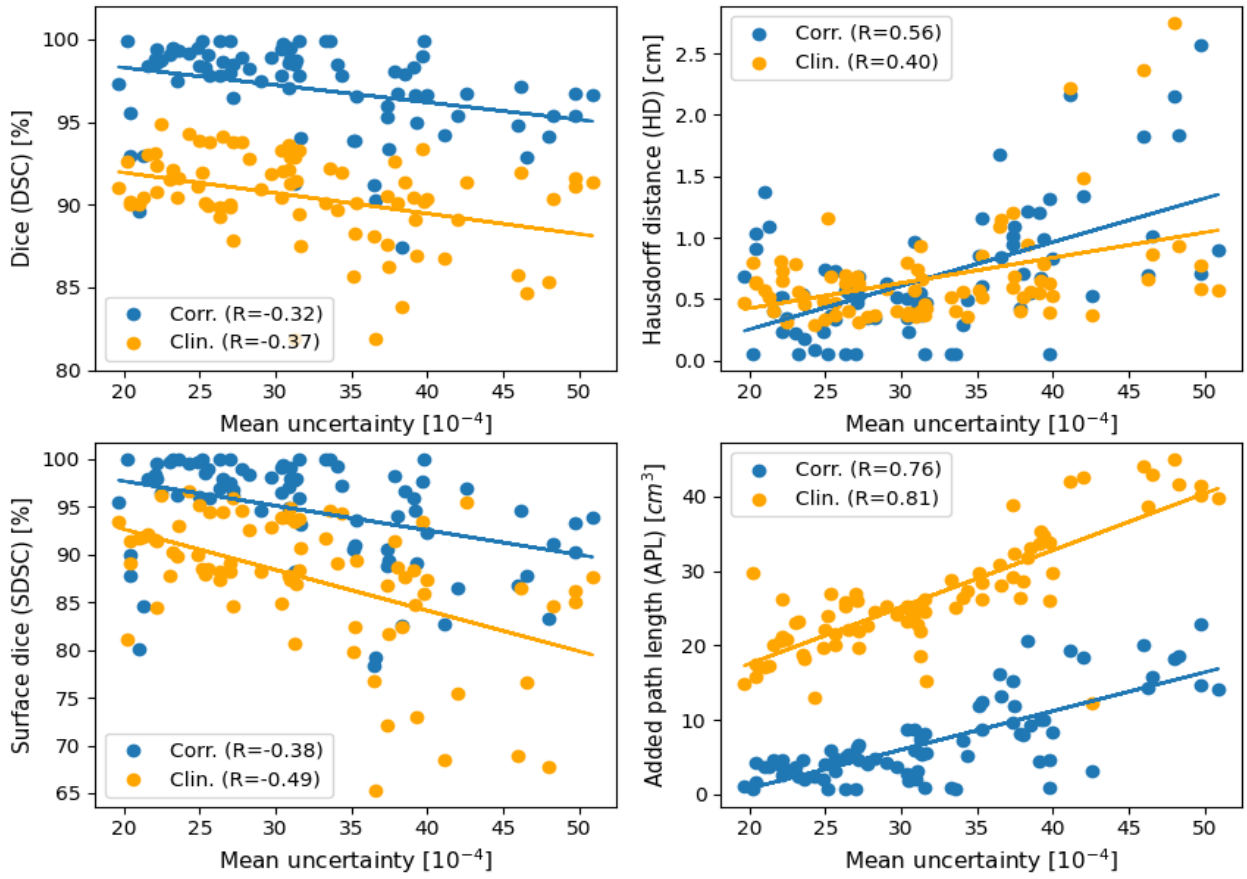


Figure 1 Global correlation between mean uncertainty $\overline{H}$ and network performance metrics.

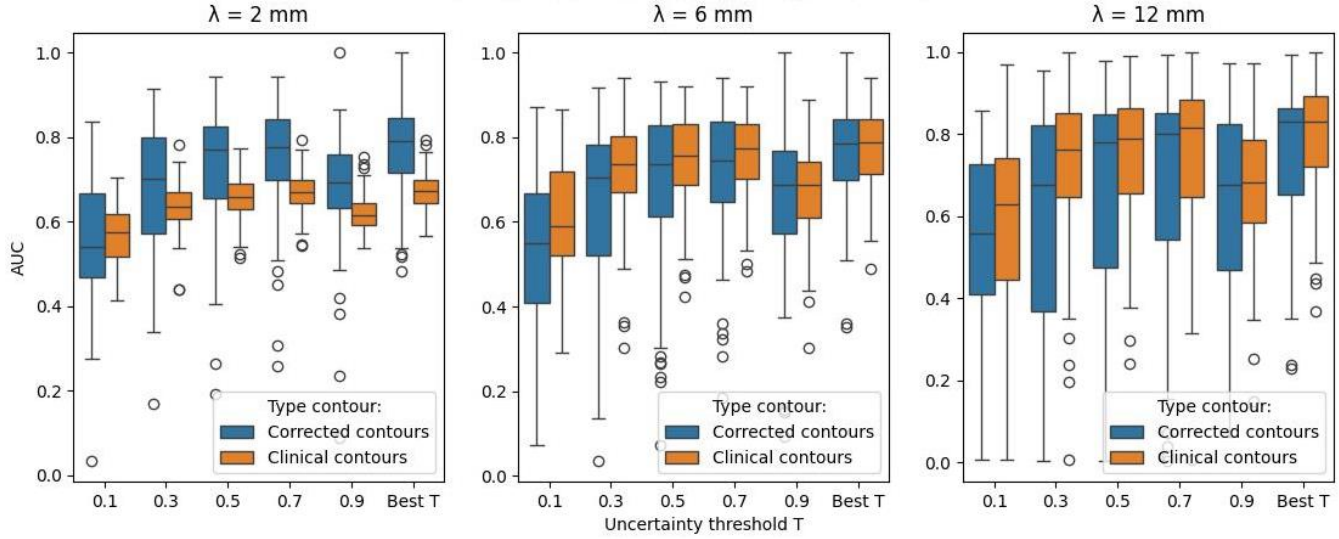


Figure 2 AUC-based performance of the local uncertainty metric used as a binary classifier for edits larger than λ.

12 mm were used due to a decrease in the number of scans with edits that were binary classifiable for values of λ outside this range.

What stands out in Figure 2 is the influence of the uncertainty threshold *T* on the performance. The AUC exhibits an increase until *T*=0.7, followed by a subsequent decrease. This trend remains consistent while varying λ and holds for both *corrected* and *clinical* contours. For minor edits (λ=2mm), the classifier performed better on the *corrected* contours compared to the *clinical* contours, whereas results are comparable at medium and large λ values. Additionally, the variability of the performance increases when increasing the editing threshold λ. The final boxplots in each subplot, labelled '*Best T*,' present the performance of the classifier when utilizing patient-specific T. Selecting the optimal T for each patient enables to obtain a median AUC value above 0.8 for *corrected* contours and around 0.68 for *clinical* contours at λ=2mm.

4 Discussion

This study aimed to study the differences between *clinical* and *corrected* contours in the context of uncertainty application. Conventional analyses typically only investigate how well uncertainty estimates correlate with

segmentation quality on a global level [8]. Our approach also involved local investigation. Specifically, we examined the correlation between network uncertainty and the amount of editing on the *predicted* contours in order to build a binary classifier for edits.

The results of the global analysis show no significant differences in the correlation with uncertainty between *corrections* and *clinical* contours. The geometric similarity improves when comparing *predictions* to *corrections*, as opposed to *clinical* contours. This suggests that *corrections* are biased towards the *prediction*, which may be due to reduced inter-observer variability in the *corrections*. Possibly, only the truly relevant edits are addressed in the *corrections*, while less critical errors may be ignored. Furthermore, there is a clear correlation increase as the examination moves from volumetric measures, such as the volumetric Dice score, to more local metrics, like the APL. This emphasizes the potential of a local analysis of the correlation.

When studying the local association, the predictive value of local uncertainty is stronger in the *corrected* contours compared to the *clinical* ones for small edits (λ=2mm). This demonstrates that uncertainty can capture the relevant edits that are prevalent in the *corrections* but may be intertwined with less significant errors in the *clinical* contours. Conversely, for higher editing thresholds, the classifier performance is comparable for both types of contours. Larger errors are more likely to be relevant and thus considered in both *corrections* and *clinical* delineations.

Additionally, employing extreme uncertainty thresholds, such as T=0.1 and T=0.9, leads to poorer performance of the binary classifier. Intermediate values seem to be more favourable for this analysis. Performance is further improved by utilizing patient-specific thresholds ('*Best T*'), involving the selection of the threshold that maximizes the correlation between editing vectors and uncertainty envelopes for each patient. However, it is noteworthy that the distribution of AUC using patient-specific T is very similar to the one at T=0.7, indicating that the model is properly calibrated [3].

Our work exhibits some limitations. First, the dataset used included a limited number of scans. Moreover, the *clinical* contours were made under time pressure and may be biased by different non-DL models. Additionally, the computation of both local metrics, namely envelope width and edit length, relied on BLD which may yield suboptimal results for intricate contour shapes. Finally, the use of a 2D network posed challenges in extending the analysis to the sagittal plane since the model works on an axial level.

5 Conclusion

In this study, we explored the correlation between uncertainty estimates and both manual *corrections* and inconsistencies with the *clinical* contours. Our novel methodology offers a local analysis, resulting in an uncertainty-based binary classifier that predicts which region requires editing. We obtained a median AUC of around 0.8 to identify manual edits of 2 mm or larger. The findings highlight that for small edits, the uncertainty maps did correlate with manual *corrections* but only weakly to inconsistencies with the *clinical* contours. Assuming only truly relevant mistakes were manually adjusted, this demonstrates the capability of the uncertainty maps to capture these errors.

Despite the limitations, the study contributes to taking OART a step further by leveraging the uncertainty estimates to help clinicians to more efficiently evaluate and correct DL auto-segmentation. However, future work is still needed to assess the relevance and predictability of edits, further enhancing the potential to reduce the physician's workload and save time in OART through uncertainty-guided segmentation corrections.

Acknowledgements

We thank Anja Betgen, Lisa Wiersema, Corrie Marijnen, and Femke Peters for contouring the mesorectum CTVs, and Nicole Ferreira Silvério for providing the DL model.

Conflicts of Interest

The Netherlands Cancer Institute receives institutional funding and research support by Elekta AB (Stockholm).

References

[1] L. G. W. Kerkmeijer, V. Valentini, C. D. Fuller, and B. J. Slotman, "Editorial: Online Adaptive MR-Guided Radiotherapy," *Frontiers in Oncology*, vol. 11. Frontiers Media S.A., Aug. 30, 2021. doi: 10.3389/fonc.2021.748685.

[2] W. van Rooij, W. F. Verbakel, B. J. Slotman, and M. Dahele, "Using Spatial Probability Maps to Highlight Potential Inaccuracies in Deep Learning-Based Contours: Facilitating Online Adaptive Radiation Therapy," *Adv Radiat Oncol*, vol. 6, no. 2, Mar. 2021, doi: 10.1016/j.adro.2021.100658.

[3] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On Calibration of Modern Neural Networks," Jun. 2017, [Online]. Available: http://arxiv.org/abs/1706.04599

[4] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation," *Nat Methods*, vol. 18, no. 2, pp. 203–211, Feb. 2021, doi: 10.1038/s41592-020-01008-z.

[5] N. Ferreira Silverio *et al.*, "MO-0798 Evaluation of target autocontouring for MRI-guided online adaptive treatment of rectal cancer," 2023.

[6] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning," Jun. 2015, [Online]. Available: http://arxiv.org/abs/1506.02142

[7] H. S. Kim, S. B. Park, S. S. Lo, J. I. Monroe, and J. W. Sohn, "Bidirectional local distance measure for comparing segmentations," *Med Phys*, vol. 39, no. 11, pp. 6779–6790, 2012, doi: 10.1118/1.4754802.

[8] K. Hoebel *et al.*, "An exploration of uncertainty information for segmentation quality assessment," SPIE-Intl Soc Optical Eng, Mar. 2020, p. 55. doi: 10.1117/12.2548722.

Virtual patient-specific quality assurance: a prospective two-year analysis

Alon Witztum[1], Phillip Wall[2], Emily Hirata[1] and Gilmer Valdes[1]

[1]Department of Radiation Oncology, University of California – San Francisco, California, USA
[2]Department of Radiation Oncology, Washington University in St Louis, Missouri, USA

Abstract In modern radiation oncology clinics, patient-specific quality assurance (QA) has been shown to have questionable utility in relation to the disproportionate amount of resources it demands. Predictive machine learning models can be designed to indicate the likelihood a treatment plan will fail QA based on its complexity, which can maximize efficiency by focusing resources less on "simple" plans and more on "complex" plans. Our implementation of this method – Virtual QA (VQA) – has been integrated into a clinical workflow to run prospectively alongside regular QA operations for over two years, storing a predicted QA result alongside the measured result for every treatment plan. Results from this study show VQA can reduce the number of QA measurements by approximately two-thirds, with a specificity of 100% and sensitivity of 67% in identifying plans with an absolute QA result greater than or equal to 4%. The model maintained a mean absolute error of 0.97% and a maximum absolute error of 2.99%. The area under the ROC curve for VQA to predict QA results that fail at a 3% threshold is 0.77, and at a 4% threshold it is 0.94.

1 Introduction

Measurement-based patient-specific quality assurance (QA), despite being both labor and resource intensive, has long been the most popular method to QA patient plans, posing challenges in a clinical setting where timeliness and efficiency are crucial. As far back as 2012, patient-specific quality assurance was identified as being the least effective at identifying treatment plan issues [1]. Since the seminal paper described the use of machine learning (ML) to predict patient-specific QA results via plan complexity in 2016 [2], several academic institutions have reinforced the viability and accuracy of such ML models [3].

In 2021 our group developed a machine learning based clinical decision support system which was deployed as a web-platform, Virtual QA (VQA), at the University of California, San Francisco (UCSF). VQA provides clinicians with a predicted QA result without the need to deliver the treatment plan. VQA was deployed to demonstrate the first long-term prospective validation of a patient-specific QA prediction model. With the initial publication describing the platform and initial clinical implementation, we also reported initial results from the first three-month period [4]. We now present the results of our long-term study, with a dataset covering 26 months of prospective validation. This work aims to provide a comprehensive analysis and evaluation of the long-term effectiveness and reliability of the VQA model in a clinical setting.

2 Materials and Methods

Patient specific QA data was extracted for 3798 unique VMAT treatment plans delivered between February 1st 2021 and April 10th 2023. The data includes treatment plans delivered on one Truebeam STx and three matched Truebeams (Varian, a Siemens Healthineers Company, Palo Alto, CA).

Patient-specific QA at our institution pairs a 3D independent dose recalculation performed by Mobius3D (Varian, a Siemens Healthineers Company, Palo Alto, CA) with a single-point ion chamber measurement within the Mobius solid water verification phantom (Figure 1). This phantom has seven inserts for ion chamber placement (positions A-G). Treatment plans were all created in the RayStation (Raysearch Laboratories, Sweden) treatment planning system (TPS). A Python script was developed in the TPS to automatically create QA plans on the Mobius phantom and send the dose at each ion chamber insert position to Research electronic data capture (REDCap), an online HIPAA-compliant research database.

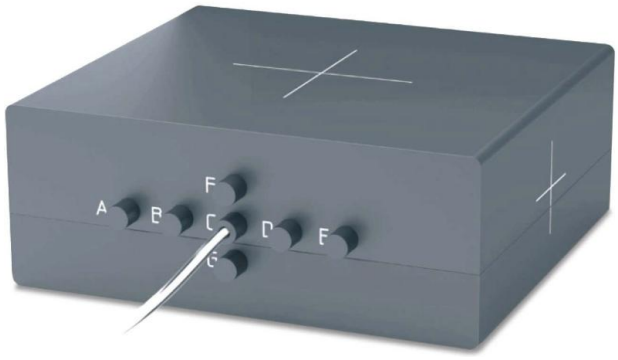


Figure 1: Mobius solid water phantom with seven ion chamber inserts labelled A-G. Reproduced from www.varian.com.

Following the recommendations of AAPM Task Group 218 [5], the QA team are instructed to use one of the seven ion chamber positions that meets the following high-dose-low-gradient conditions: 1) dose greater than 80% of the prescription dose, 2) standard deviation of less than 5% across the chamber volume. If there is no such point, the isocenter is shifted in the QA plan so that the high-dose-low-gradient region can be measured in one of the seven ion chamber positions. The QA team record the ion chamber dose in REDCap, which automatically calculates the percent difference between the TPS dose and the measured ion chamber dose (QA result). The QA team also record the percent difference predicted by the deployed VQA system (VQA prediction).

VQA utilizes an XGBoost machine learning (ML) model which was trained on 579 clinical treatment plans and tuned to predict QA results of VMAT plans. The ML model incorporates 30 classes of complexity features (details of which can be found in established literature [2, 4]) as input for the model. Model tuning was performed using a grid search over learning rate and tree depth hyperparameters and evaluated with 10-fold cross-validation. Further details on the platform are available [4].

For the 3798 VMAT plans extracted, the percent difference between the TPS dose and the measured QA dose (QA result), this percent difference predicted by VQA (VQA prediction), and the difference between these two results (VQA error) were extracted. Initial data cleaning involved remeasurement of the poorest predicted (absolute VQA error > 3%) treatment plans in order to eliminate setup errors and poor ion chamber position choice as the cause of the disagreement. The mean absolute error (MAE), maximum absolute error, and ROC curves of the VQA model are reported.

3 Results

In the initial three-month prospective validation, the machine learning model had a mean absolute error of 1.08% ± 0.77% and maximum absolute error of 2.98% [4]. Of the 3798 unique treatment plans, 87 plans (2.3%) initially fell outside of the expected 3% maximum absolute error and were remeasured. After remeasurement, all 87 plans improved, with the mean absolute error of all plans remaining around 1% (1.02% prior to remeasurement, 0.97% after remeasurement), but the maximum absolute error falling from 8.65% to 2.99%. The distribution of VQA errors is shown in Figure 2 with an average and standard deviation of 0.02% ± 1.20%.

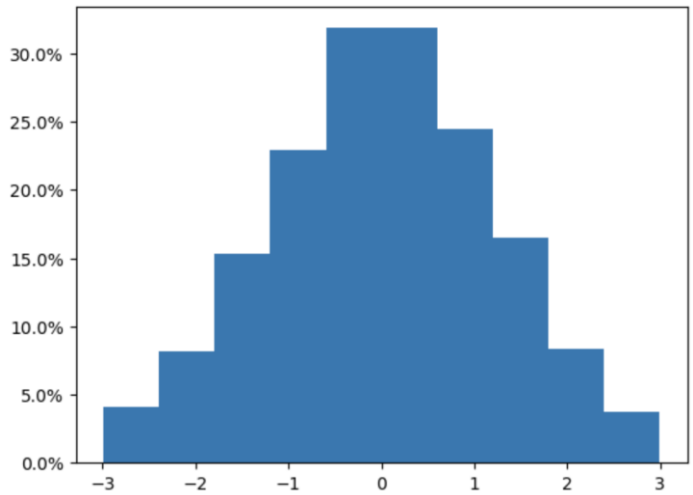


Figure 2: Distribution of VQA Error, the difference between the VQA prediction and the QA result.

The distribution of the QA results is shown in Figure 3 with an average and standard deviation of 0.62% ± 1.26%. There were 11 QA results with an absolute percent dose difference greater than or equal to 4% and 108 with an absolute percent dose difference greater than or equal to 3%.

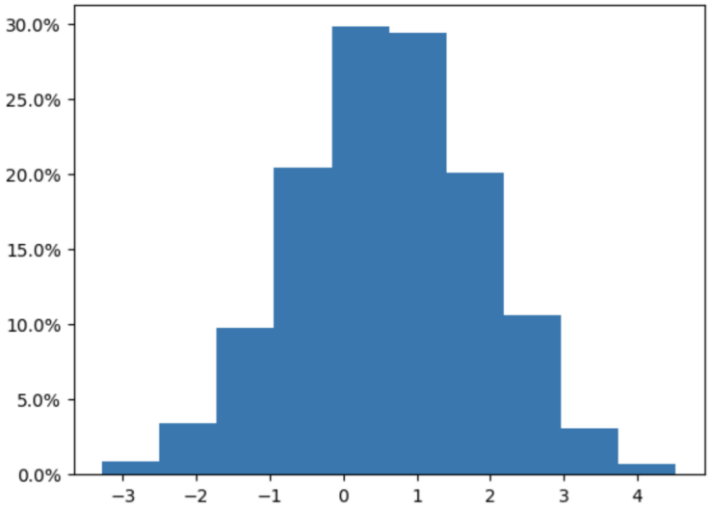


Figure 3: Distribution of QA measurement results

Considering QA results of above 3% or 4% absolute percent difference as 'failing', the VQA prediction model has an AUC of 0.77 or 0.94 as seen in Figure 4.

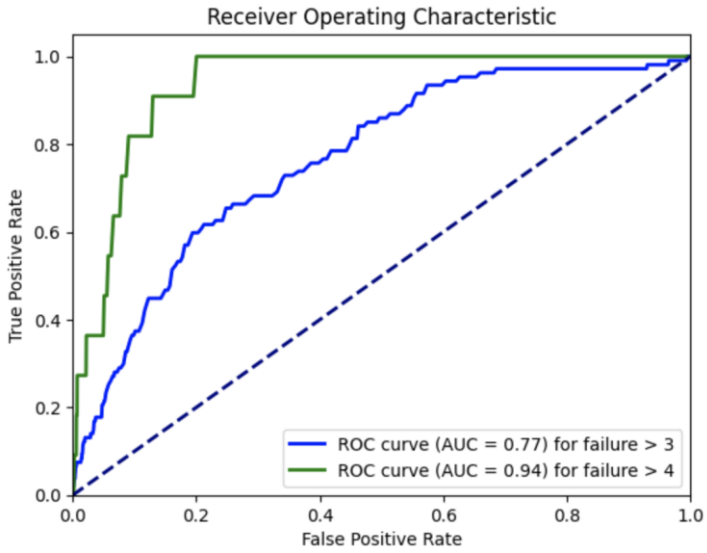


Figure 4: ROC curves and AUC values for VQA model using 3% and 4% as clinical QA failure thresholds.

For this prospective study, a threshold of 1% on the VQA prediction was chosen, meaning that if the absolute percent difference predicted by VQA was less than or equal to 1%, a QA measurement would not be required. Applying this retrospectively to the 3798 plans, 2529 (66.6%) would not need to be measured. Of those 2529, 33 had a QA result greater than or equal to 3%. Of the 1269 (33.4%) plans that still required measurement, 75 had a QA result greater than or equal to 3%, of which 11 were greater than or equal to 4%. This data is summarized in Table 1 below.

	Pass QA	Fail QA		Pass QA	Fail QA
Pass VQA	2496 (66%)	33 (0.01%)		2529 (67%)	0 (0%)
Fail VQA	1194 (31%)	75 (0.02%)		1258 (33%)	11 (<0.01%)

Table 1: Truth tables showing number of measurements needed and saved by using a 1% threshold on the VQA model and a 3% threshold on the QA result (left) and a 4% threshold on the QA result (right).

4 Discussion

Our continued brute-force approach to patient-specific QA, requiring the same time-consuming approach for every single plan remains an ineffective use of physicist time [1]. Fewer than 0.03% of plans (108) 'fail' QA with a 3% threshold, and even fewer (11, <0.01%) 'fail' with a 4% threshold. The VQA machine learning model can predict QA results, and therefore identify treatment plans that are more likely to fail QA. By applying a simple threshold of 1% below which a QA measurement is no longer required, we can reduce our QA load by approximately two thirds. By building an ML model that predicts a continuous numeric QA result, rather than a pass/fail classification model, we also allow for a customizable threshold that can be adjusted based on risk tolerance and periodic data analysis.

An unfortunate limitation of QA prediction models is the lack of gold standard data due to the uncertainties introduced through set-up as well as poor measurement point selection. This is demonstrated by the fact that in all 87 remeasured plans, the VQA Error was reduced. An additional limitation is the small number of treatment plans that actually 'fail' with a 3% or 4% threshold, giving machine learning models few data points in the region of the most interest. This limitation is another reason why prediction of a continuous numeric QA result is preferred over a classification system.

A barrier to removal of measurement-based patient-specific QA is that it is often used as an end-to-end test to test data transfer and data integrity. This is especially true of clinics that operate in a multi-vendor environment. As part of setting up a comprehensive QA program, these needs should also be addressed. An efficient patient-specific QA pipeline could feasibly involve three components: 1) independent dose recalculation, 2) VQA prediction model, 3) computerized data integrity checks. None of these three components require delivery of the treatment plan prior to treatment, making it also an attractive option for online adaptive radiotherapy.

5 Conclusion

A machine learning based clinical decision support system has been successfully deployed and integrated into the clinical workflow. While at this time, QA measurements are continuing on all treatment plans, VQA has been shown to maintain the accuracy quoted during initial testing and validation.

References

[1] Ford EC, Terezakis S, Souranis A, et al (2012) Quality Control Quantification (QCQ): A tool to measure the value of quality control checks in radiation oncology. International Journal of Radiation Oncology*Biology*Physics. doi: 10.1016/j.ijrobp.2012.04.036

[2] Valdes G, Scheuermann R, Hung CY, et al (2016) A mathematical framework for virtual IMRT QA using machine learning. Medical Physics 43:4323–4334. doi: 10.1118/1.4953835

[3] Chan MF, Witztum A, Valdes G (2020) Integration of AI and machine learning in Radiotherapy QA. Frontiers in Artificial Intelligence. doi: 10.3389/frai.2020.577620

[4] Wall PDH, Hirata E, Morin O, et al (2022) Prospective clinical validation of virtual patient-specific quality assurance of volumetric modulated arc therapy radiation therapy plans. International Journal of Radiation Oncology*Biology*Physics 113:1091–1102. doi: 10.1016/j.ijrobp.2022.04.040

[5] Miften M, Olch A, Mihailidis D, et al (2018) Tolerance limits and methodologies for IMRT measurement- based verification QA: recommendations of AAPM task group no. 218. Medical Physics. doi: 10.1002/mp.12810

Model Size Reduction in CNNs through Sparse Regularization: Balancing Efficiency and Diagnostic Accuracy in Medical Imaging.

Carlos Garcia-Fernandez[1], Wilmer Arbelo-Gonzalez[3], Jerome Friedman[4] and Gilmer Valdes[1,2]

[1]Department of Radiation Oncology, University of California, San Francisco, USA
[2]Department of Epidemiology & Biostatistics, University of California, San Francisco, USA
[3]Department of Image Analysis and Computational Biology, Altos Labs, Redwood City USA
[4]Department of Statistics, Stanford University, Stanford, USA

Abstract Convolutional neural networks (CNNs) are at the forefront in medical imaging diagnosis and prognostics. Typically, networks contain millions of parameters and are predominantly designed for the ImageNet database, which consists of millions of images across 1000 categories. In contrast, medical imaging tasks usually involve fewer categories and a smaller number of training images. In this study, we introduce an innovative method for automatic model size reduction in CNNs that preserves performance in computer vision classification tasks.

1 Introduction

The application of deep learning models in medicine has brought transformative changes to healthcare [1,2]. Larger models have demonstrated remarkable capabilities, including superior generalization and prediction performance [3,4], heightened ability to capture intricate patterns in complex data [5-8], and increased versatility across various applications [9]. These advancements have been instrumental in pushing the boundaries of machine learning—ranging from radiological images to genomics—identifying patterns and features that are often imperceptible to human experts [10]. However, this scaling up of model sizes introduces several challenges. One such challenge is that training models with numerous parameters on small datasets can result in overfitting, where the model performs well on the training data but fails to generalize effectively to new, unseen data [11]. This limitation is particularly significant in healthcare, where large datasets are often either not readily available or are costly to acquire [12,13]. Additionally, large models require substantial computational resources, which may be prohibitively expensive and have a significant environmental impact due to energy consumption [14,15]. Another important constraint is memory usage, which is determined by the number of mathematical operations the hardware can perform. With fewer parameters, the memory requirements are lessened, making it possible for instance to handle larger images resolutions (e.g. CT-scans) and to run models for training or inference on less powerful hardware. This is a crucial consideration, especially in environments where high computational resources are not readily available or are cost prohibitive. Additionally, the complexity of these models often makes them "black boxes," thereby raising concerns about interpretability and trust in clinical settings. These challenges amplify the need for more efficient, interpretable, and adaptable deep learning solutions tailored for medical applications.

The idea of reducing the size of deep learning models while maintaining similar performance is not new in the realm of machine learning and it has been extensively researched [15, 16]. Currently, there are three primary strategies for achieving less computationally intensive models. The first approach involves the manual design of adaptive architectures, such as MobileNet [17] and EfficientNet [18]. These CNN architectures were specifically engineered to achieve nearly equivalent performance using only a fraction of the computational resources required by their full-sized counterparts. However, the use of these models is restricted to some specific computer vision tasks and require a thorough testing and tweaking of their size configurations to find an optimal balance between performance and computational efficiency. The second approach, known as Magnitude Pruning, seeks to simplify network architectures by setting smaller parameters in absolute value to zero based on specific heuristics [15,16,19]. The biggest limitation of Magnitude pruning is the fact that the size of the parameter is not indicative of its impact into the loss function and as such results in suboptimal solutions. Moreover, it is also important to note that these methods do not automatically lead to faster inference times or more efficient memory usage. This is primarily because modern hardware is optimized for dense matrix operations, and the sparsity introduced by pruning tends to disrupt these operations, reducing their efficiency. In simpler terms, the hardware still computes the zeroed or pruned parameters, meaning the reduction in the number of parameters doesn't translate to proportional gains in computational speed or memory efficiency. While pruning may reduce storage requirements, the practical benefits in terms of computational efficiency, inference time, and memory usage are often limited. Additionally, some pruning techniques require complex mask matrices or indices to remember which connections have been pruned, adding an extra layer of computational overhead. Moreover, magnitude pruning techniques are prone to underperformance, which may require cessation of the pruning process, hindering researchers from reaching their original objectives. Ultimately, the ability of these ad hoc techniques to generalize across various tasks or datasets is not assured, making them an unreliable solution for model simplification and computational efficiency. Lastly, the third strategy is Second-order pruning techniques [20] which utilize information from the loss function but are computationally expensive due to the

need to calculate second-order derivatives, like the Hessian matrix, making them impractical for large models.

We propose a novel methodology to reduce CNN model size while maintaining near-equivalent performance in a computer vision classification task. Our core algorithm introduces sparse regularization into nonlinear functions, solving a long-standing constrained optimization problem without any prior knowledge of the regularization strength. Unlike existing regularization methods that require trial-and-error based cross-validation to find the regularization strength, we automatically identify a set of network architectures based on dataset-specific regularization constraints.

2 Materials and Methods

We solve the optimization problem:

$$\hat{\boldsymbol{\omega}} = \text{argmin}_{\boldsymbol{\omega}} \mathbb{E}_{\boldsymbol{x},y}(L[y, f(\boldsymbol{x};\boldsymbol{\omega})]) \quad \text{s.t. } P(\boldsymbol{\omega}) \leq t$$

where L is a loss function, f is a non-linear function (e.g. CNN), and P(w) is a general constraint function that monotonically increases with |w| so that:

$$\frac{\partial P(\boldsymbol{w})}{\partial |w|} \geq 0.$$

For instance, $P(\omega) = \Sigma|\omega|$ for Lasso (L1) and $P(\omega) = \Sigma\omega^2$ for Ridge regression (L2).

In our approach, we impose varying values to the constraint function $P(w)$ and solved through sparse regularization (L1). By solving the optimization problem for each specified value of regularization strength, we systematically generate a series of different model architectures. Each architecture is then analyzed to assess its performance and structural characteristics. During this analysis, we identify and remove filters within the architecture where all the weights have become zero. Effectively, these weights no longer contribute to the model, and their elimination helps to reduce the complexity of the network. After pruning these non-contributive elements, we redefine the architecture to include only the remaining, or 'surviving,' convolutional channels. This structured pruning and restructuring process is a critical step towards achieving a more efficient architecture that retains the necessary parameters for accurate predictions while discarding the redundancies that do not contribute to the model's performance. The result is a set of streamlined CNN architectures, each corresponding to a particular regularization strength, optimized for both performance and efficiency.

3 Results

Our initial experiment compared the performance of our algorithm with three magnitude-based pruning techniques using L1-regularization on the pre-trained ResNET-50 architecture within the MURA dataset. This dataset comprises 14,863 studies from 12,173 unique patients, totaling 40,561 multi-view radiographic images. The images encompass seven types of standard upper extremity radiographic studies: elbow, finger, forearm, hand, humerus, shoulder, and wrist. Each study was categorized into binary classifications, with '1' denoting a normal result (target variable) and '0' indicating an abnormal finding. The classification problem presented a class balance of 50/50 percent.

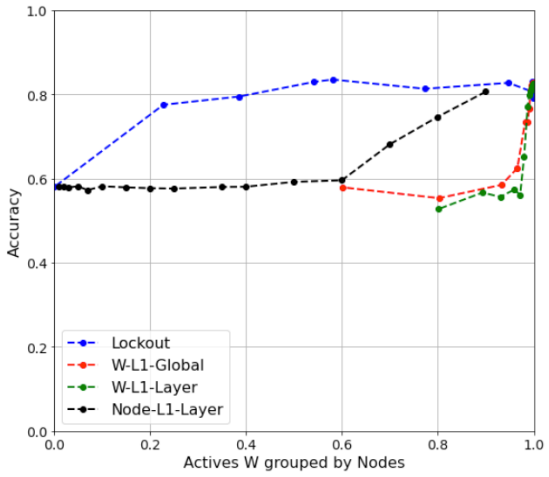


Figure 1: Obtained performance for different model compression rates when using Lockout algorithm (blue) vs magnitude-based pruning techniques (black, red and green curves).

The results, depicted in Figure 1, show our algorithm achieving state-of-the-art performance with a substantially reduced model size. Our approach significantly outperforms standard pruning techniques by reducing the number of active neurons required during the learning process while maintaining near-equivalent performance. This important reduction on the model size translates to fewer operations, lower inference time, and decreased memory consumption (Table 1).

Model	Parameters	Nodes	Inference (s)	Mult-adds (G)	Weights size (mb)	Total size (mb)
Full	25512130	26560	7.72	246	94	10583
Compressed	1627741	8774	4.75	71	6	7390

Table 1: Comparison of computational resources employed for the full ResNet50 architecture vs our compressed network.

An additional significant finding of our study is the clear interpretability gained from examining sparsity patterns across the convolutional network layers automatically induced by our algorithm. Illustrated in Figure 2 is the stratified architecture that emerges under our algorithm's optimization, charting a course from the original model size (depicted in blue) to a reduced model retaining only 20% of the nodes (visualized in black). The visualization reveals a

discernible trend: the encoder portion of the network largely retains its structure, indicating its critical role in performance, whereas the algorithm preferentially prunes channels closer to the output layer.

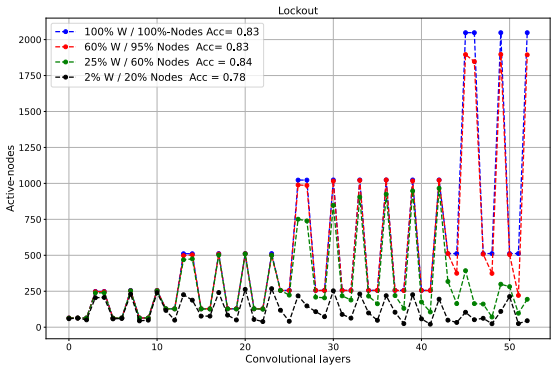


Figure 2: Number of CNN filters employed by convolutional layers for different model compression rates.

This result suggests that in a wide range of scenarios, modern computer vision architectures may be over-parameterized, particularly for tasks beyond the scope of large-scale datasets like ImageNet. Such extensive parameterization, while beneficial for broad datasets (lottery hypothesis), appears excessive for the more specialized requirements of clinical settings. Our findings confirm that a leaner architecture could be equally or more effective in a clinical context, where the data often differ markedly from the vast and varied examples found in large datasets.

Domain specific transfer learning

To test the general character and transferability of the architectures automatically designed by the algorithm we conducted experiments in a similar classification task but using a different dataset. Thus, we used two streamlined architectures, obtained to classify normal versus abnormal activity in multi-view radiograph MURA images, as starting point (pre-trained models) for training a binary classifier on the ChexPert X-ray dataset [22]. The new dataset consists of 24,316 chest radiographs from 65,240 patients labeled for 14 diseases as negative, positive or uncertain. Each study was labeled as either 1 or 0 which represents finding (target-variable) or non-finding activity. The class imbalance in this dataset is 90/10.

As detailed in Fig 3, our two selected compact models small-1 (black) and small-2 (yellow) with 20% and 50% of the nodes respectively achieve equivalent performance to a large model, as the full ResNet50 architecture, both in terms of accuracy and required calculation time (epochs) for an early stopping solution, when trained on the new dataset.

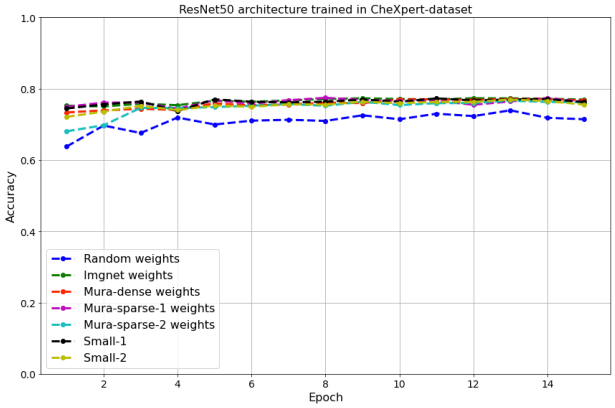


Figure 3: Performance achieved by the leaner architectures obtained (small-1, small-2) in the Mura dataset with respect to the full architecture (ResNet50).

4 Discussion

This study presents a significant advance in the field of deep learning for medical imaging, addressing the common challenge of model over-parameterization without compromising diagnostic accuracy. Our methodology directly confronts the trade-off between model size and performance, offering a solution that caters to the resource-constrained environments typical in healthcare settings. The induced sparsity within the CNNs not only reduces the computational load but also enhances model interpretability—a crucial aspect in clinical decision-making.

Our approach diverges from traditional pruning by incorporating sparse regularization directly into the learning process, thereby optimizing the network architecture intrinsically and eliminating the need for post-hoc adjustments. This automated process allows for a seamless determination of the most efficient architecture for a given dataset, which is particularly beneficial in the medical domain where data diversity and volume may vary substantially.

Moreover, the application of our streamlined models to the ChexPert dataset demonstrates the versatility and transferability of our methodology. The ability of the derived architectures to maintain high performance on a different dataset underscores the potential of our algorithm to serve as a foundation for developing robust, generalizable medical imaging tools.

Our findings also contribute to the ongoing discourse on the necessity of vast parameterization for deep learning models. While larger models have their merits, our results suggest that smaller, more efficient architectures can achieve comparable or superior performance, especially in domain-specific tasks. This aligns with the "lottery ticket hypothesis," which posits that certain subnetworks are inherently well-suited for a given task and once identified, can operate independently of the original, larger network.

5 Conclusion

In conclusion, our novel methodology for CNN model reduction successfully addresses several challenges in medical imaging: the need for large datasets, high computational resources, and model interpretability. By introducing sparse regularization into the training process, we automatically derive compact yet effective network architectures that demonstrate near-equivalent performance to their full-sized counterparts. Our results not only advocate for a paradigm shift toward more computationally efficient models in healthcare but also pave the way for further research into model optimization techniques that do not sacrifice accuracy for efficiency.

The implications of this work are far-reaching, with the potential to democratize access to advanced medical imaging diagnostics across varied clinical environments. Future studies could explore the application of our methodology across other domains of medical imaging and further validate the transferability of the optimized architectures to different tasks and datasets. As deep learning continues to permeate medical technology, our research offers a sustainable, efficient path forward, ensuring high-quality care without the burden of excessive computational demands.

References

[1]- Jiang, F. et al. “Artificial intelligence in healthcare: past, present and future”. Stroke Vasc. Neurol. 2, (2017), p. 230–243

[2]- Yu, K.-H., Beam, A. L. & Kohane, I. S. “Artificial intelligence in healthcare”. Nat. Biomed. Eng. 2, (2018), p. 719–731

[3]- Chen, M., Hao, Y., Hwang, K., Wang, L. & Wang, L. “Disease prediction by machine learning over big data from healthcare communities”. IEEE Access 5, (2017), p. 8869–8879.

[4]- Wang, H. et al. “Predicting hospital readmission via cost-sensitive deep learning”. IEEE/ACM Trans. Comput. Biol. Bioinforma. 15, (2018), p. 1968–1978.

[5]- Lysaght, T., Lim, H. Y., Xafis, V. & Ngiam, K. Y. “AI-assisted decision-making in healthcare”. Asian Bioeth. Rev. 11, (2019), p. 299–314

[6]- Esteva, A. et al. “Dermatologist-level classification of skin cancer with deep neural networks”. Nature 542, (2017), p. 115–118

[7]- Poplin, R. et al. “Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning”. Nat. Biomed. Eng. 2, (2018), p. 158

[8]- Coudray, N. et al. “Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning”. Nat. Med. 24, (2018), p. 1559–1567.

[9]- Lee, J. et al. “BioBERT: a pre-trained biomedical language representation model for biomedical text mining”. Bioinformatics 36, (2020), p. 1234–1240.

[10]- Shen, J. et al. “Artificial intelligence versus clinicians in disease diagnosis: systematic review”. JMIR Med. Inform. 7, e10010 (2019)

[11]- Misha Denil, Babak Shakibi, Laurent Dinh, Nando de Freitas, et al. “Predicting parameters in deep learning”. In Advances in Neural Information Processing Systems, (2013), p. 2148–2156.

[12]- Cho, J., Lee, K., Shin, E., Choy, G. and Do, S. “How much data is needed to train a medical image deep learning system to achieve necessary high accuracy? ”. Preprint at https://arxiv.org/abs/1511.06348 (2015).

[13]- Gentil, M. L. et al. “Factors influencing the development of primary care data collection projects from electronic health records: a systematic review of the literature”. BMC Med. Inform. Decis. Mak. 17, (2017), p. 139.

[14]-Strubell, Emma, Ananya Ganesh, and Andrew McCallum. “Energy and policy considerations for deep learning in NLP”. Proceedings of the 57 th Annual Meeting of the Association for Computational Linguistics, (2019), p. 3645–3650.

[15]- Frankle, M Carbinar. “The lottery ticket hypothesis: Finding sparse, trainable neural networks”. arXiv preprint arXiv:1803.03635, (2018) arxiv.org.

[16]- Stephen Jose Hanson and LorienY Pratt. “Comparing biases for minimal network construction with back-propagation”. In Advances in neural information processing systems, (1989), p. 177–185,

[17]- A G Howard, M Zhu, B Chen, D Kalenichenko, W Wang. “Mobilenets: Efficient convolutional neural networks for mobile vision applications”. - arXiv preprint arXiv:1704.04861, (2017)

[18]- M Tan, Q Le. “Efficientnet: Rethinking model scaling for convolutional neural networks”. International conference on machine learning, (2019) proceedings.mlr.press.

[19]- S Han, J Pool, J Tran, W Dally. “Learning both weights and connections for efficient neural network”. Advances in neural information processing systems, (2015).

[20]- Christian Szegedy, et al. “Going deeper with convolutions”. arXivpreprint arXiv:1409.4842, (2014).

[21]- Yann Le Cun,John S Denker, and Sara A Solla. “Optimal brain damage”. Advances in neural information processing systems, (1990), p.598–605.

[22]- Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo and Robyn Ball and Katie Shpanskaya and Jayne Seekins and David A. Mong and Safwan S. Halabi and Jesse K. Sandberg and Ricky Jones and David B. Larson and Curtis P. Langlotz and Bhavik N. Patel and Matthew P. Lungren and Andrew Y. Ng . “CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison”. arXivpreprint arXiv:1901.07031, (2019).

The implementation of Gaussian Process Regression to Inform Voxel-Based Analysis when Dealing with Sparse and Irregular Time-series Longitudinal Data

Jane Shortall[1], Eliana Vasquez Osorio[1], Andrew Green[2], David Wong[3], Tanuj Puri[1], Peter Hoskin[1,4], Ananya Choudhury[1,4], Marcel van Herk[1] and Alan McWilliam[1]

1 Division of Cancer Sciences - Faculty of Biology, Medicine and Health, The University of Manchester, Manchester, United Kingdom
2 EMBL's European Bioinformatics Institute, United Kingdom.
3 Division of Informatics, The University of Manchester
4 Christie Medical Physics and Engineering, The Christie NHS Foundation Trust, Manchester, United Kingdom

Abstract
Short-term Prostate Specific Antigen (PSA) trajectory has been shown to be prognostic of future biochemical recurrence (BCR), which in turn could be related to too low radiotherapy dose in certain anatomy. PSA follow-up is often irregular and sparse, presenting challenges in population-based studies. We implement Gaussian Process (GP) regression to inform voxel-based analysis aiming at investigating the effect of dose on BCR. Routinely collected PSA data was modelled using GP regression. Interpolated PSA in 6 steps from 0 – 3 years was used as outcome in voxel-based analysis to find regions with significant association of PSA and dose. For this, we used dose distributions mapped to a common reference anatomy, and Spearman's rank test for voxel-wise statistics in a cohort of 219 patients with prostate cancer treated with conformal hypo-fractionated radiotherapy. GP informed voxel-wise analysis was successfully implemented, and revealed a significant correlation between dose and follow-up PSA data at all tested time-points. The identified region identified became larger between 0 – 18 months follow-up, and then shrank. Regardless of the changes over time, dose in the seminal vesicles was consistently identified as significantly correlated with PSA.

1 Introduction

Voxel-based analysis is an established methodology becoming a standard for spatial modeling of outcomes after radiotherapy. It has been used in different anatomical regions for different outcomes[1–5]. This technique has been recently used to show an association between radiotherapy dose and biochemical recurrence (BCR) in prostate cancer patients treated with curative intent radiotherapy, identifying a region in the seminal vesicles where lower dose resulted in worse disease control[6].

BCR, which will occur in one in three men with prostate cancer, is detected via rises in Prostate Specific Antigen (PSA). Current routine follow-up for prostate radiotherapy consists of six-monthly PSA monitoring. Despite regular collection of data, BCR is typically detected via rises in PSA of ≥2ng/ml above nadir (the lowest concentration), which is a patient-defined PSA threshold that can only be defined after the event (NHS NICE guidelines[7]). Further, data is often recorded irregularly and sparsely. This means that recurrence can go undetected between PSA tests for some patients, and more intense follow-up or intervention can be delayed by months.

Studies indicate routinely collected short-term follow-up PSA data utilized differently could offer additional prognostic value, with PSA dynamics acting as a surrogate end-point for future BCR. Several authors suggest close following of PSA can allow for more targeted follow-up and early intervention with salvage treatment[8–13].

In this work, we aim to identify associations between radiotherapy dose and dynamic short-term follow-up PSA data. The main challenge is the way to effectively deal with irregular time-series data. We present a methodology based on Gaussian Process (GP) regression, a common technique for dealing with irregular longitudinal data, to interpolate regularly sampled PSA values. These are then used to inform voxel-based analysis to associate follow-up PSA at determined time-points with planned radiotherapy dose in prostate cancer patient, and follow the evolution of association over time.

2 Materials and Methods

Planning Computed Tomography's (CTs), dose distributions and clinical information, including repeat post-radiotherapy PSA measurements, were collected for prostate cancer patients treated with conformal hypo-fractionated radiotherapy (50 Gy in 16 fractions) between 2005 and 2007 at a single academic centre. Institutional approval had been granted to use this data (research ethics committee reference: 17/NW/0060).

To consider short-term PSA information, only patients with 3-years PSA follow-up, who did not experience BCR within this period, were included. Three years was chosen as a pragmatic and appropriate balance between the small number of patients who are expected to recur within this time period, and the number of patients with sufficient PSA data available[8,9]. All patients had at least two readings during this period.

Modelling PSA data

Log-transformed PSA trajectories were modelled using Gaussian Process regression (MagmaClustR, R (version 4.0.2) in RStudio (desktop version 1.3.1073)). To ensure at least 3-years of data was available for all patients, the first PSA reading > 3-years follow-up was included.

A population-based model was created using the whole cohort of data. The posterior mean for each patient was then sampled at regular time-points (0.5 years) between 0 and 3-years, for use in voxel-based analysis. The posterior mean and 95 % confidence band was plotted for all models and patients and visually inspected to assure a good fit for the data.

To test the robustness of results, individually informed models, where a separate GP model was created for each patient, were also created and analysis repeated and results compared visually.

Dose Mapping

CTs were deformably registered to an arbitrarily chosen reference patient and dose distributions mapped to the same anatomy (in-house software). A region of interest forming a sphere of approximately 10cm radius, centered on the centre of the reference prostate, was chosen for analysis to remove regions where spurious results could occur due to differences in radiotherapy plans depending on patient anatomy[6].

For dose mapping, dose distributions of all patients were flipped in the left-right direction and included in analysis twice[3–5]. Target Registration Error[14] analysis was performed to account for registration uncertainty, and dose distributions blurred using a 3D Gaussian filter of width corresponding to the standard deviation if variation in each cardinal direction of landmarks set at the seminal vesicles and apex (≤0.9cm). All mapped dose distributions were visually inspected.

Voxel-Based Analysis

To assess the correlation between planned radiotherapy dose and post-radiotherapy GP informed PSA at each 6-monthly time-point (0 – 3 years), a Spearman's rank test was performed for each voxel of the dose distribution in the region of interest (in-house software). Correlation coefficients (ρ) for each voxel were collected to create ρ-maps.

To correct for multiple comparisons, and to determine regions of significant correlation, permutation testing was performed (n=1000). PSA values were randomly permuted and the voxel-wise tests performed, recording extreme ρ-statistics for each permutation[2]. From the distribution of extreme ρ-statistics, ρ value corresponding to p≤0.05 were identified and iso-ρ levels plotted on the observed ρ-map to identify the region with significant correlation between dose and PSA at each time-point. Each PSA time-point was analyzed separately.

3 Results

PSA data

Figure 1 illustrates the characteristics of the collected PSA data. The median number of PSA measurements collected within the 3-year period was six [2 - 27], collected an average of 1.2 years apart [0.4 - 4.4], for an average of 6.8 years [3.4 - 10.3].

All PSA data was successfully modelled, and interpolated PSA values used as input for voxel-based analysis. GP informed PSA at 0 months follow-up was, on average, within 0.16 ng/ml of the recorded PSA at 0 months ([0.03 – 1.7 ng/ml], standard deviation 3.1 ng/ml). An example GP sampled posterior mean (0 - 3 years) for one patient is shown in Figure 2.

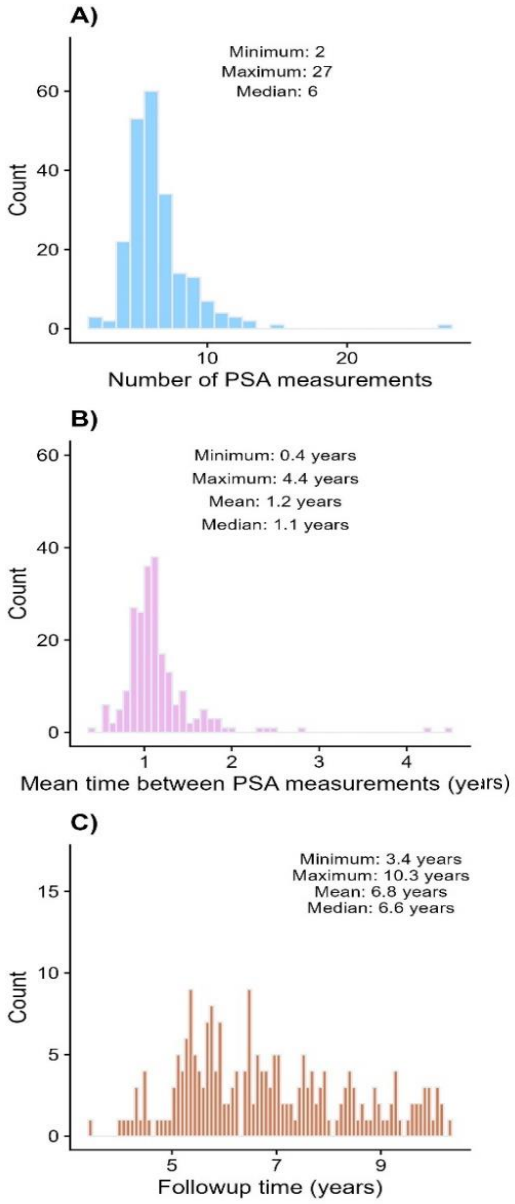


Figure 1: Histograms displaying the distribution of PSA collection across all patients (n=219). Data was highly irregular and, in some cases, sparse. PSA was recorded between 2 and 27 times, on average at 1.2 year intervals, for between 3.4 and 10.3 years.

Voxel-Based Analysis

Figure 3 shows regions of significant correlation between planned radiotherapy dose and PSA at 6-montly intervals between 0 and 48-month follow-up, using A) population-based and B) individually-based GPs ($p \leq 0.05$). For both methods, a large region, indicated by the red overlay, superior to the prostate, covering the seminal vesicles, was detected for all time-points up to 36-months. These regions indicate where a dose and GP informed PSA were negatively correlated, i.e. lower dose correlated with higher PSA.

The region is seen to change between time-points. It appears to grow larger between the beginning of follow-up and 18 months, and then become smaller and eventually disappear at 48 months follow-up. This could indicate that PSA within 18-months of follow-up is most informative of potential treatment failure. Results are consistent between GP methods.

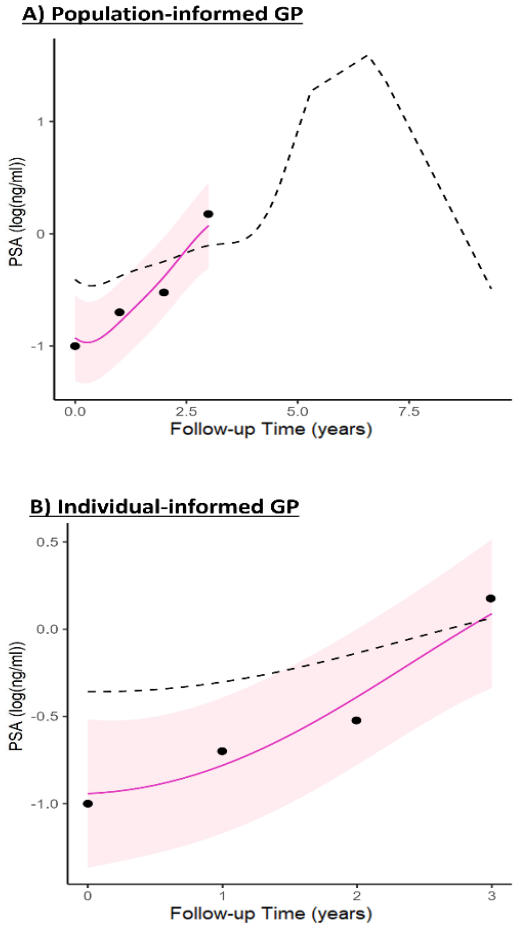


Figure 2: An example population-based (A) and individually-based (B) GP for one patient. Black dotted line represents the mean of each model respectively informed GP model. The pink line shows the posterior mean and 95 % confidence interval for this patient. Note the different scales. Note that the model extends over the 3-year time-period, as the first PSA measurement after 3-years was included to ensure adequate data.

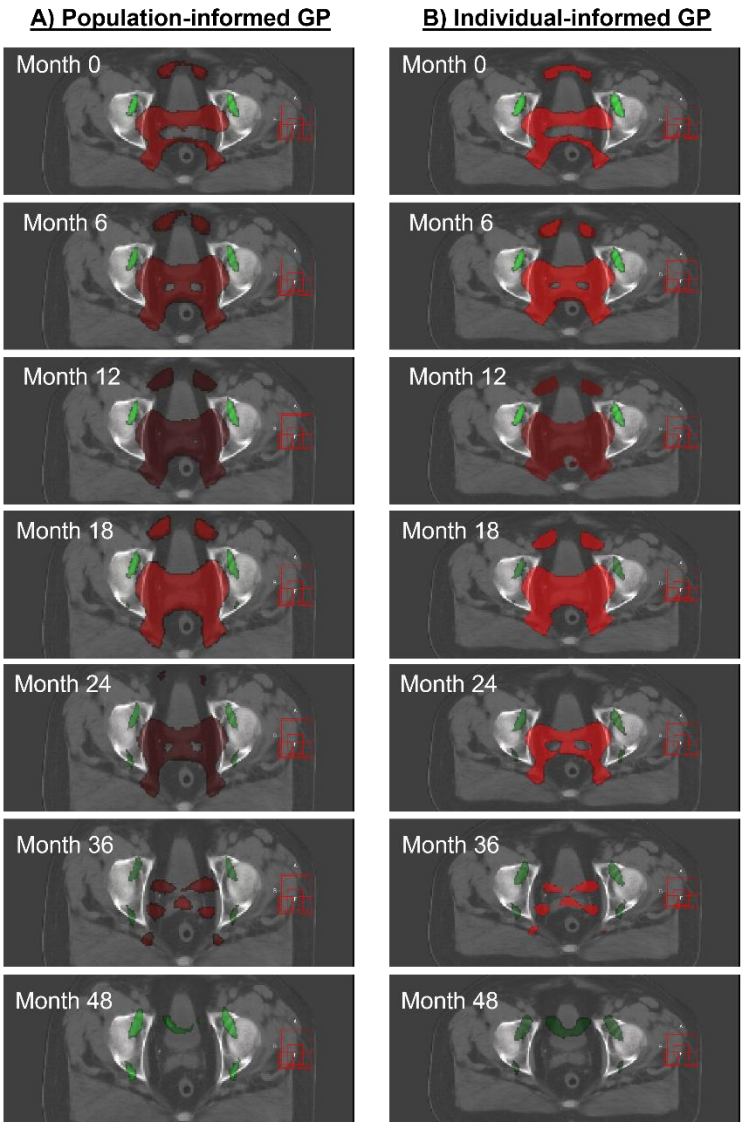


Figure 3: Regions of significant association between planned radiotherapy dose and A) population-based and B) individually-based GP informed follow-up PSA at 6 month intervals between 0- and 48-month post-radiotherapy follow-up. The red and green overlays indicate regions where dose was negatively or positively associated with dose, respectively. Results were consistent across both population- and individually-based GP models.

4 Discussion

Our results demonstrate that irregular and sparse time-series data can be successfully incorporated into GP informed voxel-based analysis to inform outcome models for prostate cancer patients treated with radiotherapy.

We used both population- and individually-based GP models and found similar results after IBDM. The combination of GP and voxel-wise analysis allows the association between dose and irregular time series PSA data to be explored for the first time.

As reported, our results were consistent across population- and individually-based GP models. A simulation study where the ground truth is known, may be a good method to determine which approach is more reliable. The posterior

sampled means of the two methods were, on average, within 0.2 ng/ml ([0.04 – 7.2 ng/ml], standard deviation 2.9 ng/ml), which is not of a magnitude that would change clinical management of a patient. It should be investigated whether the two methods create models that lie within a common confidence bound, however. The robustness of these results could further be tested using a "leave one out" population-based model, where the model is used to predict PSA in unseen data at each available time-point for a direct comparison with measurement.

The identified region varies at across time, with the largest region occurring at 12–18 months into follow-up, before disappearing at 48 months. This could indicate a time window of 0 – 36 months to consider meaningful value from PSA measurements during follow-up, and could be related to recovery after hormone treatment. This should be investigated further, accounting for the time between the end of radiotherapy and beginning of follow-up, which was irregular in our data ([0.01–3.0 years], mean 0.8 years, standard deviation 0.4 years).

Regardless of the changes over time, dose in the seminal vesicles and their surrounding is consistently identified as significantly correlated with PSA, with association growing up until 18 months follow-up. This supports recent work that uses voxel-based analysis to identify a region in the seminal vesicles where radiotherapy dose was significantly associated with BCR[6]. The large size of the region is likely due to spatial correlation of the dose distributions.

We use short-term post-radiotherapy PSA as a surrogate for treatment failure. Both PSA at defined time-points within follow-up, and PSA dynamics have been associated with the prediction of treatment failure for prostate cancer patients[8,10,13,15]. Our results suggest that radiotherapy dose could be influencing follow-up PSA. It should further be explored whether our results can be related to treatment outcome. Once validated in other external datasets, these results could be used to create a tool to re-stratify patients after radiotherapy, allowing for more informed follow-up schedules, earlier intervention, and refined target volumes.

We consider absolute PSA values at defined time-points. Recent work demonstrates how short-term PSA dynamic information can be characterized as one of three typical trajectories using clustering techniques. This work further indicates significant association between short-term PSA trajectory and future BCR, allowing for potential re-stratification during early follow-up[13]. It would be interesting to repeat our analysis using cluster specific prediction curves[16], to associate radiotherapy planned dose with each typical PSA trajectory.

5 Conclusion

We present a method to successfully incorporate sparse time series data into GP informed voxel-based analysis to inform outcome models for prostate cancer patients treated with radiotherapy. Population and individual GP were implemented showing good agreement in inference of PSA and in the VBA.

References

1. Jane Shortall, Giuseppe Palma, Hitesh Mistry, Eliana Vasquez Osorio, Alan McWilliam, Ananya Choudhury, Marianne Aznar, Marcel van Herk, A. G. Flogging a Dead Salmon? Reduced Dose Posterior to Prostate Correlates With Increased PSA Progression in Voxel-Based Analysis of 3 Randomized Phase 3 Trials-Marcello et al. *Int. J. Radiat. Oncol. Biol. Phys.* **110**, 696–699 (2021).
2. Green, A., Vasquez Osorio, E., Aznar, M. C., McWilliam, A. & van Herk, M. Image Based Data Mining Using Per-voxel Cox Regression. *Front. Oncol.* **10**, 1178 (2020).
3. Shortall, J. *et al.* Dose outside of the prostate is associated with improved outcomes for high-risk prostate cancer patients treated with brachytherapy boost. *Front. Oncol.* **13**, (2023).
4. Beasley, W. *et al.* Image-based Data Mining to Probe Dosimetric Correlates of Radiation-induced Trismus. *Int. J. Radiat. Oncol. Biol. Phys.* **102**, 1330–1338 (2018).
5. Mcwilliam, A. *et al.* Radiation dose to heart base linked with poorer survival in lung cancer patients. *Eur. J. Cancer* **85**, 106–113 (2017).
6. Shortall, J. *et al.* Dose outside of the prostate is associated with improved outcomes for high-risk prostate cancer patients treated with Brachytherapy Boost. *Front. Oncol.* **13**, 2461.
7. Prostate cancer Diagnosis and treatment Clinical guideline CG175 Prostate cancer: diagnosis and treatment Contents. (2014).
8. Parr, H., Porta, N., Tree, A. C., Dearnaley FRCR, D. & Hall, E. A Personalised Clinical Dynamic Prediction Model to Characterise Prognosis for Patients with Localised Prostate Cancer: analysis of the CHHiP Phase III Trial. *Int. J. Radiat. Oncol. Biol. Phys.* **0**, (2023).
9. Lorenzo, G. *et al.* Patient-specific forecasting of postradiotherapy prostate-specific antigen kinetics enables early prediction of biochemical relapse. *iScience* **25**, (2022).
10. Johnstone, P. A. S., Williams, S. R. & Riffenburgh, R. H. The 100-day PSA: usefulness as surrogate end point for biochemical disease-free survival after definitive radiotherapy of prostate cancer. *Prostate Cancer Prostatic Dis. 2004 73* **7**, 263–267 (2004).
11. Hall, W. H., Jani, A. B., Ryu, J. K., Narayan, S. & Vijayakumar, S. The impact of age and comorbidity on survival outcomes and treatment patterns in prostate cancer. *Prostate Cancer Prostatic Dis.* **8**, 22–30 (2005).
12. Saad, F. *et al.* Deep Prostate-specific Antigen Response following Addition of Apalutamide to Ongoing Androgen Deprivation Therapy and Long-term Clinical Benefit in SPARTAN. *Eur. Urol.* **81**, 184–192 (2022).
13. Shortall, J. *et al.* GAUSSIAN PROCESS CLUSTERING TO STUDY PSA DYNAMICS IN PROSTATE CANCER PATIENTS. *Phys. Medica* **104**, S66–S67 (2022).
14. Brock, K. K., Mutic, S., McNutt, T. R., Li, H. & Kessler, M. L. Use of Image Registration and Fusion Algorithms and Techniques in Radiotherapy: Report of the AAPM Radiation Therapy Committee Task Group No. 132. *Med. Phys.* (2017) doi:10.1002/mp.12256.
15. Lorenzo, G. *et al.* Patient-specific forecasting of postradiotherapy prostate-specific antigen kinetics enables early prediction of biochemical relapse. *iScience* **25**, (2022).
16. Leroy, A., Latouche, P., Guedj, B. & Gey, S. Cluster-Specific Predictions with Multi-Task Gaussian Processes. (2020).

Voxel-wise causal inference to detect location and dose-response of organs at risk in lung cancer

Marcel van Herk, Azadeh Abravan, Alan McWilliam, Eliana Vasquez Osorio

Division of Cancer Sciences, University of Manchester, UK
The Christie NHS Foundation Trust, Manchester, UK

Abstract. Dose correlations between neighbouring organs complicate NTCP modeling. Sometimes even the dose-sensitive region and NTCP function are unknown. We propose a novel voxel-wise NTCP modelling approach based on instrumental variable analysis, a causal inference technique, to address this problem and apply it to lung cancer patients. Our instrumental variable is the random perturbation of the delivered dose due to an action threshold during historical image-guided radiotherapy (IGRT). We test many NTCP models and perform Cox regression per voxel between ΔNTCP due to perturbation and hazard of death. The largest AIC-significant voxel cluster over all parameters defines region and NTCP model. In a cohort of 609 stage III patients, a dose-response was found in a location close to the sinoatrial node in the heart, with an integrated Gaussian dose-response with D_{50} of 20+/-1Gy and γ of 2+/-0.5Gy. This is the first time voxel-wise causal inference is used in radiotherapy research.

1 Introduction

Estimating dose-effect relationships poses challenges due to various issues. First, it is not clearly known which part of an organ at risk (OAR), such as the heart, is most sensitive [1]. Second, overall survival is often confounded, the outcome not only dependent on the disease itself but also on the toxicity of normal tissues, e.g. the heart in lung radiotherapy (RT) [2]. Such toxicity is often under-reported. The main factors driving dose to OARs are tumour size and location. Through these variables, OAR dose is indirectly associated to cancer survival independent of toxicity (**Figure 1**). To estimate a dose-effect relationship for toxicity we must, therefore, carefully correct for such associations.

Causal inference is the process of determining the causal effect of a particular exposure that is component of a larger system, to a given outcome. We aim, for the first time, to combine a form of causal inference used for natural experiments, an instrumental variable method, with a voxel-wise proportional hazard analysis to allow direct estimation of voxel-wise dose-effect relationships. We detect the most significant dose-sensitive heart region and model its normal tissue complication probability (NTCP) relationship for hazard of death in lung cancer patients.

2 Materials and Methods

As usual, dose distributions were first mapped to a template patient using deformable registration (Niftyreg). Here we mapped both the planned dose distribution and an estimate of the delivered dose distribution based on historic IGRT data with an action threshold of 5mm. The causal directed acyclic diagram (DAG) that represents our method per voxel that incorporates NTCP modeling is shown in **Figure 1**. A change of NTCP due to residual setup error is calculated for an assumed dose-effect relationship and fitted to time-dependent survival data using Cox regression. However, if performed voxel-wise, the change in dose due to residual setup affects only a small number of voxels depending on the steepness of the NTCP function and dose gradients (**Figure 2**).

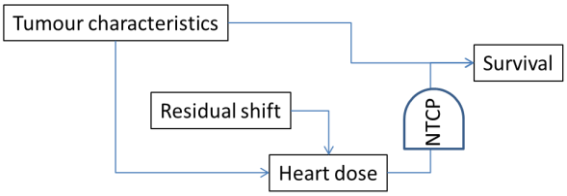


Figure 1: Adapted causal Directed Acyclic Graph (DAG) representing our voxel-wise assumptions and incorporating NTCP modeling. 'Residual shift' is the instrumental variable affecting heart dose randomly. Tumour characteristics affect survival but also voxel dose, because patients with more advanced tumours tend to receive higher doses everywhere. Therefore the planned voxel dose cannot be used directly to estimate NTCP from survival data. Instead the change in NTCP is estimated from random perturbations in dose (here due to the use of image guidance with an action threshold).

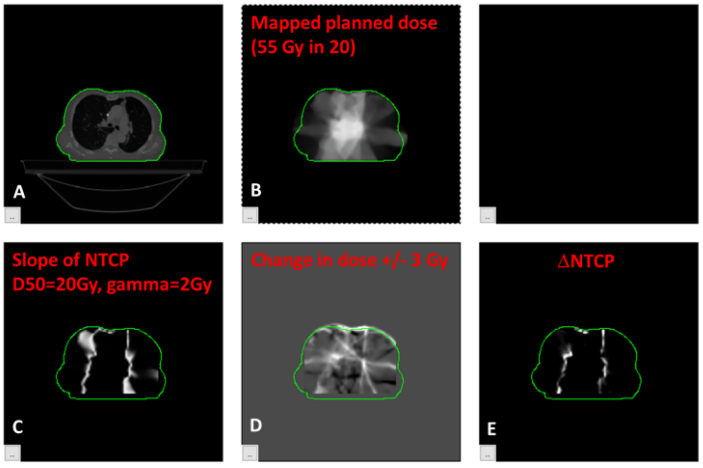


Figure 2: Visualisation of image processing in voxel-wise causal inference. The dose distribution of each patient in the cohort (B) is mapped to a template patient (A) using deformable registration. The slope of an NTCP curve is highest for doses that are close to D_{50}. Effectively, the ΔNTCP is the product of the derivative of the NTCP function (C), bright where high, and the perturbation in dose due to the use of an action threshold (D). As a result, ΔNTCP (E) only differs from zero (black) in few voxels, making its average signal small even if analysing hundreds of patients. For this reason the resolution must be matched to the unknown scale of anatomical structures.

For this reason, the dose resolution is reduced, and the cluster size of significant voxels with a change in Akaike Information Criterion (AIC) over a preset threshold is used as test variable. AIC summarizes the relative quality of a statistical model, and a reduction of 2 points or more indicates a model has better discrimination. The pipeline is illustrated in **Figure 3**. We limited the analysis to the heart in the template patient with 2cm margin, at a voxel size of 2.3x2.3x3 mm^3. This means that for each parameters combination about 300,000 voxel-wise Cox regressions are performed. We used an efficient multithreaded implementation as described in [3]. A previously studied cohort of 609 patients with stage III lung cancer treated with 55Gy in 20 fractions was analysed [4].

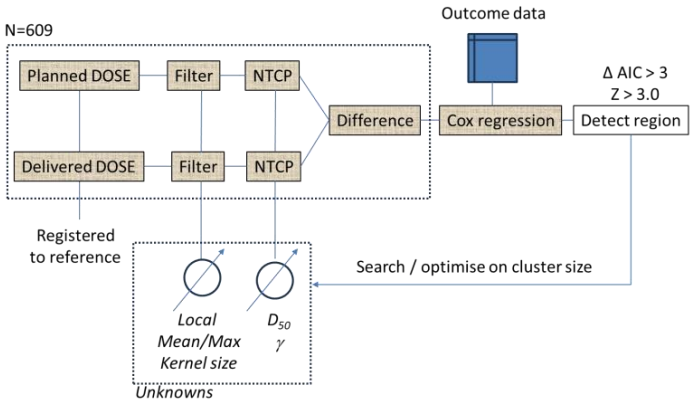


Figure 3: Pipeline for voxel-wise causal inference. Shaded boxes are applied voxel-wise. Unknown parameters are the type of the local filter used to reduce the resolution of the dose distributions and the shape parameters (D_{50} and γ) of the NTCP model. The Cox regression utilizes time-dependent survival from start of radiotherapy of 609 lung cancer patients. An exhaustive search over about 2000 parameter combinations was performed based on the cluster size of significant voxels at a pre-set significance level.

3 Results

Residual setup errors ranged from -5 to 5 mm and were not correlated with age, tumour size, performance status or comorbidities [4]. The combination of applying a single integrated normal dose response-model for planned and delivered dose distributions and performing a univariable Cox regression between their difference and survival for 300,000 voxels took 45-200s on an Intel i7-9700 CPU, depending on the convergence of regressions, which was poor for some parameter combinations. Note that we cached the filtered dose distributions of all patients in memory to improve performance.

We arbitrarily used an AIC reduction of 3 and Z-value over 3 to define significant voxels. The largest clusters were found using a local maximum filter with a kernel size of 10x10x10 mm.



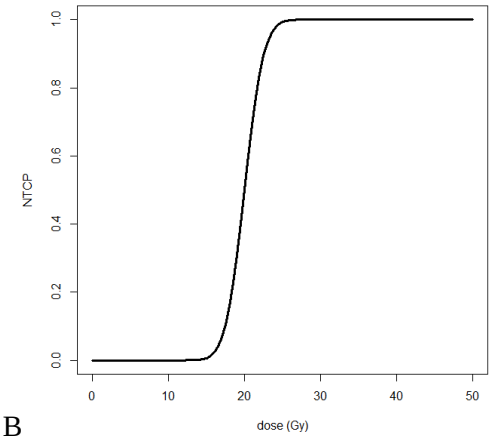


Figure 4: A) Detected heart region at a dose-response of (20Gy, 2Gy) in red. The template patient has a pacemaker and the detected region is located at the end of one lead. Contours illustrate the current heart sparing regions in our rapid-learning trial. The detected significant voxels are near to it but cover less volume. B) Estimated NTCP curve.

Use of a 7x7x7 mm or 12x12x12 mm reduced the number of significant voxels by about a factor 2. Use of a local mean filter reduced the number of the significant voxels over 5 times. The ultimate detected region is close to the sinoatrial node and has its steepest dose response at 20Gy (**Figure 4**).

4 Discussion

This is the first application of voxel-wise causal inference based on instrumental variables that directly integrates NTCP modeling. We were able to pinpoint both a sensitive location in the heart and define its dose-response. By adopting causal inference methods with an instrumental variable, confounding is implicitly dealt with.

There are several limitations to this method. First we rely on an instrumental variable that may not always be available. Moreover, by nature, many voxels will not have a detectable impact on outcome simply by having a dose in an incorrect range or having too small dose perturbations.

To deal with this, we had to lower the resolution of the analysis and found that filter parameter choice was very sensitive. It is likely that with much larger cohorts than our 609 patients, we may need less filtering. Running 6 billion Cox regressions took about 48 hours. To explore the resulting data, a specific browser was developed but exploring the 6-dimensional solution space remains time consuming. For this reason, the browser detects the largest cluster for each tested dose level given a D_{50} and γ.

Our results are in agreement with the published voxel-based-analysis literature for dose sensitive regions of the heart. The region defined here however is smaller and may be more targeted at the clinically relevant region driving early mortality. Once our methodology and results are validated they can help refine ongoing clinical trials of targeted heart sparing for patients with lung cancer treated with radiotherapy.

5 Conclusion

This is the first application of voxel-wise causal inference in radiotherapy outcome modeling and we were able to pinpoint a sensitive location close to the sinoatrial node in the heart as well as its dose-response.

Acknowledgements

We acknowledge the contributions of Corinne Johnson-Hart and Gareth Price in analyzing the IGRT data. Andrew Green is acknowledged for developing the Cox-per voxel software.

References

[1] Banfill K, Giuliani M, Aznar M, Franks K, McWilliam A, Schmitt M, Sun F, Vozenin MC, Faivre Finn C; IASLC Advanced Radiation Technology committee. Cardiac Toxicity of Thoracic Radiotherapy: Existing Evidence and Future Directions. J Thorac Oncol. 2021 Feb;16(2):216-227. doi: 10.1016/j.jtho.2020.11.002

[2] McWilliam A, Abravan A, Banfill K, Faivre-Finn C, van Herk M. Demystifying the Results of RTOG 0617: Identification of Dose Sensitive Cardiac Subregions Associated With Overall Survival. J Thorac Oncol. 2023 May;18(5):599-607. doi: 10.1016/ j.jtho.2023.01.085

[3] Green A, Vasquez Osorio E, Aznar MC, McWilliam A, van Herk M.Image Based Data Mining Using Per-voxel Cox Regression. Front Oncol. 2020 Jul 21;10:1178. doi: 10.3389/fonc.2020.01178

[4] Johnson-Hart C, Price G, McWilliam A, Green A, Faivre-Finn C, van Herk M. Impact of small residual setup errors after image guidance on heart dose and survival in non-small cell lung cancer treated with curative-intent radiotherapy. Radiother Oncol. 2020 Nov;152:177-182. doi: 10.1016/j.radonc.2020.04.008

Pelvic bone marrow dose-volume predictors of late lymphopenia following pelvic lymph node radiation therapy for prostate cancer

Maddalena Pavarini[1], Lisa Alborghetti[1], Stefania Aimonetto[2], Angelo Maggio[3], Valeria Landoni[4], Paolo Ferrari[5], Antonella Bianculli[6], Edoardo Petrucci[7], Alessandro Cicchetti[8], Bruno Farina[9], Paolo Salmoiraghi[10], Eugenia Moretti[11], Barbara Avuzzi[12], Tommaso Giandini[13], Fernando Munoz[14], Alessandro Magli[15], Giuseppe Sanguineti[16], Justyna Magdalena Waskiewicz[17], Luciana Rago[18], Domenico Cante[19], Giuseppe Girelli[20], Vittorio Vavassori[21], Nadia Gisella Di Muzio[22,23], Tiziana Rancati[8], Cesare Cozzarini[23*], Claudio Fiorino[1*]

1 IRCCS San Raffaele Scientific Institute, Medical Physics Dept, Milano, Italy; 2 Ospedale Regionale Parini-AUSL Valle d'Aosta, Medical Physics Dept, Aosta, Italy; 3 Istituto di Candiolo - Fondazione del Piemonte per l'Oncologia IRCCS, Medical Physics Dept, Candiolo, Italy; 4 IRCCS Istituto Nazionale Tumori Regina Elena, UOSD Laboratorio di Fisica Medica e Sistemi Esperti, Roma, Italy; 5 Comprensorio Sanitario di Bolzano, Medical Physics Dept, Bolzano, Italy; 6 IRCCS Crob, Medical Physics Dept, Rionero in Vulture, Italy; 7 ASL TO4 Ospedale di Ivrea, Medical Physics Dept, Ivrea, Italy; 8 Fondazione IRCCS Istituto Nazionale dei Tumori, Unit of Data Science, Milano, Italy; 9 Ospedale degli Infermi, Medical Physics Dept, Biella, Italy; 10 Cliniche Gavazzeni-Humanitas, Medical Physics Dept, Bergamo, Italy; 11 Azienda sanitaria universitaria Friuli Centrale, Medical Physics Department, Udine, Italy; 12 Fondazione IRCCS Istituto Nazionale dei Tumori, Radiotherapy Department, Milano, Italy; 13 Fondazione IRCCS Istituto Nazionale dei Tumori, Medical Physics Department, Milano, Italy; 14 Ospedale Regionale Parini-AUSL Valle d'Aosta, Department of Radiation Oncology, Aosta, Italy; 15 Azienda Ospedaliero Universitaria S. Maria della Misericordia, Department of Radiotherapy, Udine, Italy; 16 IRCCS Regina Elena National Cancer Institute, Department of Radiation Oncology, Roma, Italy; 17 Comprensorio Sanitario di Bolzano, Radiotherapy, Bolzano, Italy; 18 IRCCS Crob, Radiotherapy, Rionero in Vulture, Italy; 19 ASL TO4 Ospedale di Ivrea, Radiotherapy, Ivrea, Italy; 20 Ospedale degli Infermi, Department of Radiotherapy, Biella, Italy; 21 Cliniche Gavazzeni-Humanitas, Radiotherapy Dept, Bergamo, Italy; 22 IRCCS San Raffaele Scientific Institute, Department of Radiation Oncology, Vita-Salute San Raffaele University, Milano, Italy; 23 IRCCS San Raffaele Scientific Institute, Department of Radiation Oncology, Milano, Italy. *These authors share the last authorship

Abstract

To assess clinical/dosimetric predictors of late hematological toxicity in a prospective multi-institute cohort study of prostate cancer (PCa) patients undergoing pelvic-nodes irradiation (PNI).

Clinical, dosimetric and blood test data were prospectively collected, including lymphocytes count (ALC) at different time points. DVHs with irradiated volumes of the body and of four pelvic bones were extracted: ilium (BMILEO), lumbosacral spine (BMLS), lower pelvis (BMPELVIS) and whole pelvis (BMTOT). Current analysis focused on 2-year CTCAE v4.03 grade≥2 (G2+) lymphopenia (ALC<800/μL). Mann–Whitney U tests selected DVH parameters that better discriminate patients with/without toxicity. Twenty-four clinical/dosimetric covariates were considered as potential predictors, including best-selected DVH parameters. Acute CTCAE v4.03 grade≥3 (G3+) lymphopenia was also separately considered as a potential predictor. After data pre-processing, a multi-variable logistic regression model was identified. Bootstrap-based internal validation was finally performed.

Forty-six out of 499 patients (9.2%) experienced Late G2+ lymphopenia. ALC at baseline [HR=0.997, 95% CI 0.996-0.998, p<0.0001], smoke (yes/no) [HR=2.9, 95% CI 1.25-6.76, p=0.013] and BMLS-V≥24Gy (cc) [HR=1.006, 95% CI 1.002-1.011, p=0.003] were retained for the model. When acute G3+ lymphopenia (yes/no) was considered, it was included in the model [HR=4.517, 95% CI 1.954-10.441, p=0.0004]. Performances were moderately high and internal validation confirmed the robustness of the results.

Two-year lymphopenia after PNI for PCa is not negligible and mostly modulated by baseline ALC. Multivariate analysis showed BMLS-V24 as the most prominent dosimetry predictor with baseline ALC and smoking. An independent role of acute G3+ lymphopenia was also ascertained.

1 Introduction

The interest in quantifying hematological toxicity (HT), with a special focus on severe dropout of blood counts, after radiotherapy (RT) without chemotherapy is high, aiming to estimate the impact of the "RT-induced" symptoms [1, 2, 3, 4, 5].

In particular, as the reduction of lymphocytes relates to the immune response, stimulating cytokine secretion and helping anti-tumor immunity [6], lymphopenia could translate into detrimental effects of anti-cancer therapies [7]. Clearly these effects mainly depend on the delivered dose, on the anatomical location and on the extent of the Planning Target Volume (PTV). Effects are expected to occur acutely by the direct killing of the circulating lymphocytes and the incidental irradiation of BM, as primary site of hematopoiesis [8]. A situation at high risk of moderate/severe RT-induced lymphopenia concerns the irradiation of large volumes of the pelvic region; more than half of the active BM is located at this level [9] and consequently not negligible effects especially occur in case of pelvic nodes irradiation (PNI), mostly for bladder and prostate cancer (PCa) [4].

In this scenario, a multi-centric study was started in 2012 (IHU-WPRT TOX [1, 2, 10, 11, 12]) to assess predictors of intestinal, hematological and urinary toxicity after PNI for PCa, delivered with intensity-modulated radiotherapy (IMRT). Data were mature for current study, with the aim of 1) quantifying dose-volume relationships between BM structures and 2-year moderate/severe lymphopenia; 2) developing predictive models of 2-year lymphopenia combining BM (and possibly Body) dosimetry information with clinical predictors.

2 Materials and Methods

In September 2012 a national multi-centric institutional review board-approved study was activated at San Raffaele Scientific Institute, Milan, to evaluate clinical/dosimetric predictors of intestinal, hematological and urinary toxicity in WPRT (ClinicalTrials.gov identifier NCT2803086).

Eight hundred eighty-seven PCa patients were enrolled up to the end of 2021, irradiated with radical or post-surgery IMRT, including PNI, and followed for 5-years: the trial involved 14 clinical centers, the coordinating Institute included. Details on aims, patient characteristics, planning and volumes, dose/fractionation, delivery and methodology

of the study were previously reported [1, 2, 10, 11, 12] as well as preliminary results regarding the single institute pilot study on HT [1, 2].

Four pelvic bone volumes were prospectively contoured: ilium (BMILEO), lumbosacral spine (BMLS), lower pelvis (BMPELVIS) and whole pelvis (BMTOT) [13].

Full planning data (planning CT, RT-Plan, RT-Dose, RT-Structure DICOM files) of the PNI part of the treatment were collected to incorporate dosimetric predictors into the analysis at a later stage.

Dose-volume histograms (DVHs) of bones and body were extracted in absolute values (cc) in the range 0-70Gy with a 1Gy pitch.

The cranial extent (mm) from the L5 vertebrae lower edge to the PTVLN upper extremity was automatically computed through a proper script in Python as the vertical distance obtained by the combination of the baseline CT image and the contoured structures. To achieve this, L5 vertebras were previously delineated with TotalSegmentator [14].

The PTVLN volume (cc) was retrieved, along with the cranial extent, to evaluate the impact of spatial extension of irradiated volumes - target and organs at risk (OARs) - on late HT.

Blood tests, including lymphocytes counts (ALC), were collected at various intervals over 5 years to analyze late grade≥2 (G2+) lymphopenia, being ALC at 2 years (or 30 months if missing) after RT conclusion<800/µL (CTCAE v4.03) [15].

Patients with available 3D planning, baseline clinical data and with complete baseline, RT mid- and/or endpoint, 3 or 6 months, 1 year or 18 months, 2 years or 30 months blood tests were included.

Mann–Whitney U tests were preliminarily performed to assess the DVH regions of each BM and Body structure that better discriminated patients with/without toxicity (corresponding to p-values<0.05). The selected volumes were then checked for Pearson's pairwise correlation. When Pearson's correlation coefficient was deemed to be sufficiently high (ρ>0.7) the volume with the lowest p-value at the Mann–Whitney U test was kept as the best DVH predictor.

Twenty-four clinical/dosimetric covariates were considered as potential predictors, including best-selected DVH parameters. In addiction, acute CTCAE v4.03 grade≥3 (G3+) lymphopenia (ALC at its lowest value between mid and end RT<500/µL) was separately considered as a potential predictor.

Preprocessing and training used the information available for the included patients while 1000 bootstrap replicates were generated for internal validation. Univariate logistic regression assessed the effect of each covariate on the binary outcome of toxicity. For those variables with a significance level p<0.3, pairwise Spearman correlation was measured. In case of high-level correlation (ρ>0.7) between two variables maintained at univariable, the one with the lowest p-value was considered.

The training phase tuned a backward stepwise multiple logistic regression, aimed to identifying feature subsets that maximize the average precision (AP) as the primary selected performance indicator. The AP consists in the area under the precision-recall curve (AUPR) which, when maximized, assures the best tradeoff between high precision and high recall for different thresholds, particularly suitable when the classes are very imbalanced.

During the internal validation procedure, bootstrap resampling determined the robustness of the model by iteratively resampling the training population with replacement [16]. The chosen performance indicators (average precision and ROC curve AUC) were then corrected for optimism.

Raw data manipulation and analysis were performed with an in-house Python 3.7.9 code: medicalAI library in mAItre (Medical Artificial Intelligence Toolkit for REsearch).

3 Results

In current analyses, 499 patients respecting the selection criteria were included. Forty-six patients (9.2%)

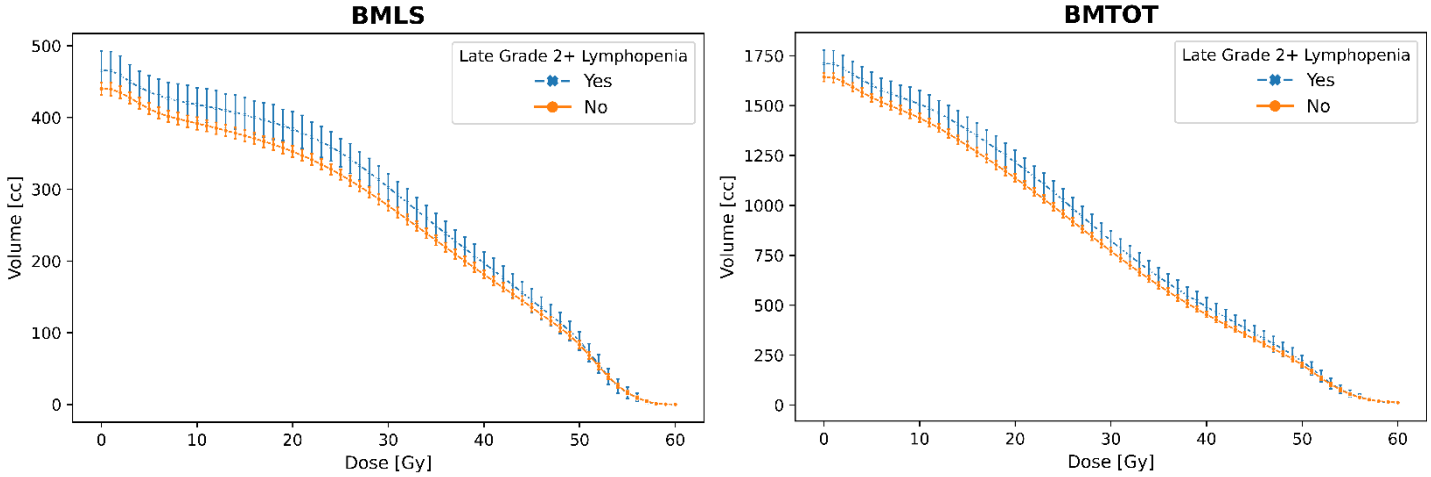


Figure 1: Mean absolute DVHs of BMLS (left) and BMTOT (right) for patients with and without late G2+ lymphopenia with 95% confidence intervals. Significant differences (p-values<0.05) at the Mann–Whitney test were found in the ranges V0-V36Gy for BMLS and V0-V5Gy and V10-V29Gy for BMTOT.

experienced late G2+ lymphopenia. DVHs of structures were stratified by the presence of the endpoint and significant differences (p-values<0.05) at the Mann–Whitney test were found in the ranges V0-V36Gy, V15-V24Gy, V0-V5Gy and V10-V29Gy, V20-V32Gy and V38-V43Gy for BMLS, BMPELVIS, BMTOT and BODY respectively (Figure 1). The latters were compared pairwise to measure the strength of their inter-correlation. Of them, BMLS-V24 was finally selected as the best independent DVH variable so that eleven features were found to be eligible to enter the training phase.

The variables which the backward sequential feature selection retained for late G2+ lymphopenia were: ALC at baseline (μL^{-1}) [OR=0.997, 95% CI 0.996-0.998, $p<6.1*10^{-11}$], smoke (yes/no) [OR=2.840, 95% CI 1.254-6.432, p=0.01] and BMLS-V≥24Gy (cc) [OR=1.006, 95% CI 1.002-1.011, p=0.002]. The ROC and PR AUC (hereinafter referred to as the "apparent" performance) were 0.865 and 0.415, respectively, suggesting that the model was performing well in distinguishing between classes overall, with modest precision in the positive class. The McFadden's pseudo-R-squared (0.259) and LLR p-value (p<0.0001) assessed the goodness of fit.

In Figure 2, the resulting 2-year risk of G2+ lymphopenia is shown against ALC at baseline, stratified by Smoke and dichotomized BMLS-V24 (above/below the Youden optimal cutoff value of 307cc).

During the internal validation stage, the "apparent" performance in the original sample was corrected for the mean "optimism" across the bootstrap replicates, resulting in a more realistic performance of 0.859 and 0.379 for ROC and PR AUC, respectively.

When acute G3+ lymphopenia (yes/no) was added among the possible predictors it was retained in addition to the previous three variables: ALC at baseline (μL^{-1}) [OR=0.997, 95% CI 0.997-0.998, $p=5*10^{-9}$], smoke (yes/no) [OR=2.902, 95% CI 1.246-6.761, p= 0.014], the fraction of lumbosacral volume receiving≥24Gy (cc) [OR=1.006, 95% CI 1.002-1.011, $p=3.5*10^{-3}$] and acute G3+ lymphopenia [OR=4.517, 95% CI 1.954-10.441, $p=4.2*10^{-4}$].

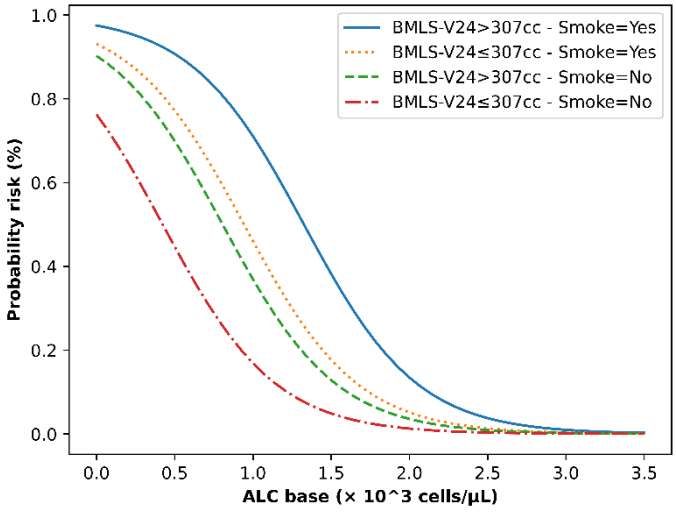


Figure 2: Probability risk (%) for 2-year G2+ lymphopenia at increasing levels of ALC at baseline, stratified by Smoke and dichotomized BMLS-V24 (above/below the Youden optimal cutoff value of 307cc.

In this case, the performance on the original dataset for ROC and PR AUC slightly improved with a value of 0.878 and 0.489, respectively, as well as McFadden's pseudo-R-squared (0.308) and LLR p-value (p<0.0001).

The internal validation statistics (ROC AUC=0.870, PR AUC=0.446) accounted for the natural decrease in performance in the bootstrap sample compared to the original sample.

DVH constraints to reduce late G2+ lymphopenia were determined for BMLS (V24<307cc) and BMTOT (V10<1520cc, V20<1250cc, V30<850cc) via the Youden index (J) method: the optimal cut-points simultaneously maximized sensitivity and specificity for toxicity classification.

4 Discussion

To our knowledge, current study for the first time shows that, in a large multi-centric patients' cohort treated with IMRT, moderate/severe lymphopenia still affects patients 2 years after RT with a G2+ rate of 9.2% (G3+=0.8%) against a baseline (pre-RT) rate of 1.8%. On the other hand, the rates of acute G2+ and G3+ lymphopenia (86.8% and 39.5% respectively) were in line with other investigations [5] and with our previous single institution reports [1, 2].

A second major outcome is the quantification of the dose-volume effect of pelvic BM and sub-structures. Interestingly, the lumbosacral BM showed the strongest association with late lymphopenia and its corresponding DVH surrogate value (V24Gy) was retained in the resulting multi-variable models. Our results suggest that, along with an expected large risk reduction when limiting the cranial PTV limit, sparing BM may translate in a minor but still significant risk mitigation even for patients' sub-groups with similar cranial limit.

A significant association between the whole pelvic BM and the risk of late lymphopenia was also found, leading to suggest the following DVH constraints: V10<1520cc, V20<1250cc and V30<850cc.

The presence of two clinical predictors corroborates the evidence that G2+ lymphopenia is largely modulated by the baseline patient situation, being baseline ALC the strongest predictor. The smoking effect seems to suggest a likely relevant negative role in micro-circulation modifications.

The final model was also slightly improved when considering acute G3+ lymphopenia as a potential predictor of 2-year lymphopenia, suggesting a quite intuitive relationship between the entity of ALC in the acute phase and recovery ability and speed.

The information that patients with lower baseline ALC value and/or smoking have a much higher risk of experiencing persistent G2+ lymphopenia could help the clinician during treatment planning. Considerations could

also be made on the real expected benefit in including PNI, maybe small or questionable for a not negligible fraction of patients, especially in the post-operative setting [17]. On the other hand, the detrimental effect of persistent lymphopenia in terms of decreased QoL and/or anti-tumor effect for PCa is not clear yet.

As pelvic BM is not usually included in the planning optimization phase, there could be substantial improvements in reducing the risk of 2-year lymphopenia thanks to refined planning strategies [3] or considering, when feasible and justified, the use of protons [18]. The choice of the cranial limit of the fields could also significantly impact this risk and could be discussed by clinicians when considering the cost-benefit of the therapy, especially in patients with low ALC and/or smokers.

5 Conclusion

The risk of 2-year lymphopenia after PNI for PCa is not negligible and mainly depended on pre-RT ALC, smoking and pelvic BM DVH, with evidence of a major impact of the cranial extension of PTVLN including a larger portion of lumbar vertebrae in the irradiated volume.

The knowledge of the risk factors is important for the radiation oncologist in the treatment decision phase and its optimization. There should be a large room to improve the sparing of pelvic BM without detriment to other OARs and/or PTV coverage, at least in patients with unfavorable clinical factors (i.e.: baseline ALC, smoking).

References

[1] Cozzarini C, et al. Hematologic Toxicity in Patients Treated With Postprostatectomy Whole-Pelvis Irradiation With Different Intensity Modulated Radiation Therapy Techniques Is Not Negligible and Is Prolonged: Preliminary Results of a Longitudinal, Observational Study, Int J Radiat Oncol Biol Phys, 2016, 95(2):690-695, DOI: 10.1016/j.ijrobp.2016.01.022.

[2] Sini C, et al. Dose–volume effects for pelvic bone marrow in predicting hematological toxicity in prostate cancer radiotherapy with pelvic node irradiation, Radiother Oncol, 2016, 118(1):79-84, DOI: 10.1016/j.radonc.2015.11.020.

[3] Papanikolaou P, et al. NTCP modeling and dose-volume correlations of significant hematocrit drops 3 months after prostate radiation therapy, Advances in Radiation Oncology, 2023, 101393, DOI: 10.1016/j.adro.2023.101393.

[4] Danckaert W, et al. Impact of radiotherapy parameters on the risk of lymphopenia in urological tumors: A systematic review of the literature. Radiother Oncol. 2022;170:64-69. DOI: 10.1016/j.radonc.2022.02.030.

[5] Schad MD, et al. Radiation-related Lymphopenia after Pelvic Nodal Irradiation for Prostate Cancer, Advances in Radiation Oncology, 2019, 4(2):323-330, DOI: 10.1016/j.adro.2019.01.005.

[6] Paganetti H, A review on lymphocyte radiosensitivity and its impact on Radiotherapy, Front. Oncol., 2023, Volume 13, Article number 1201500, DOI: 10.3389/fonc.2023.1201500

[7] Terrones-Campos C, et al. Hematological toxicity in patients with solid malignant tumors treated with radiation – Temporal analysis, dose response and impact on survival, Radiother Oncol, 2021,158:175-183, DOI: 10.1016/j.radonc.2021.02.029.

[8] Ebrahimi S, et al. A hybrid deep learning model for forecasting lymphocyte depletion during radiation therapy, Med Phys. 2022; 49: 3507–3522, DOI: 10.1002/mp.15584

[9] Hayman JA, et al. Distribution of Proliferating Bone Marrow in Adult Cancer Patients Determined Using FLT-PET Imaging, Int J Radiat Oncol Biol Phys, 2011, 79(3):847-852, DOI: 10.1016/j.ijrobp.2009.11.040.

[10] Sini C, et al. Patient-reported intestinal toxicity from whole pelvis intensity-modulated radiotherapy: First quantification of bowel dose–volume effects. Radiother Oncol, 2017, 124(2):296-301, DOI: 10.1016/j.radonc.2017.07.005.

[11] Bresolin A, et al. Predictors of 2-Year Incidence of Patient-Reported Urinary Incontinence After Post-prostatectomy Radiotherapy: Evidence of Dose and Fractionation Effects. Front Oncol, 2020, Volume 10, Article Number 1207, DOI: 10.3389/fonc.2020.01207.

[12] Bresolin A, et al. Acute patient-reported intestinal toxicity in whole pelvis IMRT for prostate cancer: Bowel dose-volume effect quantification in a multicentric cohort study Radiother Oncol, 2021, 158, 74-82, DOI: 10.1016/j.radonc.2021.02.026.

[13] Mell LK, et al., Dosimetric predictors of acute hematologic toxicity in cervical cancer patients treated with concurrent cisplatin and intensity-modulated pelvic radiotherapy, Int J Radiat Oncol Biol Phys, 2006, 66(5):1356-1365, DOI: 10.1016/j.ijrobp.2006.03.018.

[14] Wasserthal J, et al. TotalSegmentator: Robust Segmentation of 104 Anatomic Structures in CT Images, Radiology: Artificial Intelligence, 2023, DOI: 10.1148/ryai.230024

[15] National Cancer Institute, National Institutes of Health, US Department of Health and Human Services. Common Terminology Criteria for Adverse Events (CTCAE) Version 4.0. Published May 28, 2009; Revised version 4.03 June 14, 2010; https://evs.nci.nih.gov/ftp1/CTCAE/CTCAE_4.03/CTCAE_4.03_2010-06-14_QuickReference_8.5x11.pdf

[16] Steyerberg EW, et al. Internal validation of predictive models: Efficiency of some procedures for logistic regression analysis, J Clin Epidemiol, 2001, 54, 774–781, DOI: 10.1016/S0895-4356(01)00341-9.

[17] Cozzarini C, et al., Accurate prediction of long-term risk of biochemical failure after salvage radiotherapy including the impact of pelvic node irradiation, Radiother Oncol, 2022, 175, 26-32, DOI: 10.1016/j.radonc.2022.08.001.

[18] Kim N, et al. Reduced radiation exposure to circulating blood cells in proton therapy compared with X-ray therapy in locally advanced lung cancer: Computational simulation based on circulating blood cells. Front Oncol, 2023;13:1119173. DOI: 10.3389/fonc.2023.1119173.

Inter-patient registration methods for voxel-based analysis in lung cancer

Chloe Min Seo Choi[1], Jue Jiang[1], Maria Thor[1], Joseph Deasy[1] and Harini Veeraraghavan[1]

[1]Department of Medical Physics, Memorial Sloan Kettering Cancer Center, New York, United States

Abstract Accurate registration between patients is crucial for reliable radiotherapy outcomes prediction that utilizes voxel-based analysis (VBA). We compared three registration methods, including: 1) an iterative non-deep learning intensity-based B-spline deformable image registration method available in open-source Elastix called SyN, 2) a published deep learning method called patient-specific anatomy and shape context (PACS), and 3) an enhanced version, called PACS-tumor aware, which specifically considers relevant structures that may change, e.g., tumors. The algorithms were tested on a dataset of 3D planning CT images from 240 patients with lung cancer. Our analysis showed that PACS-tumor demonstrated minimal differences in tumor volume (1.08%) and GTV dose (0.05Gy) between non-deformed and deformed moving images, preserving tumors effectively. Additionally, it exhibited robust registration accuracy and consistency across patients of different sizes. In conclusion, PACS-tumor is a well-performing method for inter-patient registration in VBA for lung cancer patients.

1 Introduction

Voxel-based analysis (VBA) has emerged as an important tool to understand the radiation therapy (RT) dose-response relationship for various toxicities after RT for cancer [1-4]. VBA refers to pixel- or voxel-based methods that examine how doses are distributed locally, moving beyond traditional organ-focused approaches using Dose Volume Histograms (DVH). Unlike typically used DVH-based approaches, which offer a 2D aggregation of the 3D dose distribution, VBA incorporates the complete dose distribution. This could allow the identification of specific spatial dose-response relationships.

The success of VBA relies heavily on accurate image registration. This involves aligning individual patients' images onto a common reference anatomy. To date, numerous registration algorithms have been utilized for inter-patient registration. Of these, Symmetric normalization (SyN), which is an iterative optimization-based deformable image registration (DIR) method [5], has been the most widely used DIR approach for inter-patient registration [6, 7]. More recently, deep-learning (DL)-based approaches have been developed, with one example being Transmorph, which has a hybrid Transformer-ConvNet architecture [8]. Moreover, we previously developed a Patient-specific Anatomic Context and Shape prior model (PACS), which uses 3D convolutional long short-term memory networks [9]. Using this approach, large differences in anatomy are handled by computing a spatially and temporally varying dense deformation field by incrementally refining registrations between patients.

However, registration accuracy is impacted by the anatomic variabilities between patients such as various sizes of the heart and other organs due to factors such as gender, race as well as the presence of tumors, which are arbitrarily distributed within the chest. Hence, VBA depends on the selection of canonical reference patients to whom all the other patients are aligned to perform the analysis. So far, the selection of such a patient has been performed by using a patient with median lung or heart volume without considering the impact of the presence of tumors. In fact, patient scans with tumors can adversely impact the accuracy of the analysis due to the difficulty of aligning patient scans without causing unrealistic stretching of the normal tissues as well as erosion of tumors.

Tumor erosion on the aligned moving image occurs when tumors are absent on the target or fixed image, leading to physically unrealistic stretching of normal tissues. The spatial locations of tumors on the fixed image contribute to this issue. In previous studies, patients with unsuccessful reregistration were excluded from the analysis, aiming to mitigate inaccuracies [10, 11]. However, excluding patients is problematic because it creates an inaccurate representation of the general population.

To address this issue, we introduce a novel registration model, "PACS-tumor," designed to prevent inadequate tumor warping while still yielding satisfactory registration outcomes. The aim of this study was to compare the accuracy of different registration methods to assess how the choice of model affects inte-rpatient registration for VBA analysis. Additionally, we explore the potential demographic bias inherent in these methods.

2 Materials and Methods

This retrospective study was approved by the local institutional review board at the Memorial Sloan Kettering Cancer Center. The dataset included 240 locally advanced non-small cell lung cancer (LA-NSCLC) patients treated with intensity-modulated RT to 60 Gy in 2 Gy fractions. Each patient had a treatment planning CT (pCT) with associated gross tumor volume (GTV) and organ-at-risks (OARs) including the heart, esophagus, normal lungs, and cardiac substructures (Inferior Vena Cava (IVC), left and right ventricle (LV, RV), left and right atrium (LA, RA), pulmonary artery (PA), superior vena cava (SVC)).

For each patient, the pCT was re-sampled to a 1x1x3mm resolution. The pCT was thereafter cropped in the z-axis to extract a 3D volume that includes the visible lung regions with an additional 5 slices kept above and below the visi-

ble lung region to have a uniform field of view of all the patient scans. A reference patient with a population median lung volume was chosen as the initial reference anatomy. To ensure pCT alignments, all pCTs were affinely pre-registered to the reference patient before applying any DIR. Air and table occurring outside the body were excluded by first extracting the body region using automatic thresholding of the CT density followed by morphological processing. The largest connected component was extracted, and the image volume was cropped with a 30-pixel extension around the body region to focus deformable image registration towards relevant parts of the image that contain the body and the organs.

Three registration configurations were compared: 1) SyN, 2) PACS [9], and 3) novel PACS-tumor model, which also input tumor conditioning in addition to bi-directional tumor rigidity, thereby preserving tumor volumes and preventing unrealistic stretching of the adjacent normal tissues.

To calculate the registration accuracy, the Dice similarity coefficient (DSC) between the OARs of the aligned patients and the canonical patient was computed. DSC is defined by $DSC = (2 \times |X \cap Y|) / (|X| + |Y|)$ where DSC=0 means no overlap and DSC=1 means perfect overlap between the moving and reference. We also computed the difference in the Planned Tumor Dose (ΔPTD) between undeformed and deformed moving images. The smaller the difference, the better the algorithm's ability in preserving the tumor and its associated dose. Next, to estimate the potential impact of registration quality on VBA, the number of patients with lung segmentation DSC scores ≤ 0.7 was computed, which would indicate poor registration. Furthermore, the tumor volume change (ΔT) was calculated and defined by $(|V_{def}-V_m| / |V_m|) \times 100$ where V_{def} and V_m indicate tumor volume after and before DIR respectively.

We further investigated the algorithm's robustness to patient size variations to mitigate potential population bias arising from the reference patient selection. Here, we observed the correlation between the DSC of the heart and lungs and the patient's body size. The body size was determined by calculating the volume of voxels corresponding to the body region in a single CT slice (slice number 60).

3 Results

Standard PACS showed (Figure 1a) physically unrealistic tissue stretching in locations containing tumors whereas PACS-tumor effectively preserved tumors while also avoiding unrealistic tissue stretching (Figure 1b). The DSC scores measuring registration accuracy for the thoracic OAR is shown in Table 1. The overall DSC was the highest for PACS followed by PACS-tumor and SyN. PACS and PACS-tumor produced comparable accuracies. Registration quality indicated by lung DSC ≤ 0.7, showed that PACS resulted in the least number of excluded patients followed by PACS-tumor (Table. 2). SyN resulted in an exclusion of the greatest number of patients.

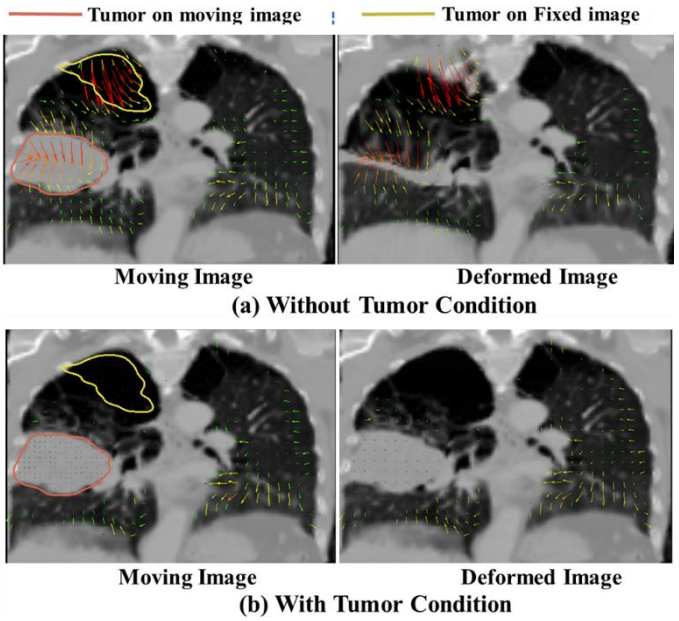


Figure 1: Rationale for tumor conditioning. Large deformations and physically unrealistic stretching of normal tissue in the regions occupied by tumors occurring without tumor conditioning (a) is avoided when using tumor conditioning (b).

	SyN	PACS	PACS-tumor
Heart	0.67	$0.77^{*\dagger}$	$0.74^{\dagger}$
Esophagus	0.18	$0.24^{*\dagger}$	$0.20^{\dagger}$
Lungs	0.71	$0.87^{*\dagger}$	$0.81^{\dagger}$
IVC	0.25	$0.29^{\dagger}$	$0.29^{\dagger}$
LV	0.31	$0.44^{*\dagger}$	$0.41^{\dagger}$
LA	0.38	$0.41^{\dagger}$	$0.41^{\dagger}$
RA	0.26	$0.34^{*\dagger}$	$0.32^{\dagger}$
RV	0.22	$0.28^{*\dagger}$	$0.25^{\dagger}$
PA	0.29	$0.34^{*\dagger}$	$0.32^{\dagger}$
SVC	0.16	$0.26^{*\dagger}$	$0.19^{\dagger}$

Table 1: DSC scores of the OARs for registration methods measured for multiple structures. IVC: inferior vena cava; LV: left ventricle; LA: Left atrium; RA: right atrium; RV: right ventricle, PA: pulmonary artery; SVC: superior vena cava. †: significantly greater than SyN ($p<0.05$), *: significantly greater than PACS-tumor ($p<0.05$)

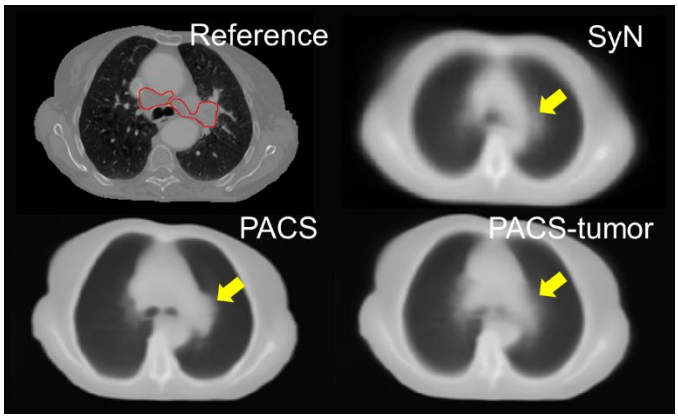


Figure 2: Voxel-wise averages of the patients' anatomical images using different methods. The initial image depicts the reference patient with the GTV outlined in red.

PACS-tumor preserved tumor volumes to the highest degree as shown by the lowest tumor loss ΔT metrics (Table 2). PACS on the other hand was thest least effective in preserving tumor volumes, indicating the importance of tumor conditioning used in PACS-tumor. SyN method achieved minimal ΔPTD and ΔT, suggesting that overall deformation was negligible. This was also visually confirmed as in Figure 2. The tumor in the canonical reference patient was located in the medial region. The voxel-wise average of the PACS and PACS-tumor show better inter-patient agreement compared with SyN. In the tumor region, however, PACS methods induced warping of surrounding tissues whereas the extent of warping is much less with PACS-tumor. SyN shows large inter-patient discrepancies everywhere.

	Excluded cases (Lung DSC ≤ 0.7)	Tumor dose difference, ΔPTD (Gy)	Tumor volume difference, ΔT (%)
SyN	89	0.004	3.40
PACS	7	7.59	46.69
PACS-tumor	19	0.05	1.08

Table 2: Suitability for VBA. Number of excluded patients (heart and lung DSC < 0.7), tumor dose preservation ΔPTD (Gy), and tumor volume difference ΔT (%) by various methods measured.

In the analysis of the impact of patient size on registration accuracy, we found that the overall trend across all patients did not show a significant impact of size on registration accuracy. However, larger patients had notably lower DSC values for both algorithms (Figure 2). Overall, the registration models showed similar trends.

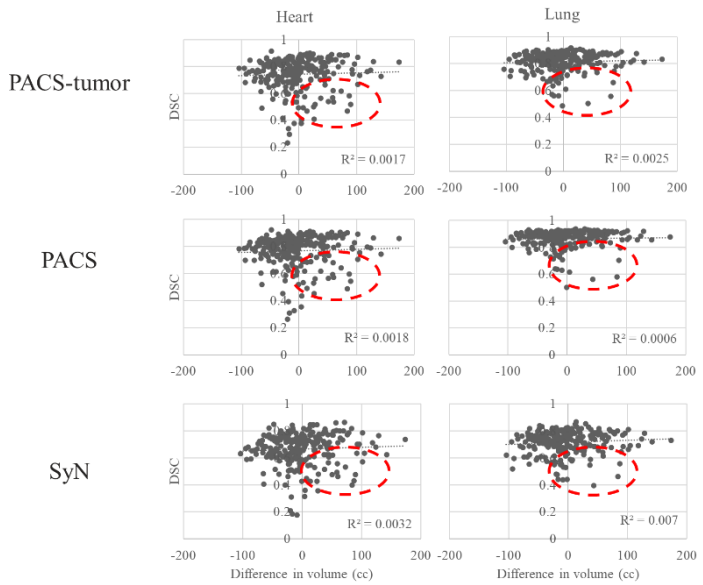


Table 2: Figure 2: Correlation between patient size and the DSC accuracy.

4 Discussion

To our knowledge, no previous studies have investigated how the choice of registration tool affects VBA in RT. This study emphasizes the importance of considering both registration accuracy and dose preservation when selecting an algorithm. Our analysis revealed that PACS-tumor demonstrated superior accuracy in preserving conditioned structures, preventing unrealistic anatomy deformations in locations with tumors, especially in comparison to the PACS registration methods.

Although PACS achieved slightly higher accuracy in terms of OAR, it unrealistically warped tumor volumes to mimic the reference patient's tumor anatomy, resulting in large variations in tumor shrinkage and subsequent dose. This is critical for VBA, as deformed dose mapping is a central analysis component. SyN, on the other hand, was the least accurate method, which resulted in the largest number of excluded patients due to poor registration. Notably, iterative registration methods are frequently used in VBA analysis as they are relatively easier to apply without requiring training on task-specific datasets. However, poor registration quality and the need to exclude patients can also diminish the utility of these analysis methods.

Considering population bias, a slight trend in patient size was observed in which patients with larger body sizes than the reference presented worse registration outcomes. While the trend may be slight, it highlights the importance of investigating the influence of patient size on registration accuracy and fairness.

There are some limitations of this study. This study did not include the downstream impact on VBA and its subsequent prediction. This will be the end goal of future studies. Furthermore, this study used a female median patient as a reference, analysis with a male reference patient will be conducted to observe gender-related differences. Additionally, exploring alternative strategies to balance accuracy and inclusivity, such as patient-specific optimization or hybrid approaches, could prove beneficial.

5 Conclusion

This study has demonstrated that the success of inter-patient VBA is strongly dependent on the algorithm. The novel PACS-tumor model showed significantly better performance in preserving tumor anatomy while maintaining robust registration accuracy across patients of varying sizes. This promising performance suggests its potential application in VBA workflows.

References

[1] O. Acosta *et al.*, "Voxel-based population analysis for correlating local dose and rectal toxicity in prostate cancer radiotherapy," *Phys. Med. Biol.*, vol. 58, no. 8, p. 2581, Mar. 2013, doi: 10.1088/0031-9155/58/8/2581.

[2] S. Monti *et al.*, "A novel framework for spatial normalization of dose distributions in voxel-based analyses of brain irradiation outcomes," *Phys. Med.*, vol. 69, pp. 164–169, Jan. 2020, doi: 10.1016/j.ejmp.2019.12.017.

[3] E. Mylona *et al.*, "Rectal and Urethro-Vesical Subregions for Toxicity Prediction After Prostate Cancer Radiation Therapy: Validation of Voxel-Based Models in an Independent Population," *Int. J. Radiat. Oncol.*, vol. 108, no. 5, pp. 1189–1195, Dec. 2020, doi: 10.1016/j.ijrobp.2020.07.019.

[4] G. Palma *et al.*, "A Voxel-Based Approach to Explore Local Dose Differences Associated With Radiation-Induced Lung Damage," *Int. J. Radiat. Oncol.*, vol. 96, no. 1, pp. 127–133, Sep. 2016, doi: 10.1016/j.ijrobp.2016.04.033.

[5] B. B. Avants, C. L. Epstein, M. Grossman, and J. C. Gee, "Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain," *Med. Image Anal.*, vol. 12, no. 1, pp. 26–41, Feb. 2008, doi: 10.1016/j.media.2007.06.004.

[6] B. B. Avants, N. J. Tustison, G. Song, P. A. Cook, A. Klein, and J. C. Gee, "A reproducible evaluation of ANTs similarity metric performance in brain image registration," *NeuroImage*, vol. 54, no. 3, pp. 2033–2044, Feb. 2011, doi: 10.1016/j.neuroimage.2010.09.025.

[7] A. Klein *et al.*, "Evaluation of volume-based and surface-based brain image registration methods," *NeuroImage*, vol. 51, no. 1, pp. 214–220, May 2010, doi: 10.1016/j.neuroimage.2010.01.091.

[8] J. Chen, E. C. Frey, Y. He, W. P. Segars, Y. Li, and Y. Du, "TransMorph: Transformer for unsupervised medical image registration," *Med. Image Anal.*, vol. 82, p. 102615, Nov. 2022, doi: 10.1016/j.media.2022.102615.

[9] J. Jiang and H. Veeraraghavan, "One shot PACS: Patient specific Anatomic Context and Shape prior aware recurrent registration-segmentation of longitudinal thoracic cone beam CTs," *IEEE Trans. Med. Imaging*, vol. 41, no. 8, pp. 2021–2032, 2022.

[10] E. Mylona *et al.*, "Voxel-Based Analysis for Identification of Urethrovesical Subregions Predicting Urinary Toxicity After Prostate Cancer Radiation Therapy," *Int. J. Radiat. Oncol.*, vol. 104, no. 2, pp. 343–354, Jun. 2019, doi: 10.1016/j.ijrobp.2019.01.088.

[11] N. Gil, M. L. Lipton, and R. Fleysher, "Registration quality filtering improves robustness of voxel-wise analyses to the choice of brain template," *NeuroImage*, vol. 227, p. 117657, Feb. 2021, doi: 10.1016/j.neuroimage.2020.117657.

Modelling progressive anatomical changes and their variability during radiotherapy treatment

Poppy Nikou[1,2,3], Eliana M. Vásquez Osorio[5,6], Marcel Van-Herk[5,6], Teresa Guerrero Urbano[7,8], Anna Thompson[4], Andrew Nisbet[3], Sarah Gulliford[3,4], Jamie McClelland[1,2,3]

[1]Centre for Medical Image Computing (CMIC) and [2]Wellcome / EPSRC Centre for Interventional and Surgical Sciences (WEISS),
[3]Department of Medical Physics and Biomedical Engineering, University College London, London, UK
[4] University College London NHS Foundation Trust Hospital, London, UK
[5]Division of Cancer Sciences, School of Medical Sciences, Faculty of Biology, Medicine and Health, University of Manchester, Manchester, UK
[6]Christie NHS Foundation Trust Hospital, Manchester, UK
[7]Guys and St Thomas NHS Foundation Trust Hospital, London, UK
[8]Kings College London, London, UK

Abstract Progressive anatomical changes are often observed in head and neck cancer patients over the course of radiotherapy. To create proton therapy treatment plans that are robust against these changes, these potential anatomical changes must be known in advance.
In this work we present a novel methodology for developing statistical population models that capture progressive anatomical changes and their variability. The models are continuous in time, and can be built from a population of patients with a varying number and frequency of images over treatment. The statistical population models are based on principal component analysis of the progressive anatomical changes, and can generate scenarios of many potential anatomical changes. In this work two principal components were used to generate nine scenarios of potential anatomical change. The evaluation found, for some patients, one scenario that matched their progressive anatomical changes better than by just considering the average change.

1 Introduction

Inter-fractional anatomical changes during proton therapy can lead to uncertainties in the delivered dose distribution. These uncertainties may reduce the efficacy of treatment, typically increasing dose to some organs at risk and/or reducing dose to the tumour. In an attempt to mitigate for inter-fraction changes, proton therapy treatment plans are optimized on uncertainty scenarios considering rigid shifts and range uncertainty of the protons. However, more complex non-rigid anatomical changes are commonly observed during treatment, such as weight loss and tumour shrinkage or growth. These changes are often progressive over treatment resulting in systematic errors in the delivered dose.

Ideally, a patient would be administered a treatment plan which is robust against predictable anatomical changes. To achieve this, a patient's potential anatomical changes need to be known in advance.

Statistical population models have been explored for this purpose[1-4]. However, most models presented in the literature do not explicitly model progressive changes and therefore, cannot efficiently represent the potential anatomical changes. The models in the literature which do explicitly model progressive changes require imaging at the same time points during treatment, which is not always available.

We present a novel methodology for building statistical population models of progressive anatomical changes. The models can predict trajectories of potential anatomical changes over the course of treatment, and do not require imaging on consistent days in the population for training.

2 Materials and Methods

2.1 Datasets

Datasets from three institutions were used for this study, as shown in Table 1, together with details of how the datasets were split between training and testing.

Dataset	Training	Testing
1	46	5
2	27	1
3	0	14

Table 1 – Table showing the split of patients provided from three institutions. The number of patients which were included for training and testing the population models are shown.

All patients were treated with photon radiotherapy. Each patient had a planning CT (pCT) and a number of CBCTs (5-26), over the course of treatment. The staging, tumour location and prescription doses varied across the patients in all three datasets. To make the population more anatomically consistent, images were flipped so that all patients had the same tumour laterality. The degree of anatomical changes, and hence the need for replanning, varied across the 3 datasets: in Dataset 2 no patients were replanned; in Dataset 3 all patients were replanned; and in Dataset 1 some patients were replanned but most were not.

The statistical population models were built from the training datasets using the following steps: 1) inter-patient correspondence was determined to spatially align all patients into a model space. 2) Intra-patient registrations were used to estimate the anatomical changes of each patient over the course of treatment. 3) Patient specific models were fit to capture the progressive anatomical changes. 4) Statistical population models were then built from the patient specific models. The statistical population models aim to capture the average progressive anatomical changes and the variability in these changes as seen in the training set.

2.2 Inter-Patient Correspondence

To build population models the data from individual patients must be mapped into a common model space so they are spatially aligned. A groupwise registration between the pCTs from Dataset 3 was used to build an anatomical atlas and define the model space.

Inter-Patient correspondence was achieved by aligning all patients to the anatomical atlas. A rigid, affine and deformable registration were performed in turn, and the resulting transformations were composed to calculate $T_{M\to p}$. The transformation $T_{M\to p}$ maps each point in the model space, M, to the corresponding point in the patient space, p, and can be used to resample the patient's images into the model space using pull interpolation. All registrations in this work were performed using NiftyReg[6]. The deformable registrations were parameterized by stationary velocity fields, v, using a regular grid of control points defining a B-Spline Free-Form-Deformation (FFD). Therefore, the registrations produced diffeomorphic transformations that ensured topology was preserved, provided both the forward ($T_{M\to p}$) and inverse ($T_{p\to M}$) transformations, and gave a rigorous framework for building the population models[7,8].

2.3 Intra-Patient Correspondence

For each patient, all CBCT images were rigidly aligned to the patient's pCT, and then transformation $T_{M\to p}$ was used to resample all images into the model space. The anatomical changes were then estimated by registering the pCT to each CBCT in the model space. Each intra-patient registration resulted in a transformation, $\tilde{T}_{reg,p,t\to 0}$, which maps from the CBCT image at time point t to the pCT image for a patient p, in model space, and is parametrized by the stationary velocity field $\tilde{v}_{reg,p,t\to 0}$.

2.4 Patient Specific Models

The patient specific models (SMs) aim to estimate $\tilde{v}_{reg,p,t\to 0}$ for any time point t in treatment, $t = 1 \dots T$. A cubic B-Spline was fit to each parameter of $\tilde{v}_{reg,p,t\to 0}$ i.e. to the *x, y, z* components of each FFD control point, over time. Four SM control points were placed equidistantly over time for each parameter, so that the B-Spline was defined between the first CBCT scan and day 35, i.e. $T = 35$. The B-Spline function was fit using a linear least squares solution. The SMs are defined by a grid of control points, $C_{SM,p}$, which define B-Splines in both the spatial and temporal dimensions. Therefore, the SMs provide very compact models of the anatomical changes over space and time. The SM can be used to estimate $\tilde{v}_{SM,p,t\to 0}$, the SM velocity field, and $\tilde{T}_{SM,p,t\to 0}$, the corresponding SM transformation, for patient p at any timepoint t between the first and last CBCT scans.

2.5 Statistical Population Models

To create the statistical population model (PM) principal component analysis was performed on $C_{SM,p}$ from all training patients. The resulting eigenvalues, λ, and principal components, w, are used to generate potential representations of C_{PM},

$$C_{PM} = \bar{C} + \sum_{i=1}^{N} a_i \sqrt{\lambda_i} w_i$$

where $\bar{C}$ is the average of $C_{SM,p}$ over all training patients, a_i defines the number of standard deviations taken in the ith principal component, and N is the number of principal components used in the model. In this work, we set $N = 2$ and used the PM to generate nine different instances of C_{PM} with $a_1,\ a_2 \in \{-1, 0, +1\}$. The nine potential anatomical changes can then be generated for any timepoint t up to day 35, by calculating $\tilde{v}_{PM,t\to 0}$ and $\tilde{T}_{PM,t\to 0}$ from C_{PM}.

2.6 Transporting Transformations

To apply $\tilde{T}_{PM,t\to 0}$ or $\tilde{T}_{SM,t\to 0}$ to a patient, the transformation needs to be transported from model space to patient space. This is done using conjugate action,

$$T_{t\to 0} = T_{M\to p}(\tilde{T}_{t\to 0}(T_{p\to M}))\,,$$

where $T_{t\to 0}$ is the transported model transformation of patient p. This is shown schematically in Figure 1.

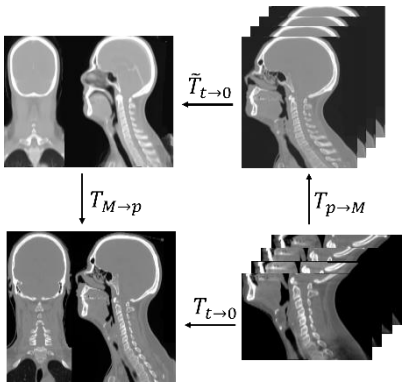


Figure 1 - Schematic representation of inter- patient correspondence $T_{M\to p}$, and it's inverse $T_{p\to M}$. Intra-patient correspondence is shown in patient space $T_{t\to 0}$ and model space $\tilde{T}_{t\to 0}$.

3. Evaluation and Results

We previously presented and evaluated an average population model[9], similar to the average component of PM proposed here, $a_1 = a_2 = 0$. This was built and evaluated using just Dataset 3. This preliminary study indicated that the average models could provide a good estimate of the progressive anatomical changes seen in most patients, but not all, due to the variability of the changes seen across the population of patients. Therefore, this study aims to test whether by including the nine different scenarios of anatomical changes of the PM, provided by the nine instances of C_{PM}, improves on the performance of the average model at estimating the actual anatomical changes seen in the test patients.

A geometric evaluation was performed to quantify the progressive anatomical changes and test the ability of the PM and average model to capture these. When evaluating

the PM models all nine C_{PM} were considered in combination, since no single instance of C_{PM} was expected to be better for all patients. Therefore, for each patient the instance of C_{PM} that provided the minimum error was used. In this evaluation, the SMs were used to provide the gold standard progressive anatomical changes for the test patients, as they have been validated in previous experiments[9]. The evaluation was performed at the last day of the models range by using the models to generate transformations with t = 35.

The transformations generated by the gold standard SM, average model, and PM were transported to the patient space and used to resample the pCT structures for each patient in the test set. The structures included the brainstem, spinal cord, left and right parotid gland, high and low dose clinical target volume (CTV), and body (only available for dataset 3). The structures resampled by the SM were compared on average surface distance to the structures (1) from the pCT, (2) resampled by the average model (3) resampled by the PM (Figure 2). The average surface distance for Datasets 1&2 is shown in Figure 3 (which are from the same institution as the patients used to build the model) and the test set from Dataset 3 in Figure 4 (from an external institution). Note that axes are scaled differently on both figures.

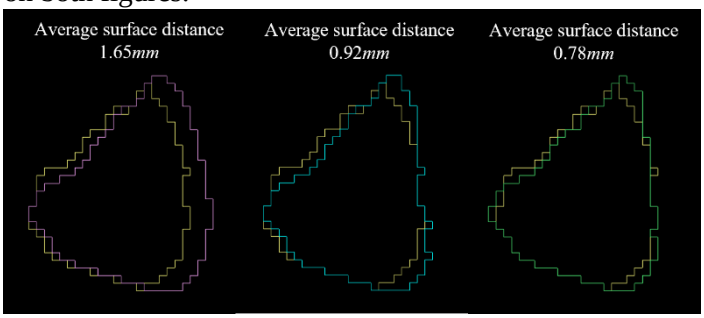


Figure 2 – Example left parotid gland structure resampled with patient specific model (yellow) compared to 1) pCT contour (pink), 2) pCT structure resampled with average model (blue), and 3) pCT structure resampled with best fitting population model (green). This structure is from a test patient in Dataset 3.

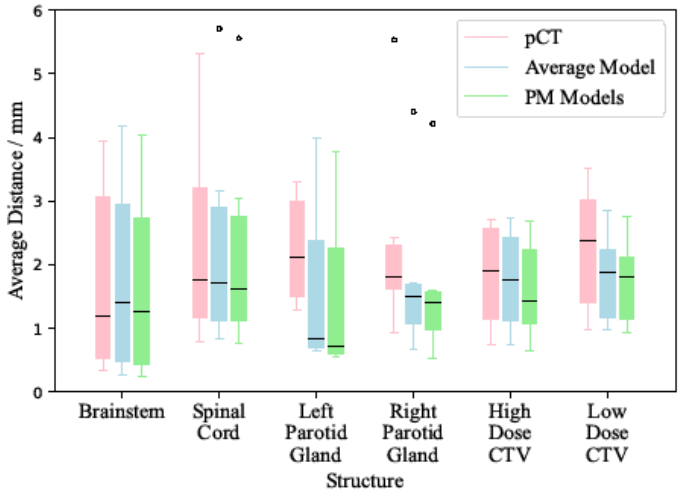


Figure 3 – Average surface distance calculated between the planning CT (pCT) structures resampled with the patient specific model and (1) the pCT structures, (2) structures resampled using the average model and (3) structures resampled using the population model (PM) which provides the best correspondence. The median, range and interquartile range are shown. The test data used in this comparison is from the same institutions from which the models were built.

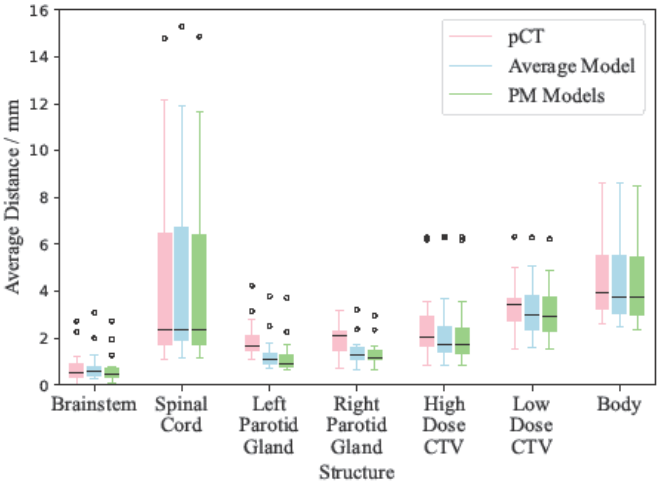


Figure 4 – Average surface distance calculated between the planning CT (pCT) structures resampled with the patient specific model and (1) the pCT structures, (2) structures resampled using the average model and (3) structures resampled using the population model (PM) which provides the best correspondence. The median, range and interquartile range are shown. The data used in this comparison is a hold-out test set; patients were treated at a different institution and all patients in this cohort were replanned.

Larger anatomical changes were seen in the pCT comparison and therefore over treatment in Dataset 3 than in Datasets 1&2. In Datasets 1&2 all anatomical changes are <2.39*mm*. In Dataset 3, the low dose CTV and body structures exceed that with median average surface distances of 3.45*mm* and 3.93*mm*. Larger anatomical changes were expected in Dataset 3, since all patients in this cohort were replanned.

The average model provides an improvement on the median average surface distance across all Datasets and structures, except for the brainstem. For Datasets 1&2, and 3, the average improvement in median average surface distance is 0.46*mm* and 0.44*mm*, respectively.

The PM model provides an improvement over the average model for all structures and datasets. For Datasets 1&2, and 3, the average improvement in median average surface distance is 0.14*mm* and 0.10*mm*, respectively. This supports the hypothesis, that for some patients there exists a scenario from the nine PM generated scenarios that matches the progressive anatomical changes better than the average model.

4 Discussion

In this work we present a novel methodology for developing statistical population models, for progressive anatomical changes and their variability as seen in head and neck cancer patients over the course of treatment. The models are continuous, despite the studied population of patients having a different number of images taken at different days.

The datasets used in this study were from three different institutions and contained patients which had a wide range of tumour types, locations and sizes. Although an attempt was made to partially mitigate for any location differences, by flipping left-right, these factors introduce additional variability into the models.

All patients used in this work were treated with photon radiotherapy. The aimed application of these models however, is to enable the generation of proton radiotherapy plans that are robust to the expected anatomical changes. As these changes are strongly linked to nutritional deficiency and primary tumour/nodal shrinkage, we expect the changes to present similarly in the two patient cohorts. This will be investigated in future work, and if necessary new models could be built from a cohort of proton therapy patients using the same methodology.

Anatomical changes experienced by a patient were estimated using deformable image registrations. Running the registration with the parameters used in this work typically required up to 3 hours and 3GB RAM per registration. So for the approximately 1500 registrations required for this study a considerable amount of computation was required.

For each patient, a SM was built using the registration results, to extract the progressive anatomical changes and remove day to day random variations in anatomy. A cubic B-Spline with four temporal control points was fit to the registration results over time. Preliminary experiments indicated this could approximate the progressive changes well without overfitting the data, even when relatively few images were available (the minimum number of CBCTs was five in this study). However, future work will investigate the number of control points used.

The anatomical changes experienced by a patient, were encoded into a regular grid of spatially and temporally aligned control points, which define B-Splines. This produces a very compact model, which is smaller than the daily voxel wise information by a factor of 10^4.

A geometric evaluation, using pCT structures showed that the statistical population models are able to capture progressive anatomical changes, and that considering the nine scenarios generated by the PM improves on just considering the average model for some patients.

The PM models considered in this work involve only two principal components. These generated nine scenarios of potential anatomical changes per patient. By considering more principal components, the total variability included in the model would increase and therefore, so would the number of scenarios. The aim of this work is to incorporate the predicted anatomical scenarios into robust optimization for proton therapy treatment planning. It is expected that the more scenarios which are included in the optimization, the more robust the treatment plan becomes. This, however, often comes at the cost of conformality and computation time. In future work, we will search for an optimal balance, by investigating how the number of scenarios impacts the robustness and conformality of a proton therapy treatment plan.

5 Conclusion

Head and neck cancer patients commonly experience anatomical changes over the course of their radiotherapy. In this work we present a novel methodology to model the progressive anatomical changes and their variability. The models are continuous in time and can estimate the changes for any day during treatment.

The statistical population models were based on principal component analysis. The evaluation showed that by using just two principal components to generate nine scenarios, the PM improves on using the average model for some patients.

In future work we aim to investigate the use of the generated scenarios in robust proton therapy treatment planning.

Acknowledgements

The authors would like to acknowledge Matthew Tedder, Sarah Mission-Yates, Victoria Butterworth and Marina Khan for helping to provide patient data. This work is supported by the EPSRC-funded UCL CDT in Intelligent, Integrated Imaging in Healthcare (i4health) (EP/S021930/1), the Dept. of Medical Physics and Biomedical Engineering, UCL and the Dept of Health's NIHR funded Biomedical Research Centre at UCLH.

References

1. *Argota-Perez, R., et.al. E.: Evaluating principal component analysis* models for representing anatomical changes in head and neck radiotherapy. Physics and Imaging in Radiation Oncology 22, pp.13–19 (2022).
2. Chetvertkov, M.A., et.al..: Use of regularized principal component analysis to model anatomical changes during head and neck radiation therapy for treatment adaptation and response assessment. Medical Physics 43(10), 5307–5319. (2016).
3. Tsiamas, P., et. al..: Principal component analysis modeling of Head-and-Neck anatomy using daily Cone Beam-CT images. Medical Physics 45(12), 5366–5375 (2018).
4. Zhang, Y., et al., E.: DIR-based models to predict weekly anatomical changes in head and neck cancer proton therapy. Physics in Medicine & Biology 67(9), 095001 (2022).
5. Veiga, C., et. al.: Atlas construction and spatial normalisation to facilitate radiation-induced late effects research in childhood cancer. Physics in Medicine & Biology, 66(10), (2021).
6. GitHub *KCL-BMEIS/niftyreg*, https://github.com/KCL-BMEIS/niftyreg, last accessed 2022/05/23.
7. Arsigny, V., Commowick, O., Pennec, X., Ayache, N. (2006). A Log-Euclidean Framework for Statistics on Diffeomorphisms. In: Larsen, R., Nielsen, M., Sporring, J. (eds) Medical Image Computing and Computer-Assisted Intervention – MICCAI 2006. MICCAI 2006. Lecture Notes in Computer Science, vol 4190. Springer, Berlin, Heidelberg.
8. J. Ehrhardt, R. Werner, A. Schmidt-Richberg and H. Handels,: Statistical Modeling of 4D Respiratory Lung Motion Using Diffeomorphic Image Registration. IEEE Transactions on Medical Imaging 30(2), 251-265 (2011)
9. P.Nikou et.al. (2023, 18-21 April). Modelling anatomical changes during radiotherapy treatment. IEEE International Symposium of Biomedical Imaging, Cartagena de Indias, Colombia
10. Cacicedo, J., et.al.: (2013). A prospective analysis of factors that influence weight loss in patients undergoing radiotherapy. *Chinese Journal of Cancer*.

An anatomically-informed correspondence initialisation method to improve learning-based registration for radiotherapy

Edward G. A. Henderson[1], Marcel van Herk[1], Andrew F. Green[2], and Eliana M. Vasquez Osorio[1]

[1]Division of Cancer Sciences, The University of Manchester, M13 9PL, UK
[2]European Bioinformatics Institute, EMBL-EBI, Cambridge, UK

Abstract We propose an anatomically-informed initialisation method for interpatient CT non-rigid registration (NRR), using a learning-based model to estimate correspondences between organ structures. A thin plate spline (TPS) deformation, set up using the correspondence predictions, is used to initialise the scans before a second NRR step. We compare two established NRR methods for the second step: a B-spline iterative optimisation-based algorithm and a deep learning-based approach. Registration performance is evaluated with and without the initialisation by assessing the similarity of propagated structures. Our proposed initialisation improved the registration performance of the learning-based method to more closely match the traditional iterative algorithm, with the mean distance-to-agreement reduced by 1.8mm for structures included in the TPS and 0.6mm for structures not included, while maintaining a substantial speed advantage (5 vs. 72 seconds).

1 Introduction

Non-rigid registration (NRR) is widely used for many applications in radiotherapy to spatially align images, including fusion of complementary images, contour propagation, evaluating treatment response, dose mapping [1, 2]. Traditional image registration algorithms function by performing iterative updates to parameters of a chosen deformation model to optimise either an intensity- and/or feature-based metric [1]. Intensity-based metrics use the images greyscale intensity information directly to evaluate similarity. Feature-based metrics use features extracted from the images such as anatomical landmarks, structure boundaries, contours and fiducial markers to evaluate alignment.

Hybrid image registration techniques have been developed which aim to combine the strengths of intensity- and feature-based algorithms, leveraging greyscale information as well as anatomical structures or landmarks to guide the registration [3, 4]. However, hybrid methods tend to be slow and labour intensive, taking a long time to set up as they require landmark annotation or structure contouring.

Recently image registration methods using deep learning (DL) have garnered much attention as they can offer large time savings over iterative algorithms [5]. DL approaches are trained ahead of time on large datasets of examples; once trained, these methods can estimate a full deformation vector field in a single function evaluation, often taking less than a second on modern hardware [6]. However, current learning-based approaches are predominantly applied to easier tasks such as the registration of brain MRI scans where the inter-subject variability is relatively low and the soft-tissue contrast is high [5, 7]. There has been little evidence to show that learning-based registration methods are suitable for large deformations, such as inter-subject CT registration [6, 8].

Many learning-based registration approaches are trained to optimise intensity-based metrics. Some can leverage segmentations present in the images, but only at training time through semi-supervised learning [5]. In this study we incorporate ideas from hybrid registration methods to improve learning-based registration performance. To achieve this we introduce a new anatomical correspondence-informed non-rigid initialisation to align labelled anatomical structures prior to using established methods to register non-labelled tissue.

2 Materials and Methods

We used an open-access dataset of 31 head and neck CT scans with highly consistent organ-at-risk (OAR) segmentations [9].

2.1 Registration pipelines

We propose the use of a previously developed learning-based correspondence model [10] and a thin plate spline (TPS) to form an anatomically informed method to initialise inter-patient non-rigid registration. The impact of the proposed initialisation was compared for two established NRR methods: a) NiftyReg, a traditional B-spline iterative-optimisation based image registration method [11] and b) Voxelmorph, an unsupervised, DL-based approach [12]. From now on, the proposed initialisation method is referred to as CorrTPS.

2.1.1 Anatomically informed initialisation (*CorrTPS*)

Correspondence model In a previous study, we showed an unsupervised geometric learning model could be trained to identify dense, point-to-point, correspondences for head and neck OARs [10]. The developed model can automatically identify corresponding anatomical points on the surface of 3D organ shapes and was shown to be effective for the brainstem, mandible, parotid glands, spinal cord and submandibular glands. The input for the correspondence model is two organ shapes which are converted to triangular meshes. The model predicts an assignment linking all vertices from the source to the target organ. Correspondence predictions are based on the organ geometry alone, however, CT imaging was utilised in the training process to improve prediction quality [10].

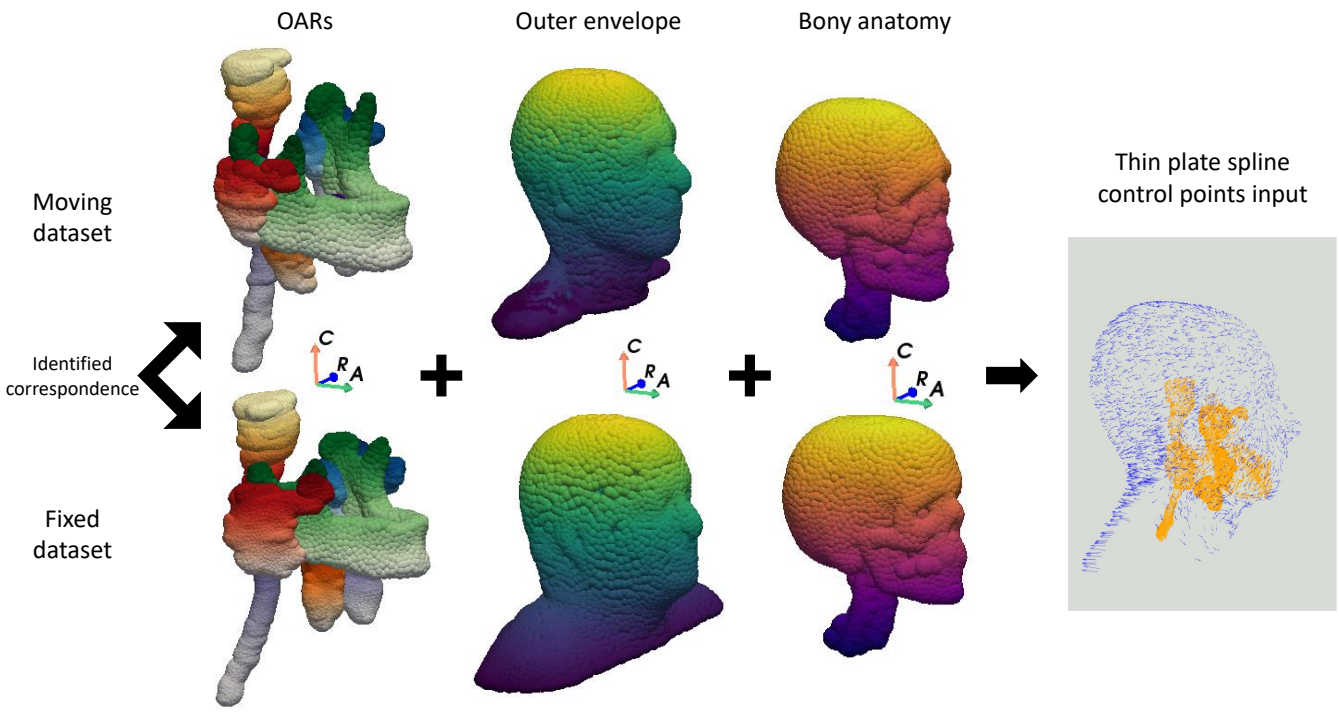


Figure 1: Correspondence of each structure between the fixed and moving datasets, represented by matching colours. Different colour maps are used for each structure and identified correspondences are only shown in the cranio-caudal direction for clarity. The resulting deformation vectors (OARs in orange, bony anatomy and envelope in blue) are the input for the TPS to set up the non-rigid initialisation.

Mesh generation To create triangular meshes of the OAR structures, we used the marching cubes algorithm followed by Taubin smoothing, mesh decimation, optimisation and simplification operations. The full details of this process have been described fully in our previous work [10].
A segmentation of the bony anatomy was acquired automatically from the CT scan by removing the table, thresholding at 400 Hounsfield units (HU), cropping at the base of the neck, and meshing using the marching cubes algorithm. A mesh of the external envelope comprising of the patient's body was generated using a lower threshold level of -200 HU.

Correspondence inference For non-rigid initialisation, dense correspondences were estimated for all segmented structures in the HN, as well as the bony anatomy and outer envelope of each patient (Figure 1).

Thin plate spline setup Once estimated, the correspondences were used as the control points to set up a TPS deformation model [13]. The TPS is generalizable to 3D and has a regularisation parameter, λ_{TPS}, which controls the allowed degree of local deformations.
Control points from each structure were gathered and a single TPS was created for each registration pair. In total, the TPS deformations were performed using about 14,000 control points (range: 10,500 - 16,000). Regions that were identified as null correspondences were excluded from the control point set driving the TPS. In this way, we prevent the creation of anatomically infeasible scenarios, e.g. inverting and folding around tissue mismatches.

For each fixed-moving image pair, we generated a displacement vector field by evaluating the TPS on the grid positions of the moving image. Then, we resampled the fixed image by inverse warping using the *DisplacementFieldTransform* from the *SimpleITK* package. We refer to these as non-rigidly initialised scans. A second NRR algorithm was then used to fine-tune the registration using the image intensities for the whole anatomy.

2.1.2 Non-rigid registration

To demonstrate the added value of CorrTPS, we compared the performance of two established intensity-based NRR methods with and without our proposed initialisation.

Deep learning-based registration (Voxelmorph) We trained multiple Voxelmorph models from scratch [12]. Image volumes were automatically cropped to dimensions of $128 \times 288 \times 480$ such that the whole HN anatomy was within the field-of-view while retaining the original voxel spacing. Additionally, the images were normalised from Hounsfield units onto the range [0,1] using contrast windowing with wide window settings (W 1600HU, L 0HU). The weight of the smoothing loss term which regularises the predicted deformations was set to 0.005.
When performing the five-fold cross-validation, each fold contains 7 testing patients. This leads to a 7×6 testing pairs for each fold (trivial self-registration cases excluded). The same Voxelmorph models were applied to the non-rigidly initialised scans.

Table 1: Mean execution time for all registration pipelines. These times include all necessary pre-processing steps for each method such as rigid pre-alignment, cropping, normalisation, and execution of the second non-rigid registration. All pipelines required rigid pre-alignment. The average time added by CorrTPS is shown in brackets and separated by the contributions of mesh generation and optimisation, correspondence estimation, TPS fitting and resampling respectively.

	Rigid pre-alignment (s)	Pre-processing steps and non-rigid registration (s)	Overall time to register image pair (s)
Voxelmorph	1.6	0.3	1.9
Voxelmorph + CorrTPS		(1.2 + 0.1 + 1.4 + 0.2) + 0.3	4.8
NiftyReg		70.7	72.3
NiftyReg + CorrTPS		(1.2 + 0.1 + 1.4 + 0.2) + 70.7	75.2

Iterative registration (NiftyReg) NiftyReg was first used to non-rigidly register the CT scans after rigid pre-alignment alone [11]. Following an empirical parameter evaluation, the configuration settled upon was a GPU-accelerated cubic B-spline using normalised mutual information loss with a bending energy of 0.001. The same configuration of NiftyReg was then used to register the non-rigidly initialised CT scans.

2.2 Evaluation

For every pipeline, we report the execution time including all the necessary pre-processing steps and the results image- and segmentation-based metrics.
The mean distance-to-agreement (mDTA) was calculated for each of the structures included within CorrTPS as well as the brain, cochleae, optic nerves, orbits following non-rigid alignment of the moving anatomy to the fixed reference patient. Wilcoxon signed-rank hypothesis tests were used to compare each registration algorithm with and without the CorrTPS initialisation. A Bonferroni correction was applied to minimise the risk of type I errors from multiple hypothesis tests (corrected p value: 0.0045).

Implementation details Model training and evaluation was performed on a workstation with an AMD Ryzen 9 3950X 16-Core Processor, 64GB of RAM and an NVidia GeForce RTX 4090 GPU.

3 Results

The average runtime for each approach is shown in Table 1. Figure 2a shows boxplots of the mDTA results for the organ structures included in the CorrTPS, while 2b shows similar results for structures not included in CorrTPS. Significant improvements in registration performance were observed for most structures included in CorrTPS across both methods when using the proposed initialisation. For structures not included in CorrTPS, some improvements were observed for the learning-based registration method (the brain, orbits and lacrimal glands), however, the performance of the traditional registration algorithm was degraded.

4 Discussion

We proposed a novel method to non-rigidly pre-register an image pair prior to inter-subject registration of head and neck CT scans using correspondence of anatomical structures. The initialisation method is based on a geometric learning-based correspondence model which, when coupled with a TPS deformation, allowed the moving scans to be deformed to align segmented OARs before a final registration using established NRR methods. Our initialisation approach showed promise to improve the registration performance of the learning-based NRR method, Voxelmorph, improving the quality of the registrations in terms of the segmentation-based evaluation measure. The performance of our proposed method with the traditional iterative algorithm, NiftyReg, was less clearly defined, with some improvements for structures included in the initialisation and degradations for those not included. The use of CorrTPS required a limited increase in execution time. In terms of the mDTA, the benefit of our proposed initialisation method is evident for structures included within CorrTPS. The one exception to this was for the spinal cord when using NiftyReg, in which the use of the initialisation method harmed registration performance.

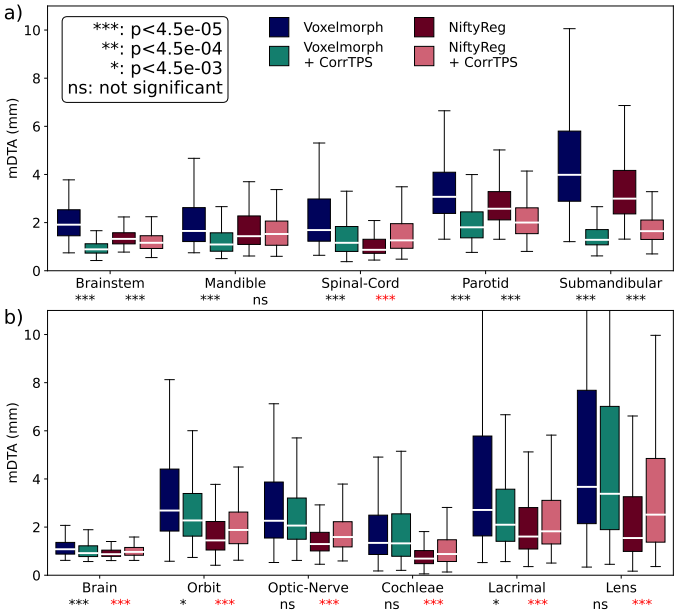


Figure 2: Mean distance-to-agreement results for each registration pipeline. The structures in a) were included in CorrTPS, i.e., their surface correspondence was used to set up the TPS, whereas the structures in b) were not. Significant deviations between methods with and without CorrTPS, as identified with a Wilcoxon signed-rank test, are indicated. Black asterisks show an improvement using CorrTPS, whereas the red show performance reductions.

Little improvement was observed when using our method for the structures not included in CorrTPS, which were predominantly in the superior region of the head. This is likely due to the TPS deformation in this region comprising of an interpolation between the envelope and bony anatomy and the controlling structures which were mainly further inferior in the head and neck region. This shows a sensitivity of the proposed method to the spatial distribution of controlling structures. For registrations with the iterative algorithm, the use of the proposed method degraded the performance for all structures not included in CorrTPS.

While we included some of the structures in this study and excluded others from CorrTPS, the proposed method is applicable to any of the OAR structures segmented for this study, i.e., the complete set of structures could also be included in CorrTPS, thereby fitting a wider distribution of control points within the head and neck anatomy. The implementation of TPS used is fast, estimating a deformation within a second for $\sim 14,000$ control points.

We deployed our proposed initialisation method with an intensity-based traditional NRR algorithm and a DL-based method. In future it would be useful to compare the performance of our hybrid-inspired registration framework against a hybrid registration methods such as ANACONDA [3] to further assess the efficacy of the proposed method.

In this study we focused solely on performing inter-patient registration, despite the task of intra-patient registration being more common within the radiotherapy treatment pathway. Inter-patient registration is important for tasks such as outcome modelling and can be challenging due to the increased variability in anatomy between subjects. We hypothesize that our initialisation method would be particularly effective in intra-patient cases such as in the re-irradiation setting, where large and complex anatomical changes are common, e.g. resulting from surgical interventions, which hinder the accuracy of currently available image registration algorithms [2, 14].

The segmentations used in this study were manually performed. Future introduction of auto-segmentation could mitigate the impact of inter- and intra-observer variation and improve correspondence reliability. Additionally, the use of auto-segmentations with the proposed framework would facilitate the implementation of a fully automated anatomically-informed registration pipeline.

5 Conclusion

The proposed method to initialise an image pair prior to NRR using anatomical correspondence shows particular promise to enhance the registration performance of DL-based registration methods to more closely match that of a traditional iterative NRR algorithm, whilst maintaining a substantial speed advantage. With further development and validation, the proposed approach could additionally raise the performance of traditional iterative NRR algorithms.

References

[1] K. K. Brock, S. Mutic, T. R. McNutt, et al. "Use of image registration and fusion algorithms and techniques in radiotherapy: Report of the AAPM Radiation Therapy Committee Task Group No. 132". *Medical Physics* 44.7 (May 2017), e43–e76. DOI: 10.1002/mp.12256.

[2] M. Murr, K. K. Brock, M. Fusella, et al. "Applicability and usage of dose mapping/accumulation in radiotherapy". *Radiotherapy and Oncology* 182 (May 2023), p. 109527. DOI: 10.1016/j.radonc.2023.109527.

[3] O. Weistrand and S. Svensson. "The ANACONDA algorithm for deformable image registration in radiotherapy". *Medical Physics* 42.1 (Jan. 2015), pp. 40–53. DOI: 10.1118/1.4894702.

[4] J. Kim, S. Kumar, C. Liu, et al. "A novel approach for establishing benchmark CBCT/CT deformable image registrations in prostate cancer radiotherapy". *Physics in Medicine & Biology* 58.22 (Oct. 2013). DOI: 10.1088/0031-9155/58/22/8077.

[5] J. Zou, B. Gao, Y. Song, et al. "A review of deep learning-based deformable medical image registration". *Frontiers in Oncology* 12 (Dec. 2022). DOI: 10.3389/fonc.2022.1047215.

[6] J. Teuwen, Z. A. Gouw, and J.-J. Sonke. "Artificial Intelligence for Image Registration in Radiation Oncology". *Seminars in Radiation Oncology* 32.4 (Oct. 2022), pp. 330–342. DOI: 10.1016/j.semradonc.2022.06.003.

[7] M. P. Heinrich and L. Hansen. "Voxelmorph++". *Biomedical Image Registration*. Springer International Publishing, 2022, pp. 85–95. DOI: 10.1007/978-3-031-11203-4_10.

[8] R. Bhattacharjee, F. Heitz, V. Noblet, et al. "Evaluation of a Learning-based Deformable Registration Method on Abdominal CT Images". *IRBM* 42.2 (Apr. 2021), pp. 94–105. DOI: 10.1016/j.irbm.2020.04.002.

[9] S. Nikolov, S. Blackwell, R. Mendes, et al. "Deep learning to achieve clinically applicable segmentation of head and neck anatomy for radiotherapy". *ArXiv e-prints* (2018). DOI: 10.48550/arXiv.1809.04430.

[10] E. G. A. Henderson, M. van Herk, A. F. Green, et al. "Unsupervised Correspondence with Combined Geometric Learning and Imaging for Radiotherapy Applications". *Shape in Medical Imaging*. Springer Nature Switzerland, 2023, pp. 75–89. DOI: 10.1007/978-3-031-46914-5_7.

[11] M. Modat, G. R. Ridgway, Z. A. Taylor, et al. "Fast free-form deformation using graphics processing units". *Computer Methods and Programs in Biomedicine* 98.3 (June 2010), pp. 278–284. DOI: 10.1016/j.cmpb.2009.09.002.

[12] G. Balakrishnan, A. Zhao, M. R. Sabuncu, et al. "VoxelMorph: A Learning Framework for Deformable Medical Image Registration". *IEEE Transactions on Medical Imaging* 38.8 (Aug. 2019), pp. 1788–1800. DOI: 10.1109/tmi.2019.2897538.

[13] F. Bookstein. "Principal warps: thin-plate splines and the decomposition of deformations". *IEEE Transactions on Pattern Analysis and Machine Intelligence* 11.6 (June 1989), pp. 567–585. DOI: 10.1109/34.24792.

[14] L. Nenoff, F. Amstutz, M. Murr, et al. "Review and recommendations on deformable image registration uncertainties for radiotherapy applications". *Physics in Medicine & Biology* (Nov. 2023). DOI: 10.1088/1361-6560/ad0d8a.

DiffuseRT: learning likely deformations of head and neck radiotherapy patients from data

Rivetti L.*[1], Smolders A.*[2], Vatterodt N.[3], Korreman S.[3], Lomax A.[2], Sharma M.[4], Studen A.[1], Weber DC.[2], Albetini F. [2], and Jeraj R.[1]

[1]Faculty of Mathematics and Physics, University of Ljubljana, Ljubljana, Slovenia
[2]Paul Scherrer Institute, Center for Proton Therapy, Switzerland
[3]Danish Center for Particle Therapy, Aarhus University Hospital, Aarhus, Denmark
[4]Department of Radiation Oncology, University of California, San Francisco, United States of America

Abstract Predicting potential deformations of patients undergoing radiotherapy can improve treatment planning for both adaptive and conventional radiotherapy.
Denoising diffusion probabilistic models (DDPMs) were developed to generate fraction-specific likely anatomical changes. Three distinct DDPM models were developed. The image model was trained to directly generate likely future CBCTs. The DVF model was trained to generate deformable vector fields (DVFs) that deform a reference CBCT. The hybrid model also generates DVFs with the advantage that it does not rely on an external deformable image registration algorithm. Visual inspection showed that the generated images exhibit random daily anatomical changes and systematic weight loss over fractions. Moreover, for the image and hybrid models, the distributions of the generated volume and position changes of the body, esophagus, and parotids followed a similar distribution to the real CBCTs. Using the generated images for robust optimization of simplified treatment plans showed increased target coverage for equal integral dose compared to treatment plans optimized with only set-up robustness.
In conclusion, we developed and evaluated three DDPMs that generate likely anatomical changes in head and neck radiotherapy patients. The results show that the models generate distributions similar to the real anatomical changes on a population level and have the potential to be used for robust anatomical optimization.

1 Introduction

Medical imaging plays a crucial role in monitoring patients before, during, and post-treatment. However, imaging is time-intensive for the medical staff and patients and often uses ionizing radiation, increasing the risk of radiation-induced cancers [1]. Image-guided radiotherapy is particularly imaging-intensive [1], exposing patients to daily volumetric imaging for positioning or treatment adaptation. Therefore, the ability to predict or generate plausible time-series of images without actual acquisition presents a valuable advancement.
Predictive modeling of anatomical changes can be helpful in retrospectively completing datasets missing some images. Using these to simulate daily dose delivery can aid in comparing treatment modalities [2]. Prospectively, such models can be used for adaptive radiotherapy with a plan library [3]. Alternatively, prospective image generation can contribute to robust anatomical optimization [4], where alternative delivery scenarios are incorporated into the optimization to ensure target coverage even when the patient deforms.
Several anatomical deformation models have been proposed in the literature [2, 5], mostly using principal component analysis (PCA) on the deformable vector fields (DVF) generated with deformable image registration (DIR) over a population of patients. However, these models depend on the DIR quality, the choice of reference anatomy and are mostly not patient-specific [2, 5].
Recently, deep-learning-based image synthesis has achieved state-of-the-art performance for various applications [6]. One recent work used deep learning to predict anatomical changes in radiotherapy [7]. As in PCA, it uses a lower dimensional distribution of anatomical changes, which is sampled to generate new deformations. Contrary to PCA, it uses neural networks to learn this distribution and to decode a low-dimensional sample, yielding better results. In natural image synthesis, however, a new type of model has gained much traction: denoising diffusion probabilistic models (DDPMs) [6, 8]. These models learn to predict likely images based on some conditional information. Additionally, such models are inherently probabilistic and can predict many different images based on the same prompt.
In this work, we aim to predict likely interfractional anatomical changes using DDPM models for patients with head and neck cancer. Based on a single, patient-specific reference CBCT and the treatment dose, the models aim to predict several fraction-specific likely anatomical deformations.

2 Materials and Methods

2.1 Data

Our dataset contained 14 patients with head and neck cancer. Each patient had one planning CT with manually annotated clinical target volume (CTV) and organs at risk (OARs), as well as the planned dose distribution and daily cone beam CTs (CBCT) ($\sim$ 30). All daily CBCTs had Raystation ANACONDA contour propagated structures. For 5 patients the structures were reviewed and corrected by an experienced radiation oncologist. These patients were used for validation, and the other 9 for training.

2.2 Denoising Diffusion Probabilistic Model

DDPMs are a category of generative models designed to approximate a given data distribution by inverting a gradual,

*These authors contributed equally to this work

multi-step noise addition process [8]. DDPMs are trained using a *forward process*, in which Gaussian noise is incrementally added to a known image x_0, until the image is transformed into pure Gaussian noise $x_T \sim N(\mathbf{0},\mathrm{I})$. This is followed by the *reverse process*, where a neural network (NN) is trained to predict the parameters of a normal distribution, $p_\theta(\mu(x_t,t),\Sigma(x_t,t))$ that gradually denoises such noisy images back to the initial image x_0. After training, new images can be generated by starting the *reverse process* with pure Gaussian noise. This process can be conditioned on specific information, e.g. a text prompt, to generate samples belonging to a subgroup of the training data [8].
Our DDPM implementation extended the *improved DDPM* [8] from 2D to 3D images. Two approaches were explored to calculate $\mu_\theta(x_t,t)$. In the first approach, the U-Net aims to directly predict the initial image x_0 from the noisy input x_t, and $\mu_\theta(x_t,t)$ is calculated with:

$$\mu_\theta(x_t,t) = \frac{\sqrt{\alpha_t}(1-\overline{\alpha}_{t-1})}{1-\overline{\alpha}_t}x_t + \frac{\sqrt{\overline{\alpha}_{t-1}}\beta_t}{1-\overline{\alpha}_t}x_{0,\theta}(x_t,t) \quad (1)$$

where $\overline{\alpha}_t$ and β_t are predefined parameters of the diffusion process. In the second approach, the network instead predicts the noise ε_θ at step t, and calculates $\mu_\theta(x_t,t)$ as [8]:

$$\mu_\theta(x_t,t) = \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{\beta_t}{\sqrt{1-\overline{\alpha}_t}}\varepsilon_\theta(x_t,t)). \quad (2)$$

To generate patient- and fraction-specific images, the first fraction CBCT_1 and the treatment dose were concatenated to the noisy x_t as conditional information before input in the 3D U-Net (Fig. 1). Since the anatomical changes in our dataset were captured on the daily CBCTs, we trained the models to generate CBCTs based on the CBCT_1, instead of the planning CT. The diffusion step t and the fraction number j were further included by adding their transformer sinusoidal position embedding into each residual block.
We developed three different DDPMs. The **Image Model** directly generates the j fraction specific CBCT_j. It was trained using the CBCT_j as x_0, and $\mu_\theta(x_t,t)$ was calculated using Eq. 2. The **DVF Model** generates DVFs that warp CBCT_1 to the new likely anatomy. It was trained using the DVF_j created with `plastimatch` as x_0, and $\mu_\theta(x_t,t)$ was calculated using Eq. 1. The **Hybrid Model** generates DVFs that warp CBCT_1 to the new likely anatomy. It was trained using the CBCT_j as x_0 but the network generates a $\mathrm{DVF}_{j,t}$ that warps the CBCT_1 to x_0 during training. $\mu_\theta(x_t,t)$ was calculated using Eq. 1 and the model included an additional bending energy penalty to smooth the generated DVF.

2.3 Evaluation

For each model, we generated 10 images for every 3rd fraction, yielding $\sim$ 100 CBCTs per patient. For the DVF and hybrid model, the generated DVF was also used to propagate the planning structures. For the image model, the generated

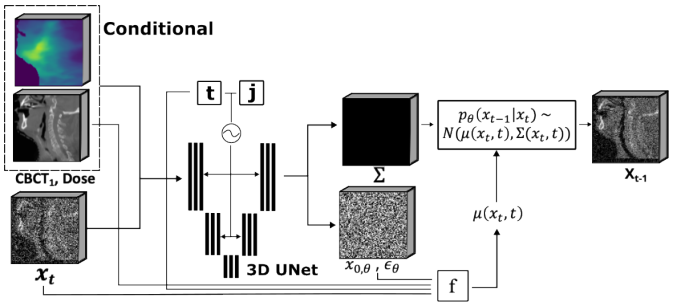


Figure 1: At each diffusion step t of the denoising process, the noisy image x_t is concatenated with the conditional CBCT_1 dose as input to a 3D U-Net. The sinusoidal position embedding [9] of the fraction number j and the step t are further included in each residual block. The network predicts Σ, and ε_θ or $x_{0,\theta}$ depending on the DDPM selected. For the image and DVF model, the function f calculates μ using x_t and t and Eq. 2 or 1. For the Hybrid model, f warps the CBCT_1 with the generated DVF to calculate x_0 and then calculates μ with Eq. 1. Finally, x_{t-1} is sampled from the normal probability density $p_\theta(\mu(x_t,t),\Sigma(x_t,t))$.

samples were registered to CBCT_1 with `plastimatch` and the resulting DVFs were used to propagate the contours.

The generated images were first qualitatively assessed with visual inspection to intuitively judge the image quality and interpret the main modes of deformation. The models were further quantitatively evaluated by comparing the changes in volume and geometrical center (GC) of the body, parotids and esophagus of the generated samples with the changes obtained from the real daily CBCTs. The real and generated distributions were compared with the Wasserstein distance (WD). A lower WD indicates more similar distributions.

As a proof of concept, we tested the suitability of the generated images for robust optimization in a simplified planning study. Since the model was trained with CBCT images, all images, both real and generated, were first converted to *corrected* CBCTs in Raystation [10]. Proton therapy plans with 2 anterior fields with $\pm$ 65° gantry angles were optimized on the *corrected* CBCT_1 to deliver 60 Gy to the CTV. No organs at risk (OARs) were considered, the only additional objective aimed at reducing the integral dose to the body. For all patients, the treatment plans were first optimized using worst-case robust optimization with 1, 3, and 5 mm set-up uncertainty (*standard robust plans*). Secondly, the treatment plans were optimized without set-up uncertainty but including 1, 2 or 3 generated images for every third fraction (*anatomical robust plans*). Lastly, we combined 1, 3 or 5 mm set-up uncertainty with 1 sample image for every third fraction (*combined robust plans*). The objective weights were kept the same for each patient and each different robustness setting. All plans were subsequently recalculated on the *corrected* daily CBCT_2 to CBCT_n. The daily doses were evaluated by calculating the integral dose and the percentage of the volume of the CTV that received 95% of the prescribed dose (V95).

3 Results

The images generated by all models looked realistic based on the author's visual inspection. Furthermore, for all three models, both systematic anatomical changes (e.g. weight loss) and random changes (e.g. changes in the oral cavity and shoulders positioning) were observed (Fig. 2). More detailed inspection indicated that changes in the initial fractions are mostly random anatomical changes, especially the positioning of the shoulders with respect to the neck. In later fractions, all models generated images showing weight loss.

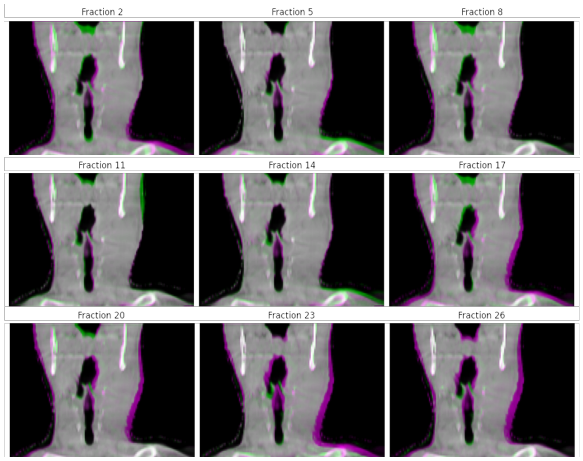


Figure 2: Coronal view of randomly generated image samples with the hybrid model for one patient of the validation dataset. The initial image is in pink, the sample is overlayed in green. The generated samples exhibit random changes and weight loss towards the later fractions.

The population-wise distribution of the changes in volume and GC position of the body generated by the image and hybrid models were overall similar to the real body changes (Fig. 3). However, for one specific patient, the real change of volume was much larger than predicted (red arrow of Fig. 3a). It was verified with the clinical personnel that the anatomical change in this patient was indeed highly exceptional and, therefore, outside the range of changes the model can predict. The distributions generated using the DVF model do not fit well to the real body changes.

The higher similarity in body change distributions for the image and hybrid models than for the DVF model is also confirmed quantitatively by a lower WD (Tab. 1). Furthermore, for the esophagus, the hybrid model performed better than the other models in both ΔVolume and ΔGC. For both parotids, the distribution of the metrics is best predicted with the image model. However, the differences in WD between the different models in the parotids are much less pronounced than for the body and esophagus.

For all the models, the anatomically robust plans improved the worse-case target coverage compared to the standard robust plans for the same integral dose (Fig. 4). Additionally, the hybrid model's anatomically robust plans also improve the average target coverage compared to the standard robust plan (Fig. 4 c). These enhancements were less prominent for the image model and negligible for the DVF model.

The anatomically robust plans are not substantially affected by including more samples per fraction in the optimization, which implies that their robustness cannot be increased this way. Robustness can, however, be increased by combining the samples with set-up uncertainty (combined robustness), which also shows improvements over only using set-up uncertainty for the image and hybrid models (Fig. 4).

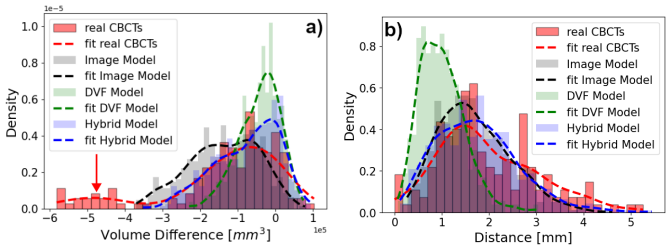


Figure 3: Population-wise distribution of changes of the body between $CBCT_1$ and $CBCT_j$ for the real changes (red) and those generated with the image (black), DVF (green) and hybrid (blue) models. a) Changes in the body volume. The red arrow points toward data points for a patient who had uncommon weight loss. b) Shift of the geometrical center (GC) of the body.

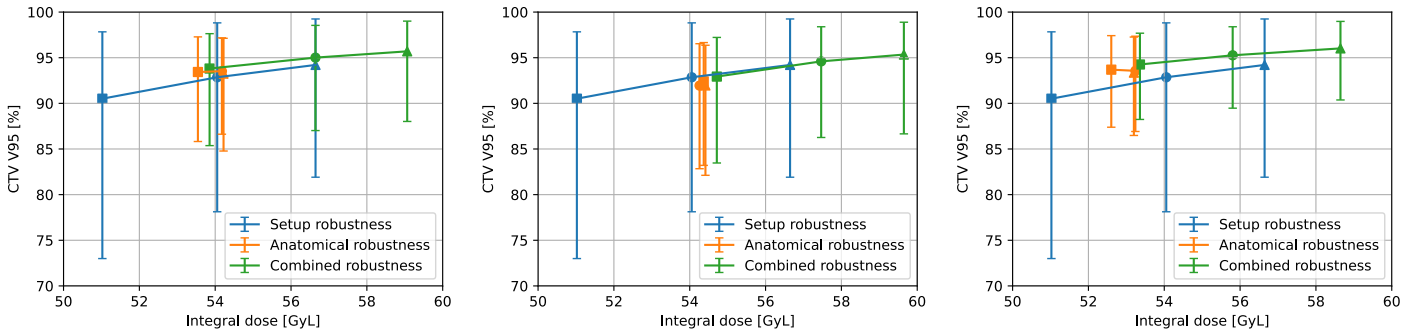
Table 1: WD between the real and predicted distributions of anatomical change over all 5 patients. ΔVolume is the change in volume of a structure between $CBCT_1$ and the subsequent images $CBCT_j$ (for the same patient). ΔGC is the distance between the geometrical center of a structure of the $CBCT_1$ and the same structure in a subsequent fraction $CBCT_j$. The best metric (lowest value) is indicated in bold.

Metric	Structure	Image Model	DVF model	Hybrid model
ΔVolume	Body	**0.21**	0.56	0.22
	Esophagus	0.40	0.44	**0.24**
	Right Parotid	**0.20**	0.28	0.23
	Left Parotid	**0.25**	0.28	0.33
ΔGC	Body	**0.24**	0.56	**0.24**
	Esophagus	0.46	0.33	**0.28**
	Right Parotid	**0.19**	0.26	0.27
	Left Parotid	**0.13**	0.17	0.19

4 Discussion

In this work, we developed three DDPMs that can generate likely anatomical changes in head and neck radiotherapy patients based only on the fraction number, one reference image, and the treatment dose distribution. Unlike other models reported in the literature [2, 7], our models generate fraction-specific changes.

The visual analysis of the generated images indicates that all models produce plausible images with systematic weight loss and random anatomical changes in the oral cavity and positioning of the patient. These results were in line with the observed changes in the real data. The image and hybrid models outperform the DVF model in terms of segmentation



(a) Image model **(b)** DVF model **(c)** Hybrid model

Figure 4: Target coverage (V95%) versus integral dose for all fractions of the evaluation patients. The markers indicate the mean, and the bounds the 5% and 95% quantiles from evaluating the dose in all the fractions. The set-up robustness includes 1 (▪),3 (●) and 5 (▲) mm set-up uncertainty (with increasing integral dose in that order). Anatomical robustness was optimized with 1 (▪), 2 (●) and 3 (▲) samples for every third fraction. Combined robustness used 1 sample for every third fraction and 1 (▪),3 (●) or 5 (▲) mm set-up uncertainty.

distribution analysis, especially regarding the volume of the body mask. Furthermore, they are more suitable for robust optimization.

Each model has unique advantages and limitations. The hybrid model does not rely on an external DIR algorithm. Furthermore, it generates a DVF that warps the reference image to the new likely anatomy, allowing to also propagate the reference structures. However, using a DVF constrains the deformations that can be predicted. Some changes, like nasal cavities or rectum fillings, cannot be modeled well with a DVF. This limitation does not impact our results because our CBCTs field of view doesn't include nasal cavities. However, it may become more crucial for images involving the abdomen or sinonasal area. The DVF model's advantage lies in its ability to be trained on a multimodal image dataset, as it only requires the DVFs that can be generated with multimodal DIR.

All models showcased the ability to enhance the robustness of the simplified treatment plan. The improvements seem to correlate with the models' proficiency in predicting body volume changes, as the hybrid and image models exhibited more enhancements compared to the DVF model. However, our simplified treatment plans did not consider OARs, and the objective functions and weights were not adjusted to obtain an optimal plan per robustness setting. For this reason, the conclusions highlighted before cannot be translated directly to clinical cases.

The models were trained to generate CBCT images based on the initial CBCT of the treatment. These images require correction before beeing suitable for dose calculation. However, since the initial CBCT is not accessible during planning at clinics, the models cannot be used for planning in clinical settings. Ideally, the models should produce daily CT images derived from the planning image, but such data was unavailable for this study. Nevertheless, the approach is analogous, and if the data becomes available, the models could be readily retrained. We anticipate that the outcomes will be comparable.

References

[1] L. A. Dawson and M. B. Sharpe. "Image-guided radiotherapy: rationale, benefits, and limitations". *The Lancet Oncology* 7.10 (2006), pp. 848–858. DOI: https://doi.org/10.1016/S1470-2045(06)70904-4.

[2] S. Thörnqvist, L. B. Hysing, A. G. Zolnay, et al. "Treatment simulations with a statistical deformable motion model to evaluate margins for multiple targets in radiotherapy for high-risk prostate cancer". *Radiotherapy and Oncology* 109.3 (2013), pp. 344–349. DOI: https://doi.org/10.1016/j.radonc.2013.09.012.

[3] B. Rigaud, A. Simon, M. Gobeli, et al. "Statistical Shape Model to Generate a Planning Library for Cervical Adaptive Radiotherapy". *IEEE Transactions on Medical Imaging* 38.2 (2019), pp. 406–416. DOI: 10.1109/TMI.2018.2865547.

[4] J. Unkelbach, M. Alber, M. Bangert, et al. "Robust radiotherapy planning". *Physics in Medicine & Biology* 63.22 (2018), 22TR02. DOI: 10.1088/1361-6560/aae659.

[5] M. A. Chetvertkov, F. Siddiqui, J. Kim, et al. "Use of regularized principal component analysis to model anatomical changes during head and neck radiation therapy for treatment adaptation and response assessment". *Medical Physics* 43.10 (Oct. 2016), pp. 5307–5319. DOI: 10.1118/1.4961746.

[6] P. Dhariwal and A. Nichol. "Diffusion Models Beat GANs on Image Synthesis". *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato, A. Beygelzimer, Y. Dauphin, et al. Vol. 34. Curran Associates, Inc., 2021, pp. 8780–8794.

[7] O. Pastor-Serrano, S. Habraken, M. Hoogeman, et al. "A probabilistic deep learning model of inter-fraction anatomical variations in radiotherapy". *Physics in Medicine & Biology* 68.8 (2023), p. 085018. DOI: 10.1088/1361-6560/acc71d.

[8] A. Q. Nichol and P. Dhariwal. "Improved denoising diffusion probabilistic models". *International Conference on Machine Learning*. PMLR. 2021, pp. 8162–8171.

[9] A. Vaswani, N. Shazeer, N. Parmar, et al. "Attention is all you need". *Advances in neural information processing systems* 30 (2017).

[10] K. Reiners, R. Dagan, A. Holtzman, et al. "CBCT-Based Dose Monitoring and Adaptive Planning Triggers in Head and Neck PBS Proton Therapy". *Cancers* 15.15 (2023). DOI: 10.3390/cancers15153881.

Choice of Colourmaps for Radiotherapy

Alex Davies[1], Mark Fleckney[1], and Marcel van Herk[2,]

[1]Department of Radiotherapy Physics, Kent Oncology Centre, Maidstone, United Kingdom
[2]Radiotherapy Related Research, University of Manchester, Manchester, United Kingdom

Abstract Colourmaps are used in medical imaging to allow the simultaneous display of multiple datasets. A large observer study was conducted to investigate whether choice of colourmap should be a consideration in radiotherapy. A simulated low-contrast phantom was displayed in eight colourmaps, and users asked which test regions were visible using each colourmap. Analysis of 214 observers showed that more test regions were visible in the rainbow, red, viridis, and orange colourmaps compared to grayscale, whilst fewer regions could be seen using blue, green, and purple. A k-means clustering analysis showed 6 distinct clusters of response type, indicating user-specific benefits to different colourmaps for this detection-type task.

1 Introduction

Over the past forty years, advances in medical imaging, understanding of oncology, and therapy techniques have led to significant improvements in cancer outcomes. Frequently, multiple patient datasets are utilised to aid with target and organ at risk delineation prior to radiotherapy planning. A number of studies have investigated the use of alternative colourmaps when viewing patient datasets to aid in diagnosis[1,2] and surgical planning[3]. In radiotherapy, colourmaps are used in image fusion and image guidance.

Within the data science community, there is an increasing awareness of the importance of choosing an appropriate colourmap to avoid misinterpretation of data. Crameri et al.[4] describe that the non-linearity of the eye can lead to visual errors >7% when observers review data presented in the "jet/rainbow" colourmaps. Additionally, such colourmaps are often poorly interpreted by viewers with Colour Vision Deficiencies (CVD). Crameri et al.[4] suggest that perceptually linear colourmaps should be used, where incremental changes in data values are displayed with incremental changes in colourmap lightness.

Despite an increase in the use of AI, auto-contouring, and auto-matching algorithms in radiotherapy in recent years, review of image segmentation and registration by human observers remains essential[5]. A literature review into choice of colourmaps found a small number of articles investigating this issue within diagnostic imaging, but did not highlight any previously published works describing the optimal of colourmaps for use within radiotherapy. This research aims to use a large observer study to investigate whether there is a "best" colourmap to review images for radiotherapy purposes, and examine whether the best colour choice is observer-specific.

2 Materials and Methods

To allow a large number of observers to contribute to the investigation, and enable control of the images viewed by participants in the study, a web survey with simulated images was designed. A low-contrast phantom designed to mimic the commercially available CATPhan™ was created in Python (v3.10.9) using the Numpy module (v1.23.5). The pixel values for eight low-contrast test regions, labelled A-H, were selected at random from a pre-defined list of values close to the phantom background. The Cv2 module (v4.8.0) was used to add both structural noise using a Gaussian filter, and image noise using a Gaussian noise array, with the noise distribution selected to visually match real CT scans. Matplotlib (v3.7.0) was used to display and export the phantom image in the PNG format in six single colour colourmaps: blue; green; orange; purple; red; grayscale, and two multi-colour colourmaps: rainbow and viridis (Figure 1).

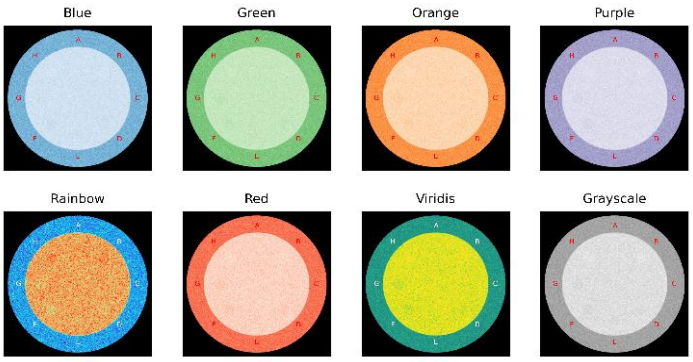


Figure 1: The simulated phantom displayed in the eight colourmaps

The script to generate simulated phantom images was run three times to create three sets of eight images in the chosen colourmaps with varying test region pixel values.

To investigate whether viewing the simulated phantom images in colour offered an improvement in the number of visible test regions, a Qualtrics™ survey was created. This displayed each of the 24 images one at a time to participants, asking them to record which of test regions A-H were visible for each image. The images were displayed in a preset random order for all sets and all respondents. At the end of the questionnaire, participants were invited to answer some additional questions that asked for their current role, and whether they worked with "*medical images*". Finally, respondents were asked "*…to the best of your knowledge do you have any colour vision deficiencies (also known as colour blindness)?*".

The Qualtrics™ survey was shared within the authors' local departments, and posted on the UK Medical Physics mailbase. No restrictions or recommendations to

respondents' backgrounds, the equipment they used to complete the survey, or environmental conditions were made to encourage as many responses as possible.

3 Results

Colourmap perceptional linearity was verified using the Colorspacious module (v1.1.2) by converting the full-colourmap RGB values into the International Commission on Illumination's (CIE) CAM02-UCS colour appearance model, and assessing the lightness component, J. Colourmaps with constantly increasing lightness are thought to have better perceptual linearity for observers[4]. The calculated J values for a series of test images displaying the full range of available colours for each colourmap were compared, and the dynamic range of the grayscale colourmap was reduced by creating a new "partial grayscale" colourmap to match the range of lightness of all colourmaps, as shown in Figure 2.

During creation of the web survey, lightness component (J) values of the simulated phantom test regions were reviewed prior to adding Gaussian noise, and the Matplotlib image rendering settings were adjusted by editing the range of pixel values that the colourmap was applied to. This method ensured that the test regions were of comparable lightness for each colourmap, and that the only variable between the images was the displayed colour.

The Qualtrics™ survey was fully completed by 214 respondents, with a range of backgrounds and experience in medical images.

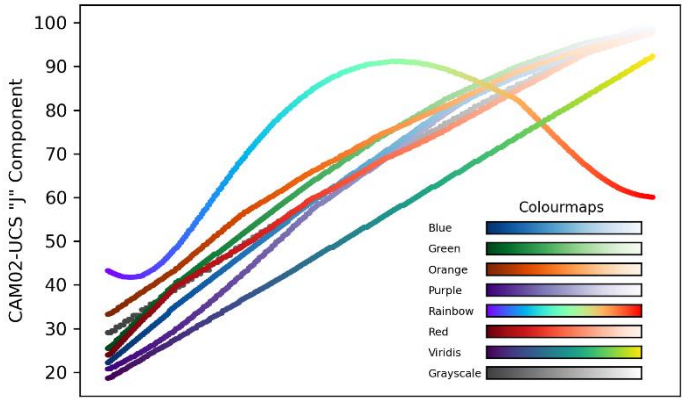


Figure 2: Colourmap lightness "J" across the full-range of each colourmap. Obviously the 'rainbow' colourmap is not perceptually linear, but the other colourmaps are approximately linear. The lightness values for the phantom test regions were found to be 86.6 < J < 92.6, where all colourmaps (excluding rainbow) are approximately linear.

The majority of survey participants were Medical Physicists (64.0%), of which 74.5% were Radiotherapy Physicists, 18% were Diagnostic Imaging Physicists, and 5% were Nuclear Medicine Physicists. Those describing themselves as fitting within the "*Other*" category (15%) included Radiotherapy Engineers, Nurses, Clinical Technologists, and PhD students (Figure 3). The majority of those surveyed (67%) said that they worked with medical images.

When asked if they had a colour vision deficiency, 6% replied "*Yes*", and 86% replied "*No*".

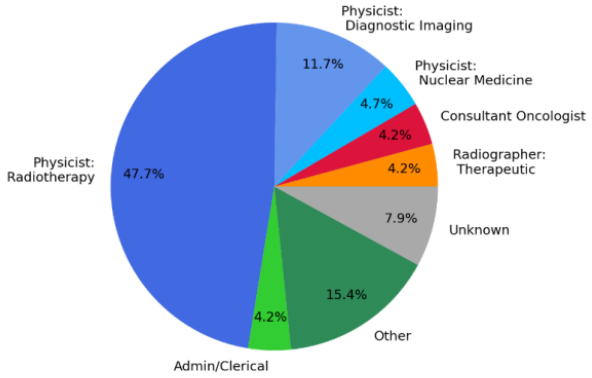


Figure 3: Pie chart showing survey participant roles

To account for differences in viewing conditions, display equipment, and end users' abilities, all analysis was performed relative to participants' own responses rather than comparing across different users. Within each set of eight images, respondents' answers for each colourmap were compared to their responses for grayscale. This showed whether a specific test region was visible when viewed in the colourmap but not in grayscale (scored as a positive improvement), or alternatively visible in grayscale but not the colourmap (scored as a negative improvement). Averaging across all respondents and all colourmaps, the most additional test regions could be seen in the rainbow colourmap, followed by red, viridis, orange, green, purple, and blue (Figure 4).

To generalise the results, and investigate whether colour improvement was more significant for brighter or darker hues, the Contrast to Noise Ratio (CNR) was calculated as the difference in mean values between a given test region and background region, divided by the background standard deviation. When results were plotted to show relative benefit for each colourmap ordered by CNR, it was noted that some colourmaps appeared to offer equal advantage and disadvantage in test region visibility dependent on end user (Figure 4). This suggested subgroups within the cohort where different colourmaps were more or less helpful in visualising the test regions.

The scikit-learn (v1.2.1) k-means clustering algorithm was used to group the respondents into 6 clusters of similar response types, where each cluster contained between 30 – 50 participants. Comparison between the clusters showed a number of features and general trends, summarised below in Table 1.

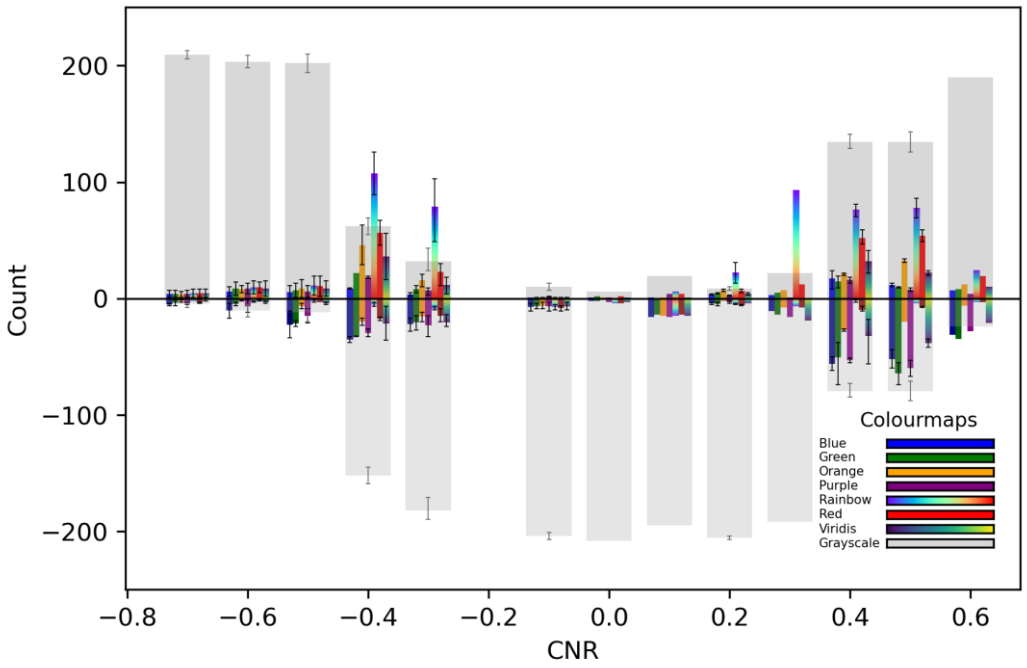


Figure 4: Improvement with colour at varying CNR. Grayscale bars show total number of visible (positive count) and non-visible (negative count) test regions in grayscale images. Positive colour bars show test regions detected by users in the displayed colour but not grayscale. Negative colour bars show colourmap specific regions not seen in colour, but detectable in grayscale. Where multiple CNR values were displayed across the entire survey, counts have been averaged, and the error bars show the range in values.

4 Discussion

This research investigated the potential benefit to using colour in radiotherapy image interpretation through a large observer study. To account for relative differences in viewing conditions, equipment, and observers all analysis was performed relative to the grayscale phantom images.

Whilst the rainbow colourmap was found to be most helpful, followed by red and orange; the blue, green, and purple colourmaps were generally not beneficial. The use of contrast to noise ratios allows these results to be applied to other colour research projects. The cluster analysis performed on the dataset shows that there were differences between participants and potentially trends to responses within user roles.

Previous studies of the benefits of using colourmaps have used small numbers of observers (Whitehead et al.[2] – a single Neuroradiologist, Saba et al.[1] – four radiologists), and reviewed clinical diagnostic images. These studies allowed participants to adjust image window and level on a patient-by-patient basis. A potential limiting factor to this previous clinically-focused research is that the results are a convolution of the inherent patient-specific image complexity and allowed windowing adjustments. In contrast, this research included a far greater number of observers and did not allow changes to windowing. Use of an artificial phantom image ensured that CNR was controlled and reproducible across repeated image sets.

There are however, a number of limitations to this work. The Qualtrics™ survey asked users if the labeled test regions were visible, leading them to only examine a specific area of the image, simplifying the task.

Cluster	Key Feature	Key Participants
0	Most significant learning/experience benefit	
1	Variable benefit in colourmaps across CNR.	30% of non-medical image users[1]. 31% of respondents with CVD[2].
2	Most significant benefit to rainbow. Colour benefit generally at negative CNR	30% of Radiotherapy Physicists[3]. 33% of Consultant Oncologists[4].
3	Consistent benefit to Rainbow & Red colourmaps	56% of Admin/Clerical respondents[5].
4	Cluster with least benefit to colourmaps	44% of Diagnostic Imaging Physicists[6].
5	Only Orange, Rainbow, Red, Viridis colourmaps beneficial	

Table 1: Summary of respondent clustering

Test regions were also reproducibly sized and positioned throughout the experiment. The results therefore do not show true detectability, although these shortcomings were necessary to allow for relative comparison between different images. Variability in test region size and position between images and colourmaps would require the experiment to be repeated in controlled conditions, limiting the number of participants. Despite some limitations, the

[1] Non-medical images users in other clusters: 8-21%
[2] CVD respondents in other clusters: 8-23%
[3] Radiotherapy Physicists in other clusters: 9-19%
[4] Consultant Oncologists in other clusters: 11-22%
[5] Admin/Clerical respondents in other clusters: 0-11%
[6] Diagnostic Imaging Physicists in other clusters: 8-20%

survey images are comparable to real-world use of low contrast phantoms where the same phantom and test region configuration is used for repeated qualitative tests.

Across all respondents, more additional test regions were visible in rainbow than any other colourmap, despite this colourmap lacking perceptual linearity as shown clearly in Figure 2. On review, this improvement is thought to be more due to coincidence rather than an inherent benefit to the rainbow colourmap. Within the survey images, the hues change rapidly from aqua to red/orange at the approximate level of the background noise. This change allows observers to infer the presence of a test region from colour changes in the noise rather than by directly visualising the test region. The rainbow colourmap therefore offers potential benefits in object detection, providing the windowing is optimised to allow change in hues at the correct level. It unlikely this benefit could improve radiotherapy delineation as this task generally involves reviewing a larger range of Hounsfield Units. Viridis is a Matplotlib multi-colour colourmap, designed to be perceptually linear. Images displayed in viridis showed a similar although less pronounced benefit than rainbow. Further research could investigate rainbow and viridis to determine whether colourmap perceptual linearity is a requirement for delineation tasks.

In general, the warmer colourmaps; red and orange, offered more benefit to the respondents, whereas fewer test regions were visible in the cooler colourmaps; blue, green, and purple. With the relative analysis approach adopted within this study, it is not possible to determine whether this trend is due to a human visual processing tendency, lighting conditions in office environments, computer monitors, or even web browser dependent. Repeating the experiment in controlled conditions would be needed to deconvolve these factors. It is hoped however, that most participants completed the survey using the equipment and environment of their routine clinical work, and that the results are therefore representative of the real-world benefits of alternative colourmaps.

Review of the clustering results showed that participants could be grouped into response type. Studies involving human subjects will always show a combination of physiological and psychological factors. This is visible in cluster 0, where the blue, green, and purple colourmaps were beneficial for these participants, but only for the first set of images. This is believed to be a learning/experience artefact, with review of further images these respondents became more confident of their interpretation of grayscale images leading to a reduced benefit to the alternative colourmaps. The majority of participants opted into sharing information about their role and whether they looked at medical images. This allowed for some generalisations about clusters and their members to be made. Cluster 4 consistently showed the least benefit to the alternative colourmaps, and contained the largest proportion of Diagnostic Imaging Physicists. This could indicate that these users felt most confident with grayscale due to their day-to-day role, or were in an environment or using equipment that had been optimised for grayscale images. Similarly, cluster 3, containing the largest proportion of Admin/Clerical participants, and self-reported non-medical Images users, showed a consistent benefit to alternative colourmaps across CNRs. This supports Saba et al.'s[1] finding that less experienced reviewers benefitted more from colourmaps than users with more clinical experience. A number of survey participants reported a CVD. These respondents were split across all clusters, although the majority was found in cluster 1. Although this could be indicative of a trend, the relatively small number of survey participants with a CVD and range of possible types of CVD suggest this might be more coincidence than correlation.

At the end of the survey, participants were asked whether they could recall which colourmaps they found easiest and hardest to interpret. A significant number of users found this question difficult to answer, suggesting that if optimal colourmaps are individual-specific, it is not possible to determine which colourmap is preferred without a feedback process.

5 Conclusion

Use of alternative colourmaps can improve test object detectability across a range of CNRs. Cluster analysis shows different trends within respondents suggesting that the most beneficial colourmap may be specific to an individual, although in general, users perform better with warmer rather than cooler colourmaps. Further research is required to investigate whether these findings are also applicable to target delineation and image fusion.

References

[1] L. Saba, G.M. Argiolas, E. Raz, S. Sannia, J.S. Suri, P. Siotto, et al. Carotid artery dissection on non-contrast CT: does color improve the diagnostic confidence? Eur J Radiol. 2014;83(12):2288-93. DOI: 10.1016/j.ejrad.2014.09.001

[2] M.T. Whitehead, L.M. Guillot, B.K. Reilly. Cochlear signal alterations using pseudo-color perceptual enhancement for patients with sensorineural hearing loss. Pediatr Radiol. 2021;51(8):1448-56. DOI: 10.1007/s00247-021-04987-z

[3] J.C. Silverstein, N.M. Parsad, V. Tsirline. Automatic perceptual color map generation for realistic volume visualization. J Biomed Inform. 2008;41(6):927-35. DOI: 10.1016/j.jbi.2008.02.008

[4] F. Crameri, G.E. Shephard, P.J. Heron. The misuse of colour in science communication. Nat Commun. 2020;11(1):5444. DOI: 10.1038/s41467-020-19160-7

[5] K. Sheng. Artificial intelligence in radiotherapy: a technological review. Front Med. 2020;14(4):431-49. DOI: 10.1007/s11684-020-0761-1

Genetically-based analysis of dose surface maps for assessing toxicity post-RT

Tiziana Rancati[1], Eliana Gioscio[1], Michela Carlotta Massi[2], Nicola Rares Franco[3], Petra Seibold[4], Barbara Avuzzi[5], Alessandro Cicchetti[1], Barry Rosenstein[6], David Azria[7], Ananya Choudhury[8], Dirk De Ruysscher[9], Maarten Lambrecht[10], Elena Sperk[11], Christopher J. Talbot[12], Ana Vega[13], Liv Veldeman[14],Adam Webb[12], Paolo Zunino[3], Anna M. Paganoni[3], Francesca Ieva[3], Andrea Manzoni[3], Sarah L. Kerns[15], Alison Dunning[16], Jenny Chang-Claude[4], Catharine M. L. West[8] and the REQUITE/RADPRECISE consortia

[1]Data Science Unit, Fondazione IRCCS Istituto Nazionale dei Tumori, Milan, Italy.
[2]Health Data Science Centre, Human Technopole, HDSD, Milan, Italy.
[3]MOX, Department of Mathematics, Politecnico di Milano, Milan, Italy.
[4]Division of Cancer Epidemiology, German Cancer Research Center (DKFZ), Heidelberg, Germany.
[5]Unit of Radiation Oncology Fondazione IRCCS Istituto Nazionale dei Tumori, , Milan, Italy.
[6]Dept. of Radiation Oncology and Dept. of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, USA.
[7]Department of Radiation Oncology, Institut du Cancer de Montpellier (ICM), Montpellier, France.
[8]The Christie Hospital NHS Trust, Department of Clinical Oncology, Manchester, United Kingdom.
[9]Department of Radiation Oncology (Maastro), Maastricht University Medical Centre, Maastricht, The Netherlands.
[10]Department of Radiotherapy, University Hospitals Leuven, Leuven, Belgium.
[11]Department of Radiation Oncology, University Medical Centre Mannheim, Mannheim, Germany.
[12]Leicester Cancer Research Centre, University of Leicester, Leicester, United Kingdom.
[13]Instiuto de Investigaciòn Sanitaria, Fundaciòn Pùblica Galega Medicina Xenòmica, Santiago de Compostela, Spain.
[14]Department of Radiation Oncology, Ghent University Hospital, Ghent, Belgium.
[15]Department of Radiation Oncology, Medical College of Wisconsin, Milwaukee, USA.
[16]Centre for Cancer Genetic Epidemiology, University of Cambridge, Cambridge, United Kingdom.

Abstract

Polygenic risk scores (PRS) can consider the cumulative impact of multiple SNPs on radiosensitivity. Epistasis is also likely to affect the toxicity risk, and it is essential to include SNP-SNP interactions in PRS.

Voxel-based analysis (VBA) [1] examines dose distributions for insights into radiobiological heterogeneity across tissues to enhance understanding of the dose-response relationships.

We propose a method to couple PRS and VBA with application to modelling late urinary toxicity after prostate cancer radiotherapy.

We included 778 patients prostate cancer patients receiving radical radiotherapy enrolled and followed up in the REQUITE and RADPrecise studies and considered three late toxicity endpoints: urinary frequency, heamaturia and rectal bleeding.

We incorporated the PRSi within VBA via proportional adjustment of the single voxel doses. On the resulting PRSi-modulated DSMs, we performed Cox VBA and identified the bladder/rectum subregions to be spared within personalised radiotherapy optimisation.

1 Introduction

Polygenic risk scores (PRS) can consider the cumulative impact of multiple SNPs on radiosensitivity. Epistasis is also likely to affect the toxicity risk, and it is essential to include SNP-SNP interactions in PRS.

Voxel-based analysis (VBA) considers full, 3-dimensional dose distributions, offering biological intuitions on the possible heterogeneity of the radiobiology across tissues. There is a need to include single-patient dose modifying factors into VBA to account for the inter-patient variability. Green et al. [1] proposed multivariable modelling at a voxel level; they produced maps of hazard ratios (HRs) for doses and patient-specific risk factors. This method is methodologically and statistically straightforward, yet the variation of HRs at a voxel level has poor biological justification.

We propose a novel approach to include patient-specific factors in VBA, allowing HRs at a patient level. As a first application, we analyse bladder and rectum dose-surface maps (DSMs) incorporating a polygenic risk score with interections (PRSi) [2,3].

2 Materials and Methods

We included prostate cancer patients enrolled in the prospective REQUITE international study. REQUITE recruited in 8 countries in Europe and the US (April 2014-

March 2017). Here, we considered the sub-cohort of patients receiving radical radiotherapy (RT).
Patients were treated according to local regimens, but centres used standardised data collection. Blood samples were collected for DNA extraction and genotyping.
We considered grade≥2 urinary frequency (Grade 0 or Grade 1 before RT), grade≥1 haematuria (Grade 0 before RT) and grade≥2 late rectal bleeding as separated endpoints.

A pool of 43 SNPs associated with late RT toxicity from the literature was tested, and a deep sparse autoencoder method identified the SNPs affecting the toxicity risk at 2-year follow-up [2]. A new method for accounting for SNP-SNP interactions (PRSi) was developed; the PRSi shows which SNPs and alleles are included, whether they increase or decrease the risk of toxicity and their combined effect sizes [3].

The PRSi for frequency, haematuria and rectal bleeding included 13 validated SNPs [2,3].

We generated DSMs using a dedicated Python tool. For the bladder, we cut DSMs anteriorly, aligned maps at the most caudal slice, and normalised their lateral extent to the population's maximum bladder diameter. In the cranial-caudal direction, we fixed the most caudal 1.5 cm (~corresponding to the bladder neck) and normalised the remaining surface to the median bladder height. For the rectum, we cut DSMs posteriorly, aligned maps at the most caudal slice, and normalised their lateral extent to the population's maximum rectum diameter. In the cranial-caudal direction, we fixed the most caudal 3 cm (~anal canal) and normalised the remaining surface to the median rectum height.

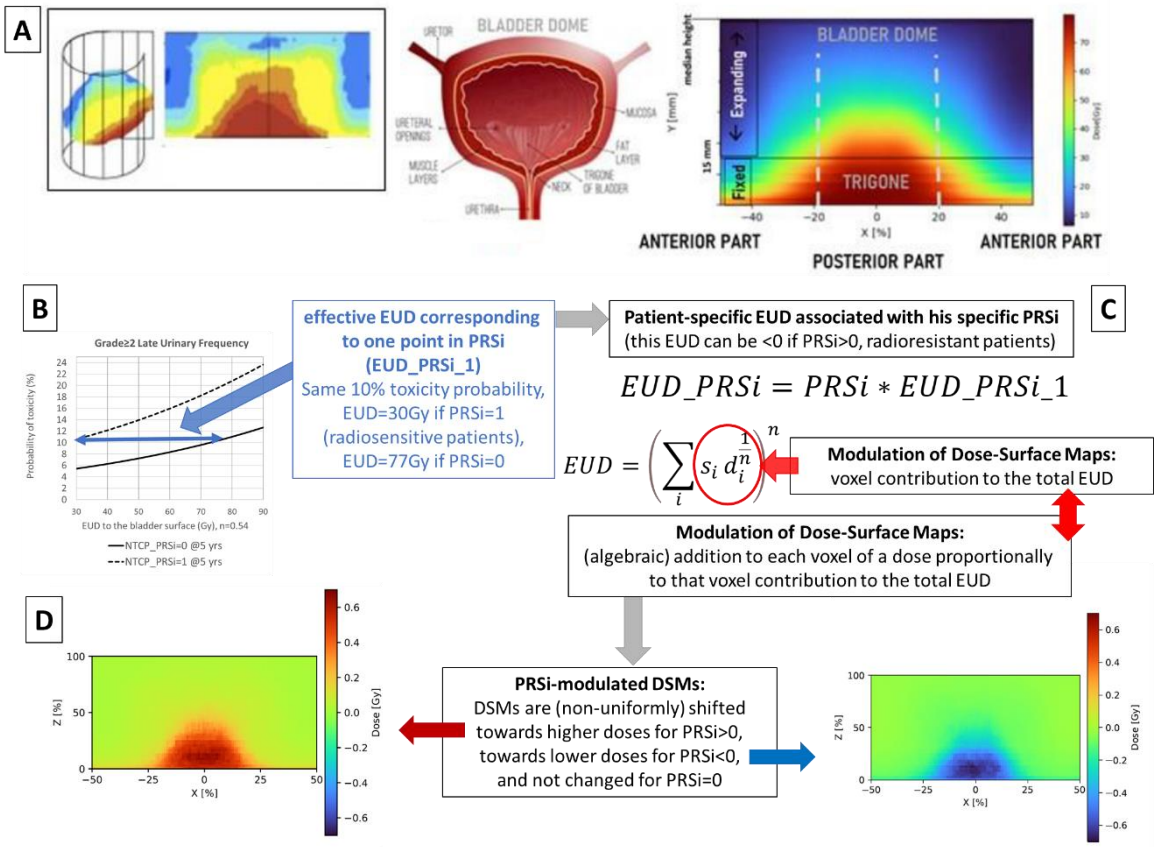


Figure 1: Steps of the approach to generate Dose Surface Map (DSMs) modulated by the Polygenic Risk Score (PRSi). A) Computation, alignment and normalisation of DSMs. B) Determination of the effective Equivalent Uniform Dose (EUD) corresponding to one point increase in PRSi through Cox-NTCP regression. C) Computation of the patient-specific EUD associated with his PRSi and the single voxel's contribution. D) Modulation of DSMs using the distributed single patient EUD_PRSi.

To establish the contribution of the PRSi, we used multivariable Cox-Normal Tissue Complication Probability (NTCP) [4] including dosimetry and PRSi. The Cox-NTCP models based on Equivalent Uniform Doses (EUDs) corrected for fractionation (alpha/beta=3Gy, gamma=0.7Gy/day, fixed) and derived from bladder and rectum dose-surface-histograms. The volume parameter n was 0.5 for the bladder (fitted) and 0.10 for the rectum (from QUANTEC).

We used the HRs for the PRSi to compute effective EUDs [5] corresponding to one point in PRSi (PRSi_EUD_1, Figure 1) for the different considered endpoints. The patient-specific EUD associated with the specific PRSi (PRSi_EUD) is the patient-specific PRSi times the PRSi_EUD_1 (i.e., for each patient PRSi_EUD=PRSi*PRSi_EUD_1). This quantity is established at a patient level.

For each patient, we included his PRSi_EUD in the DSM, adding to each voxel a dose proportionally to that voxel's contribution to the total EUD (i.e., the contribution is associated with the volume effect parameter n and the dose value, Figure 1).

We obtained PRSi-modulated DSMs: DSMs were (non-uniformly) shifted towards higher doses for PRSi>0, towards lower doses for PRSi<0, not changed for PRSi=0 (Figure 1).
We performed Cox-VBA on the obtained PRSi-modulated DSMs and considered correction for multiple testing for clusters using the permutation test coupled with Benjamin-Hockenberg (BH) correction.

3 Results

674 patients had long-term follow-up, DSMs and PRSi for the urinary endpoints, 778 for the rectal bleeding endpoint. The median follow-up was 24 months (75th percentile 60 months). 75% of patients had conventional fractionation (60-81 Gy), 25% received hypofractionation.

Toxicity crude rates were 4.8% for grade≥2 frequency, 7.5% for grade≥1 haematuria and 6.9% for grade≥2 rectal bleeding.
The PRSi associates with toxicity: mean PRSi 0.61/-0.40 in patients with/without toxicity, p<0.0001 (t-test) for frequency, 0.38/-0.02, p=0.02 for haematuria and 0.18/0.08, p=0.03 for rectal bleeding. The effective EUDs corresponding to one point in PRSi were 47 Gy for urinary frequency, 25 Gy for haematuria, and 69.3 Gy for rectal bleeding.

For grade≥2 frequency, the PRSi-modulated VBA (after BH correction) identifies a caudal-posterior region (also including the trigone) with a lateral extension covering 30% of the bladder wall and a cranio-caudal extension of 54%. HRs for the dose in this region are ~1.01-1.12 (per 1Gy), and the p-values range between 10E-8 and 0.05 (Figure 2). We found no specific area (after BH correction) without the inclusion of the modulation from the PRSi, and we had an increase in the slope of the dose-response relationship, with HRs up to 0.06 higher (per 1Gy, PRSi-modulated vs no modulation).
For grade≥1 haematuria, PRSi-modulated VBA detects a large region covering the whole extension of the bladder wall, except for two small areas located centrally in the cranio-caudal direction, one in the anterior part and the other in the posterior position. HRs for the dose are ~1.01-1.07 (per 1Gy) and p=0.0002-0.05. The extension of the region exhibiting statistically significant differences increased by 11% compared to VBA without PRSi-modulation, and HRs increased up to 0.03 (per 1Gy).

For grade≥2 late rectal bleeding, the PRSi-modulated VBA detects a region in the posterior rectal wall covering the central third of the rectum in the cranio-caudal direction. HRs for the dose are ~1.02-1.06 (per 1Gy) and p=0.0002-0.05. (Figure 3)

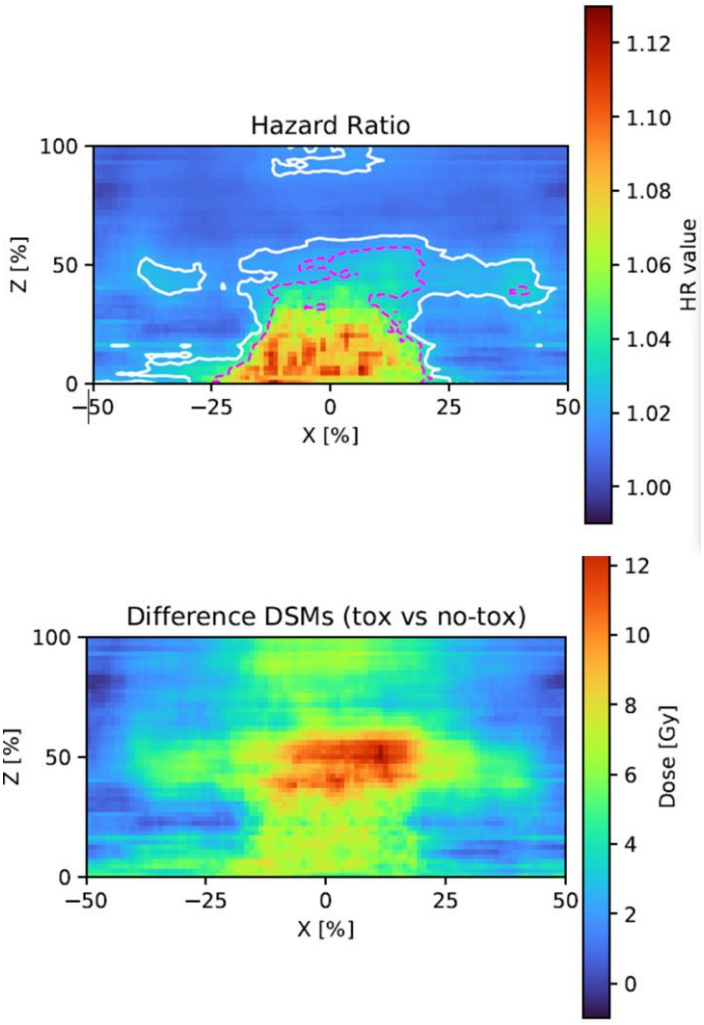


Figure 2: Results of the genetically-adjusted Voxel-Based Analysis for grade 2+ late urinary frequency: map for Hazard Ratios (left) and map of differences (right) in the genetically adjusted doses. Solid white lines depict the iso-p value level 0.05, dashed magenta lines the iso-p value level 0.05 after the correction for the false discovery rate (Benjamin-Hockenberg).

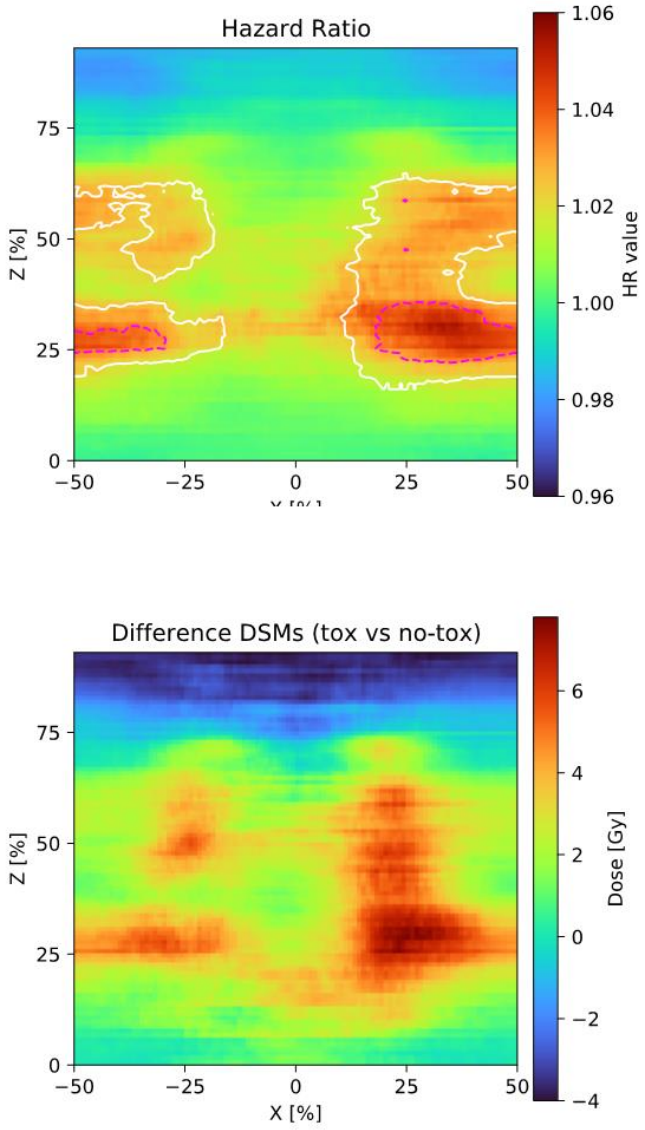


Figure 3: Results of the genetically-adjusted Voxel-Based Analysis for grade 2+ late rectal bleedingy: map for Hazard Ratios (left) and map of differences (right) in the genetically adjusted doses. Solid white lines depict the iso-p value level 0.05, dashed magenta lines the iso-p value level 0.05 after the correction for the false discovery rate (Benjamin-Hockenberg).

4 Discussion

We have here proposed a new method to include the single-patient genetic risk at a patient level in VBA and showed the benefit through an application on a large cohort of prostate cancer patients with long-term follow-up.

The method creates PRSi-modulated dose maps and can be naturally translated and extended into the inclusion of modulation from other patient-specific risk factors, even considering multiple features together.

In the present analysis, we identified specific bladder and rectal subregions associated with late urinary frequency, haematuria, and rectal bleeding.
The modulation by the PRSi increases the heterogeneity of dose distributions in the patient population, escalating the doses for radiosensitive patients and de-escalating the doses for radioresistant patients. The enhanced heterogeneity rises the areas classified as significantly different and re-gains steepness for the dose-response curve at a voxel level

.5 Conclusion

These results suggest the value of clarifying patient-specific risk factors to synergistically clarify dose-response relationships and the identification of heterogeneity in the radiobiology across tissues.

Acknowledgements

Acknowledgements to REQUITE EU 7th FP GA 601826, RADprecise ERAPERMED2018-244 and JTC ERA PerMed 2020, PerPlanRT Project - Italian Ministry of Health ERP-2020-23671125

References

[1] Green A et al. Image Based Data Mining Using Per-voxel Cox Regression. Front Oncol. 2020;10:1178. doi:10.3389/fonc.2020.01178

[2] Massi MC, et al. A Deep Learning Approach Validates Genetic Risk Factors for Late Toxicity After Prostate Cancer Radiotherapy in a REQUITE Multi-National Cohort. Front Oncol. 2020;10:541281. doi:10.3389/fonc.2020.541281

[3] Franco NR et al. Development of a method for generating SNP interaction-aware polygenic risk scores for radiotherapy toxicity. Radiother Oncol. 2021;159:241-248. doi:10.1016/j.radonc.2021.03.024

[4] Chia-Ho Hua et al. Improving Pediatric Normal Tissue Radiation Dose-Response Modeling in Children With Cancer: A PENTEC Initiative 25, 2024. Doi:10.1016/j.ijrobp.2023.11.038

[5] Iacovacci J et al.. Dependence of the AUC of NTCP models on the observational dose-range highlights cautions in comparison of discriminative performance. Phys Med. 2023;113:102654. doi:10.1016/j.ejmp.2023.102654

Prediction of survival outcomes and relevant tumor immune infiltrates using radiomic network analysis

Jung Hun Oh, Aditya Apte, Harini Veeraraghavan, Amita Shukla-Dave, and Joseph O. Deasy

Department of Medical Physics, Memorial Sloan Kettering Cancer Center, New York, USA

Abstract We propose to utilize a network-based method to analyze radiogenomic (radiomic/genomic) data. This approach enables the construction of sparse radiomic networks using the graphical LASSO (Least Absolute Shrinkage and Selection Operator) algorithm based on partial correlations among radiomic features. We developed and applied an optimal transport-based *k*-means algorithm to the resulting radiomic networks to identify clusters. The differences in survival rates and immune cell types between the resulting clusters were then assessed. In this study, we analyzed computed tomography (CT) radiomic features and RNA-Seq data used to quantify CIBERSORT scores in head and neck cancer and lung cancer patients taken from the TCIA and TCGA databases. Our proposed method identified sub-groups with the significant differences in survival outcomes and immune cell types. The promising results suggest that the geometry-based radiomic network approach could help identify patients at high risk of poor prognosis.

1 Introduction

Radiogenomics investigates the relationships between quantitative features extracted from medical images and genomic phenotypes to characterize tumor heterogeneity and understand the underlying biological correlates associated with treatment response [1]. Most radiogenomic studies have focused on developing supervised machine learning approaches to identify a set of important features and build predictive models for clinical outcomes [2, 3]. The main disadvantage of supervised machine learning is the need of parameter fitting which may cause an overfitting problem and therefore hamper validation of the predictive models previously developed [4]. Furthermore, most machine learning methods lack the capability to visualize and interpret the correlations among various features.

To address these issues, we propose a novel unsupervised method to analyze radiogenomic data [5]. First, sparse radiomic networks are constructed employing graphical LASSO based on partial correlations among radiomic features, removing likely spurious or false positive connections and visually representing the dependency among radiomic features as a form of graph [6]. Subsequently, an unbalanced optimal transport-based *k*-means method is employed to cluster samples on the resulting networks. Lastly, we assess the differences in survival outcomes and tumor immune infiltrates between the resulting clusters. In this study, we will show the usefulness of our proposed approach in radiogenomic analysis applied to CT radiomic features and CIBERSORT scores in head and neck cancer and lung cancer [7].

2 Materials and Methods

Data: We analyzed radiomic features derived from pre-treatment CT scans and CIBERSORT scores computed on RNA-Seq gene expression profiles for two cancer types: head and neck squamous cell carcinoma (HNSCC) and non-small cell lung cancer (NSCLC). From primary tumors manually contoured on CT scans, IBSI compliant 397 radiomic features, consisting of first-order statistics, shape-based features, and higher order texture features, were extracted using our in-house computational environment for radiotherapy research (CERR) radiomics toolbox [8].

For HNSCC, CT DICOM data for 77 patients were downloaded from The Cancer Imaging Atlas (TCIA) [9, 10]. For the matched 77 cases, associated RNA-Seq gene expression profiles were downloaded from The Cancer Genome Atlas (TCGA) and 22 immune cell types quantified by the CIBERSORT deconvolution method were estimated by utilizing the R immunedeconv library [11].

For NSCLC**,** CT DICOM data for 117 patients were downloaded from the TCIA. RNA-Seq data for the matched 117 cases were downloaded from the Gene Expression Omnibus (GEO) database (GSE103584) and CIBERSORT scores were estimated for immune cell abundance analysis [12].

Methods: Prior to network analysis, radiomic features highly correlated with tumor volume were removed with a threshold of Spearman correlation coefficient of 0.8. For a network representation of radiomic features, a Gaussian graphical model was then employed, which optimized the fit of the regularized partial correlation network constructed using graphical LASSO [13, 14], by minimizing the extended Bayesian information criterion (EBIC). A network-based *k*-means algorithm we have developed was adopted to cluster samples on the resulting sub-networks. This method employs an unbalanced extension of optimal transport to minimize the sum of within-cluster Wasserstein distances between the data points and their respective cluster centroid [5, 15].

For the identified clusters, Kaplan-Meier survival analysis was then conducted with log-rank test. Further, the differential assessment of CIBERSORT scores between the resulting clusters was carried out.

3 Results

In the HNSCC cohort, 55 patients (71.4%) were alive without progression at the time of assessment. The median progression free survival time was 2.0 years. The median

age of patients was 61 years. Twenty patients (26.0%) were female and 13 (16.9%) had HPV-positive tumors. The network analysis resulted in two sub-groups denoted as higher and lower-risk groups. Kaplan-Meier analysis on the largest sub-network, consisting of 65 radiomic features, showed a significant difference in progression free survival between the higher and lower-risk groups with log-rank p = 0.02 as shown in **Figure 1**. CIBERSORT analysis indicated that the lower-risk group had increased presence of neutrophils with p = 0.022.

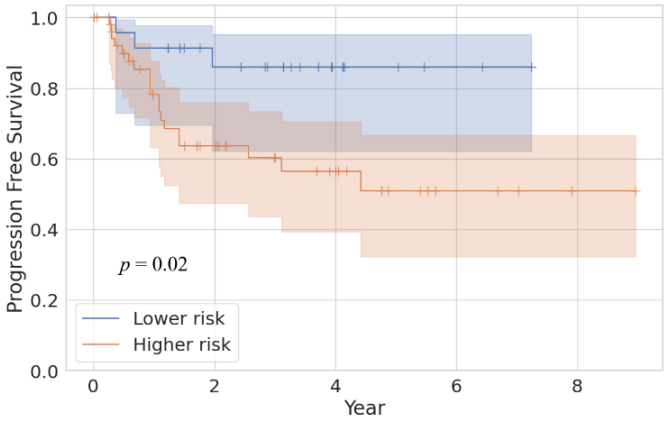


Figure 1: Kaplan-Meier survival analysis for 77 head and neck cancer patients.

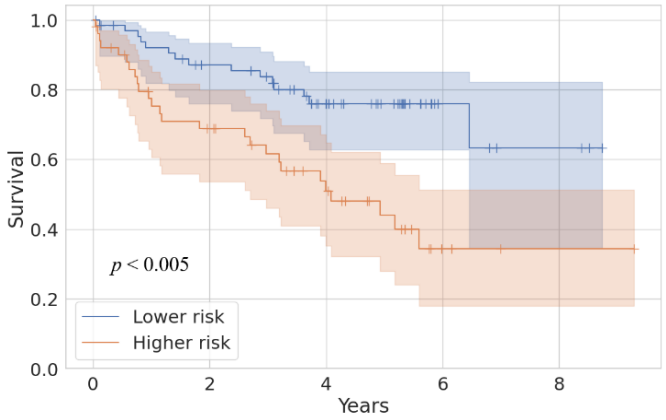


Figure 2: Kaplan-Meier survival analysis for 117 non-small cell lung cancer patients.

In the NSCLC cohort, 76 patients (65.0%) were alive at the time of assessment. The median survival time was 3.7 years. The median age of patients was 69 years and 29 patients (24.8%) were female. Kaplan-Meier analysis on the largest sub-network did not show a statistically significant difference between the higher and lower-risk groups identified using the optimal transport-based *k*-means method. However, Kaplan-Meier analysis on the second largest sub-network, consisting of 56 radiomic features, showed a statistically significant difference with $p < 0.005$ (**Figure 2**). Interestingly, 50 features were overlapping between the largest sub-network in HNSCC and the second largest sub-network in NSCLC. In CIBERSORT analysis, resting dendritic cells and activated mast cells indicated statistically significant differences between the higher and lower-risk groups with p = 0.0074 and 0.0025, respectively. For both the immune cell types, the higher-risk group had increased abundance.

4 Discussion

Radiogenomics explores the correlations between radiomic features and genomic profiles to address tumor heterogeneity and uncover the underlying biological correlates associated with clinical outcomes. However, little has been shown about the association between radiomic features and cancer immunity using a network analysis approach in an unsupervised manner.

The network form of radiomic features may be intractably complex due to their high correlations to each other. Regularized partial correlation network analysis can help remove spurious or false positive connections and make the network representation more interpretable, thereby facilitating the identification of plausible biological correlates in a downstream analysis. The optimal transport-based *k*-means method applied to the resulting sub-networks enabled the identification of patients at high risk of poor prognosis and showed that cancer immunity is likely to be relevant to the prognosis difference. The largest sub-network in HNSCC and the second largest sub-network in NSCLC resulted in the significant survival difference between the higher and lower-risk groups. Notably, 50 radiomic features between the two sub-networks were overlapping, suggesting that these are likely to be plausible biomarkers associated with survival outcomes. However, additional analysis is needed to validate the radiomic features and their associations with cancer immunity.

We plan to apply the network-based radiomic analysis approach to other imaging modalities and extend it to supervised modeling to build an individualized predictive model of clinical outcomes.

5 Conclusion

In this study, we showed the potential of using a network-based approach for radiomic analysis. The optimal transport-based *k*-means method applied to the resulting sub-networks enabled the identification of patients at high risk of poor prognosis with significantly different immune cell types. The approach is data-efficient because extensive parameter fitting is not required. This radiomic network analysis approach results in an embedding that can detect subtle differences or similarities between images, thus providing a data-efficient and robust tool to detect radiomic subtypes.

Acknowledgment

This research was funded in part through a National Institutes of Health/National Cancer Institute Cancer Center support grant P30 CA008748 and a Breast Cancer Research Foundation grant BCRF-17-193.

References

[1] M. Fan, P. Xia, R. Clarke, Y. Wang, and L. Li, "Radiogenomic signatures reveal multiscale intratumour heterogeneity associated with biological functions and survival in breast cancer," Nat Commun, vol. 11, no. 1, p. 4861, 2020, doi: 10.1038/s41467-020-18703-2.

[2] Z. Li, R. Wang, L. Wang, C. Tan, J. Xu, J. Fang, and J. Xian, "Machine Learning-Based MRI Radiogenomics for Evaluation of Response to Induction Chemotherapy in Head and Neck Squamous Cell Carcinoma," Acad Radiol, 2023, doi: 10.1016/j.acra.2023.10.054.

[3] J. Porto-Alvarez et al., "CT-Based Radiomics to Predict KRAS Mutation in CRC Patients Using a Machine Learning Algorithm: A Retrospective Study," Biomedicines, vol. 11, no. 8, 2023, doi: 10.3390/biomedicines11082144.

[4] P. Charilaou and R. Battat, "Machine learning models and over-fitting considerations," World J Gastroenterol, vol. 28, no. 5, pp. 605-607, 2022, doi: 10.3748/wjg.v28.i5.605.

[5] J. H. Oh et al., "Reproducibility of radiomic features using network analysis and its application in Wasserstein k-means clustering," J Med Imaging (Bellingham), vol. 8, no. 3, p. 031904, 2021, doi: 10.1117/1.JMI.8.3.031904.

[6] S. Epskamp and E. I. Fried, "A tutorial on regularized partial correlation networks," Psychol Methods, vol. 23, no. 4, pp. 617-634, 2018, doi: 10.1037/met0000167.

[7] A. M. Newman et al., "Robust enumeration of cell subsets from tissue expression profiles," Nat Methods, vol. 12, no. 5, pp. 453-7, 2015, doi: 10.1038/nmeth.3337.

[8] A. P. Apte et al., "Technical Note: Extension of CERR for computational radiomics: A comprehensive MATLAB platform for reproducible radiomics research," Med Phys, 2018, doi: 10.1002/mp.13046.

[9] E. Katsoulakis et al., "Radiomic analysis identifies tumor subtypes associated with distinct molecular and microenvironmental factors in head and neck squamous cell carcinoma," Oral Oncol, vol. 110, p. 104877, 2020, doi: 10.1016/j.oraloncology.2020.104877.

[10] K. Clark et al., "The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository," J Digit Imaging, vol. 26, no. 6, pp. 1045-57, 2013, doi: 10.1007/s10278-013-9622-7.

[11] G. Sturm, F. Finotello, and M. List, "Immunedeconv: An R Package for Unified Access to Computational Methods for Estimating Immune Cell Fractions from Bulk RNA-Sequencing Data," Methods Mol Biol, vol. 2120, pp. 223-232, 2020, doi: 10.1007/978-1-0716-0327-7_16.

[12] S. Bakr et al., "A radiogenomic dataset of non-small cell lung cancer," Sci Data, vol. 5, p. 180202, 2018, doi: 10.1038/sdata.2018.202.

[13] J. H. Oh and J. O. Deasy, "Inference of radio-responsive gene regulatory networks using the graphical lasso algorithm," BMC Bioinformatics, vol. 15 Suppl 7, no. Suppl 7, p. S5, 2014, doi: 10.1186/1471-2105-15-S7-S5.

[14] J. Friedman, T. Hastie, and R. Tibshirani, "Sparse inverse covariance estimation with the graphical lasso," Biostatistics, vol. 9, no. 3, pp. 432-41, 2008, doi: 10.1093/biostatistics/kxm045.

[15] M. Pouryahya, J. H. Oh, P. Javanmard, J. C. Mathews, Z. Belkhatir, J. O. Deasy, and A. R. Tannenbaum, "aWCluster: A Novel Integrative Network-Based Clustering of Multiomics for Subtype Analysis of Cancer Data," IEEE/ACM Trans Comput Biol Bioinform, vol. 19, no. 3, pp. 1472-1483, 2022, doi: 10.1109/TCBB.2020.3039511.

Radiomic Extraction and Analysis for DICOM Images to Refine Objective Quality Control (READII-2-ROQC): An open-source foundation for head and neck radiomics

Katy L. Scott[1], Sejin Kim[1,2,6], Jermiah J. Joseph[1], Matthew Boccalon[1], Mattea Welch[1,6], Mogtaba Alim[8] Umar Yousafzai[7], Ian Smith[1,2], Chris McIntosh[2,3,4], Katrina Rey-McIntyre[4], Shao Hui Huang[4,5], Tirth Patel[3,4,6], Tony Tadic[4,5,6], Brian O'Sullivan[4], Scott V. Bratman[2,4,5], Andrew J. Hope[4,5], and Benjamin Haibe-Kains[1,2,6]

[1]Princess Margaret Research, Princess Margaret Cancer Centre, Toronto, Canada
[2]Department of Medical Biophysics, University of Toronto, Toronto, Canada
[3]TECHNA Institute, Toronto, Canada
[4]Radiation Medicine Program, Princess Margaret Cancer Centre, Toronto, Canada
[5]Department of Radiation Oncology, University of Toronto, Toronto, Canada
[6]Cancer Digital Intelligence, Princess Margaret Cancer Centre, Toronto, Canada
[7]Cheriton School of Computer Science, University of Waterloo, Waterloo, Canada
[8]Department of Computer Science, University of Toronto, Toronto, Canada

Abstract Radiomics has exploded in oncology applications to develop signatures capable of predicting cancer outcomes and survival. However, radiomic analysis suffers from inconsistent methodology, insufficient data availability, and a lack of open-source processing tools to handle common radiological image formats. In this work, we present the READII-2-ROCQ pipeline: an end-to-end pipeline using DICOM-RT objects, image pre-processing, radiomic feature extraction, and publishes an organized data structure on a publicly available repository (ORCESTRA). Further, new quality control methods to assess radiomic features through negative control image generation are included. As a proof-of-concept, we explored a large head and neck cancer dataset (RADCURE) and replicated three previously published radiomics models. Our results highlight the importance of determining if radiomic signatures are using underlying imaging data. The main contributions of this work are the pipeline, an analysis-ready radiomics dataset, and open-source code to generate negative controls, so all radiomic studies can be READII-2-ROCQ.

1 Introduction

With the purported future of oncological care being precision medicine, the hunt for predictive biomarkers has become a focal point. A potential source lies in radiological imaging, which has motivated the field of radiomics for the last decade [1–4]. Radiomics research, however, has been hampered by inconsistent methodology, despite efforts to establish standard features [5]. The release of the open-source PyRadiomics toolkit [6] was a significant and necessary step to standardize radiomics analysis, but the collation and distribution of publically available radiomics datasets remains poorly organized within the community. As a result, significant overhead remains when dealing with multiple training, testing, and validation datasets from both internal and external sources. Further, a recent study has raised the question of whether radiomic features with high predictive value are surrogates for tumour volume measurements [7]. There is a need for standard methodology for radiomic feature extraction, as well as large, publicly available radiomic datasets that have undergone rigorous processing to benchmark analyses.

In this study, we have developed Radiomic Extraction and Analysis for DICOM Images to Refine Objective Quality Control (READII-2-ROCQ): a reproducible, automated, open-source processing pipeline to generate analysis-ready radiomics data. We showcase the pipeline's capabilities by processing and analyzing the largest publicly available head and neck cancer (HNC) dataset, RADCURE [8], and compare three previously published radiomics models [1,7,9] using the resulting data. Data outputs have been made available via ORCESTRA (https://www.orcestra.ca), a web-app that hosts processed 'omics data.

2 Materials and Methods

The READII-2-ROCQ pipeline leverages three main tools: Med-ImageTools [10], Radiomic Extraction and Analysis for DICOM Images (READII), and ORCESTRA [11]. While the former two are imaging-specific, we have modified ORCESTRA to work with clinical radiological data.

The proposed pipeline was tested using the RADCURE [12] dataset, accessed through the Cancer Imaging Archive (TCIA). It consists of 3,346 HNC computed tomography (CT) image volumes, corresponding radiotherapy structure sets (RTSTRUCT) containing primary gross tumour volume (GTVp) contours, and clinical data.

The Med-ImageTools library was used to generate a database for each unique CT acquisition and RTSTRUCT, and record pertinent metadata, such as series identifiers and reference CT values, to be used in subsequent processing steps.

For each CT and RTSTRUCT, preprocessing, quality checking, and radiomic feature extraction was performed using the READII package. First, the summary file generated by Med-ImageTools was scanned to automatically determine the segmentation type present and then find the associated reference CT. This updated list of images was then passed to image loading and preprocessing.

The CT and RTSTRUCT were loaded in as SimpleITK image objects [13] using corresponding functions provided by Med-ImageTools. In loading, the GTVp segmentation was converted from a discrete point set to a 3D volume. Alignment of the GTVp was performed by setting the origin, direction, and spacing for the image based on the CT values. The contour label for voxels within the GTVp was automatically found by checking for non-zero values in the image.

For each GTVp, preprocessing, quality checking, and radiomic feature extraction was performed using PyRadiomics. For preprocessing, the functions utilized were: checkMask, which checks the alignment, dimension, and contour label; and cropToTumorMask, which crops both the CT and RTSTRUCT images to the boundaries of the GTVp to optimize the feature extraction. Feature extraction settings were passed in using a configuration YAML file as described in PyRadiomics documentation. Extraction settings from the RADCURE prognostic modeling challenge [8] were applied.

In order to test negative imaging controls including those proposed by Welch et al. [7], the READII package can generate three types of negative controls with three different mask options, for a total of nine settings. The negative control options are: shuffle, where voxel values are randomly permuted; random, where random voxel values are generated within the range of the original image; and random from distribution, where voxel values are randomly selected with replacement from the original image. The mask options are full image, target and non-target. To demonstrate the performance of the negative controls, we performed feature extraction as described above with the shuffle and random options applied to the full image mask. Examples of these negative control combinations are shown in Figure 1.

The standard for data organization on ORCESTRA [11] is the MultiAssayExperiment R object [14], designed to harmonize multiple experimental assays from an overlapping patient set. To leverage this for radiomics, each set of extracted features becomes an individual experiment with clinical data included as the primary metadata describing each patient.

To demonstrate the pipeline's utility, we replicated previously published survival analysis Cox proportional hazards models with the training and test cohorts from the RADCURE challenge subset [8]. Coefficients from the MW2018 [7] and Kwan [9] models were used to calculate prognostic index values for the test cohort. For comparison, we fit a Cox model to the RADCURE training cohort using the same radiomic signature and applied it to the test cohort. A univariate model for GTVp Mesh Volume was also tested. All models were compared using the concordance index (c-index).

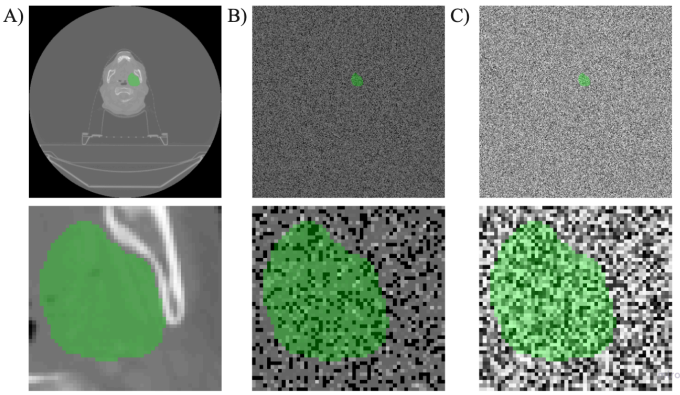


Figure 1. Original and cropped 2D axial slice extracted from example inputs to the radiomic feature extraction with the corresponding GTVp contour (green). (A) The original CT volume; (B) volume with shuffled voxel indices; (C) volume with randomly generated voxel values based on the range of original CT values.

To additionally explore the feature specific impacts of the negative controls on the PyRadiomics feature sets, cross-correlation plots were generated to display the correlation between feature extraction results for each PyRadiomics feature on both the original image and on each negative control image. The plots of correlation are grouped by PyRadiomics feature sets to show any feature group variations by feature class.

3 Results

We processed 2,994 patients with GTVp contours from varying primary tumour sites: 1076 patients with oropharynx, 713 with larynx, 319 with nasopharynx, 139 with hypopharynx, 81 with lip and oral cavity, 73 with paranasal, and 22 with esophagus. We extracted 1,317 radiomic features from the CT and negative controls for each GTVp.

The final data object containing all of these features, along with the clinical data and PyRadiomics configuration file, are available at https://www.orcestra.ca/radiomicset/10.5281/zenodo.8332910. The pipeline implementation for RADCURE is published at https://github.com/BHKLAB-DataProcessing/RADCURE_radiomics.

Results from our radiomics analysis are available in Table 1. The subset of 2,423 GTVps was split into training (n=1710) and test cohorts (n=713) based on the 'RADCURE-challenge' label in the clinical data. The Kwan model was tested with new oropharynx patients (n=366) not previously used in the Kwan paper. Model performance is similar whether the features were extracted from the CT or negative control samples, suggesting the radiomic signature is correlated with tumour volume, a known confounder of radiomics analysis, or at least invariant to the underlying imaging data. This is confirmed by the performance of the univariable volume model.

In Figure 2, we show the impact of a negative controls on different PyRadiomics features sorted by feature group by comparing the original image feature values to the shuffled or randomized image feature values. High Pearson correlations can be seen between the shape features of the original image and each of the feature groups in both the full image shuffled and randomized negative controls. In particular, features within the GLRLM, GLSZM, and GLDM groups. Correspondingly, feature groups with low correlations include first order and NGTDM, which reflects feature differences due to the underlying imaging data alterations.

	Image	c-Index (95% CI)	p.value
MW2018 [7]	Original	0.718 (0.668-0.770)	<0.001
	Shuffled	0.707 (0.657-0.757)	<0.001
	Random	0.699 (0.650-0.748)	<0.001
Kwan [9]	Original	0.273 (0.214-0.333)	<0.001
	Shuffled	0.708 (0.650-0.767)	<0.001
	Random	0.693 (0.633-0.752)	<0.001
Aerts RADCURE	Original	0.725 (0.675-0.776)	<0.001
	Shuffled	0.708 (0.659, 0.758)	<0.001
	Random	0.703 (0.653-0.751)	<0.001
Univariate Volume	Original	0.706 (0.656-0.756)	<0.001

Table 1: C-index and p-values from each of the replicated survival analysis models and a univariate volume model for comparison. 95% CI are shown for point-estimated c-index values. The replicate models all use the Aerts radiomic signature = Energy, Compactness 1, Gray Level Run Length Matrix Non-Uniformity, Wavelet HLH Gray Level Run Length Matrix Non-Uniformity.

When the 108 features shown in Figure 2 are filtered with a Pearson correlation threshold of 0.25, only 30 features pass in the shuffled image negative control and six pass for the randomized image.

4 Discussion

In this work, we present the READII-2-ROQC pipeline and demonstrate its use to explore negative radiomic controls with RADCURE by replicating three existing radiomics-based survival analysis models. In addition, we produced two of nine potential negative control images (Fig. 1) and performed correlative analysis to showcase how a common radiomic signature is impacted by negative controls.

Figure 2 highlights that there are a large number of high magnitude correlations between the original image feature values and the negative control image feature values indicating that some features may be largely invariant to the underlying imaging data.

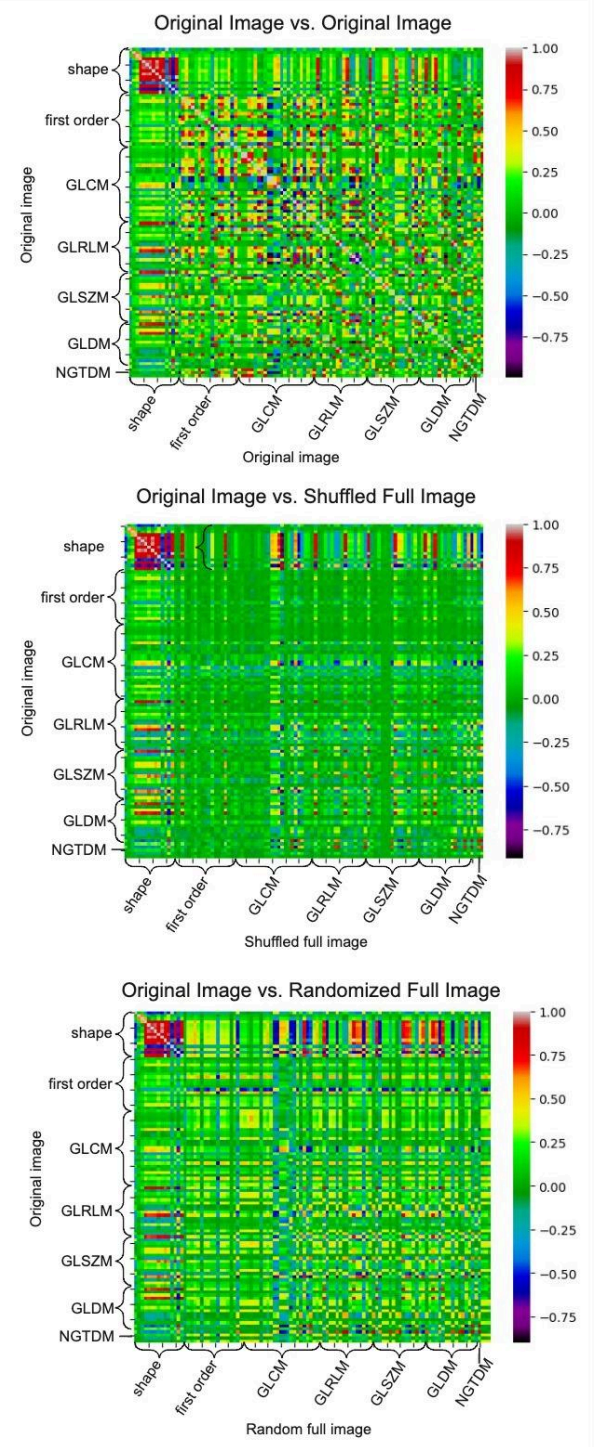


Figure 2. Cross-correlation heat map plots displaying the Pearson correlation between PyRadiomics features with no image filtering, sorted by feature groups, in the original image and the negative control images for the entire RADCURE dataset. The top heat map shows the cross-correlation of the features in the original image as a baseline reference.

Specifically, shape features, which are derived from the GTVp mask, are independent of the image information and highly correlated with the negative controls feature values. Therefore, if a radiomic signature is thought to be volume independent and is reflecting underlying image information beyond shape features, that model should perform poorly when given radiomic features extracted from negative controls generated by READII-2-ROQC.

In Table 1, we can see this is not the case for the radiomic signature we replicated. The MW2018 and Aerts RADCURE models both show comparable performance when radiomic features from original images and negative controls are used. Add in that the univariate volume model achieves similar results, and it becomes clear that this radiomic signature is not volume independent as expected.

As has been shown previously [7], radiomics features can be highly correlated with basic features such as GTV volume which limit the additional utility of radiomics models. To move forward, it is critical to implement standard methods to help identify features which are somewhat invariant relative to the underlying imaging characteristics to maximize the effectiveness of radiomics approaches.

Existing limitations of currently available analysis packages include a lack of support for raw DICOM files, the standard medical imaging format, as well as issues with implementing analysis pipelines in a consistent method. To address these issues, READII-2-ROQC provides end to end support for DICOM files via MedImageTools which then outputs the results of PyRadomics analysis with optional negative controls in an ORCESTRA object. As we have provided the final ORCESTRA objects for the RADCURE dataset, it allows anyone who has developed a radiomics model using PyRadiomics features to rapidly assess the extensibility of those models in our external dataset.

At present, the READII-2-ROQC pipeline has only been tested on CT DICOM images only but should be broadly applicable to any imaging dataset in DICOM format as the first step of the pipeline is based on Med-ImageTools has been tested on other DICOM data types (e.g. MR).

5 Conclusion

This standardized architecture framework and the publicly available processed RADCURE dataset can be used to benchmark new datasets or radiomics models semi-automatically. Future work will include organ at risk and nodal targets in the RADCURE dataset, evaluating more existing survival models with READII-2-ROCQ, and the production of ORCESTRA objects for other publicly available HNC datasets. We anticipate that this pipeline and the RADCURE objects generated could be a standard testing benchmark for future radiomics analyses and publications.

References

1. Aerts HJWL, Velazquez ER, Leijenaar RTH, Parmar C, Grossmann P, Carvalho S, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat Commun. 2014;5: 4006. doi:10.1038/ncomms5006
2. Gillies RJ, Kinahan PE, Hricak H. Radiomics: Images Are More than Pictures, They Are Data. Radiology. 2016;278: 563–577. doi:10.1148/radiol.2015151169
3. Rogers W, Thulasi Seetha S, Refaee TAG, Lieverse RIY, Granzier RWY, Ibrahim A, et al. Radiomics: from qualitative to quantitative imaging. Br J Radiol. 2020;93: 20190948. doi:10.1259/bjr.20190948
4. Lambin P, Rios-Velazquez E, Leijenaar R, Carvalho S, van Stiphout RGPM, Granton P, et al. Radiomics: extracting more information from medical images using advanced feature analysis. Eur J Cancer. 2012;48: 441–446. doi:10.1016/j.ejca.2011.11.036
5. Zwanenburg A, Vallières M, Abdalah MA, Aerts HJWL, Andrearczyk V, Apte A, et al. The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping. Radiology. 2020;295: 328–338. doi:10.1148/radiol.2020191145
6. van Griethuysen JJM, Fedorov A, Parmar C, Hosny A, Aucoin N, Narayan V, et al. Computational Radiomics System to Decode the Radiographic Phenotype. Cancer Res. 2017;77: e104–e107. doi:10.1158/0008-5472.CAN-17-0339
7. Welch ML, McIntosh C, Haibe-Kains B, Milosevic MF, Wee L, Dekker A, et al. Vulnerabilities of radiomic signature development: The need for safeguards. Radiother Oncol. 2019;130: 2–9. doi:10.1016/j.radonc.2018.10.027
8. Kazmierski M, Welch M, Kim S, McIntosh C, Rey-McIntyre K, Huang SH, et al. Multi-institutional Prognostic Modeling in Head and Neck Cancer: Evaluating Impact and Generalizability of Deep Learning and Radiomics. Cancer Research Communications. 2023;3: 1140–1151. doi:10.1158/2767-9764.CRC-22-0152
9. Kwan JYY, Su J, Huang SH, Ghoraie LS, Xu W, Chan B, et al. Radiomic Biomarkers to Refine Risk Models for Distant Metastasis in HPV-related Oropharyngeal Carcinoma. Int J Radiat Oncol Biol Phys. 2018;102: 1107–1116. doi:10.1016/j.ijrobp.2018.01.057
10. Kim S, Kazmierski M, Qu K, Peoples J, Nakano M, Ramanathan V, et al. Med-ImageTools: An open-source Python package for robust data processing pipelines and curating medical imaging data. F1000Research. 2023. p. 118. doi:10.12688/f1000research.127142.1
11. Mammoliti A, Smirnov P, Nakano M, Safikhani Z, Eeles C, Seo H, et al. Orchestrating and sharing large multimodal data for transparent and reproducible research. Nat Commun. 2021;12: 5797. doi:10.1038/s41467-021-25974-w
12. Welch, M. L. , Kim, S. , Hope, A. , Huang, S. H. , Lu, Z. , Marsilla, J. , Kazmierski, M. , Rey-McIntyre, K. , Patel, T. , O'Sullivan, B. , Waldron, J. , Kwan, J. , Su, J. , Soltan Ghoraie, L. , Chan, H. B. , Yip, K. , Giuliani, M. , Princess Margaret Head And Neck Site Group, Bratman, S. , … Tadic, T. Computed Tomography Images from Large Head and Neck Cohort (RADCURE) (Version 1) [Data set]. The Cancer Imaging Archive; 2023. doi:10.7937/J47W-NM11
13. Lowekamp BC, Chen DT, Ibáñez L, Blezek D. The Design of SimpleITK. Front Neuroinform. 2013;7: 45. doi:10.3389/fninf.2013.00045
14. Ramos M, Schiffer L, Re A, Azhar R, Basunia A, Rodriguez C, et al. Software for the Integration of Multiomics Experiments in Bioconductor. Cancer Res. 2017;77: e39–e42. doi:10.1158/0008-5472.CAN-17-0344

Predicting Radiation Pneumonitis using Pre-Radiotherapy CT Scans by Radiomic and a Pre-Trained CNN Based Feature Extraction

Matthew Gil[1,2], William Nailon[2,3], Stephen Harrow[2], Paul Murray[1], Sorcha Campbell[2], Ai Wain Yong[2], John Murchison[2], Gillian Ritchie[2], and Stephen Marshall[1]

[1]Department of Electronic and Electrical Engineering, University of Strathclyde, Glasgow, United Kingdom
[2]Edinburgh Cancer Centre, NHS Lothian, Edinburgh, Scotland
[3]School of Engineering, University of Edinburgh, Edinburgh, Scotland

Abstract A radiomic model to predict radiation pneumonitis, an adverse effect of radiotherapy, has been developed and tested using data from 66 lung cancer patients that received radiotherapy treatment, 12 of whom developed grade 2 or above pneumonitis. Image features were extracted from radiotherapy planning CT images using a radiomic approach and by leveraging a deep learning based UNet model, that had been pre-trained to segment lung volumes, as a feature extractor. Radiation dose and clinical features were also included. The AdaBoost algorithm for gradient boosted decision trees was used to create a prediction model which was trained and tested using a leave-one-out cross validation approach. When using only the dose and clinical features, an AUC ROC of 0.63 was achieved which was improved to 0.836 when including all of the CT image features. This shows that CT images, which are not currently applied clinically for pneumonitis prediction, calculated by both radiomic and our UNet based feature extraction are a source of predictive power and can greatly improve predictions from standard dose and clinical information.

1 Introduction

During radiotherapy (RT) to treat cancer, damage to certain organs, caused by the unavoidable radiation dose to healthy tissue, can have significant health implications for a patient up to and including death. Radiation pneumonitis, an inflammation of the lung tissue, is one of these potential health implications which is common among patients receiving a lung dose during their treatment. Current standard clinical RT practice includes capturing a CT image to localise anatomy and calculate the attenuation maps necessary for RT planning but this planning CT image is currently not used beyond these purposes. We have applied methods for the extraction of features from the lung region of RT planning CT scans of patients treated for lung cancer using both radiomic approaches and the utilisation of a CNN pre-trained for lung segmentation. Our methods show a large improvement to pneumonitis prediction on our dataset when using both methods for CT feature extraction compared to using only radiation dose and clinical features. Additionally, combining our method for feature extraction from a pre-trained UNet model with a radiomic feature extraction produced a further improvement to the prediction model.

Related Work Currently in clinical settings, the dose to organs at risk during RT is limited by simple dose metrics that have been determined from studies such as the QUANTIC study [1] which established dose-volume effects in all regions of human anatomy. Prediction of radiation pneumonitis from pre-RT data has previously been shown to be possible in the literature with much of the focus being on dosimetric and dosiomic approaches. Dose volume histogram features, calculated from the lung volume, including the mean lung dose (MLD), V_5 ,V_{10} and V_{20} have been shown to act as a predictor for radiation pneumonitis [2, 3]. More recently, spatial features extracted from the RT dose distribution, have been included in predictive models through radiomic [4] and CNN based [5] approaches.

Beyond the dose based prediction models, there has been previous work on imaging modalities for pneumonitis prediction. Radiomic approaches have been used to determine if a patient developed pneumonitis [6]. Less commonly, radiomic approaches have been applied to the pre-RT CT images to predict radiation pneumonitis [7–10].

Previously, pre-trained CNNs have been applied as feature extractors in the medical imaging domain for a variety of applications such as overall survival prediction from CT scans [11] and tuberculosis detection from x-ray images[12]. These works rely on CNNs previously trained for image classification tasks as opposed to our approach of using a CNN previously trained for image segmentation.

Contribution In this work, we have developed a method for feature extraction from a pre-trained segmentation CNN and applied it to the task of radiation pneumonitis prediction. Additionally we have applied more traditional radiomic approaches for feature extraction and combined this with the CNN based approach. This has advanced the sparse previous work on radiation pneumonitis prediction from CT scans. Additionally, the use of our dataset validates previous pneumonitis prediction work.

2 Method

2.1 Dataset

The dataset used in this retrospective study consists of intensity-modulated radiotherapy (IMRT) data from patients with non-small cell lung cancer (NSCLC) that were treated at the Edinburgh Cancer Centre, Western General Hospital, Edinburgh, between 01/01/2009 and 01/10/2010. During data collection, the occurrence of pneumonitis in each patient was determined from that patient's medical records by

experienced clinicians. Any patients with unclear medical notes regarding pneumonitis were not included. This generated a dataset of 66 patients, 12 of whom developed grade 2 or above radiation pneumonitis after their RT treatment where grade 2 radiation pneumonitis is pneumonitis presenting symptoms where some form of clinical intervention has been performed as defined by CTCAE (v5.0) [13]. The RT data available included the patients' RT planning CT scans, the RT planning dose distribution maps and the anatomical segmentations produced during the RT planning. Examples of a planning CT and dose map segmented to the lung volume, defined during RT planning, are shown in figure 1. The CT scans all had voxel dimensions of 1.0x1.0x3.0mm and slices in the transverse plane had a resolution of 512x512. Also available were some of the patients' clinical parameters such as their age. Of the 66 patients in the study, IMRT was applied to; 52 patients on a Varian Clinac 600C/D, 3 patients on a Varian Clinac iX and 9 patients on a Varian Clinac 21EX. All radiotherapy plans were calculated using the Varian Pencil Beam Convolution algorithm (v8.1.2).

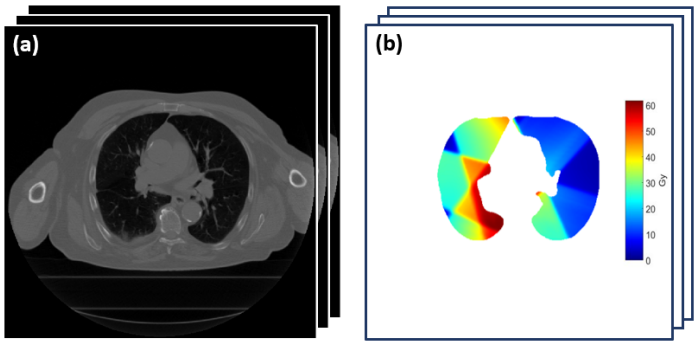


Figure 1: Examples of **(a)** a 3D planning CT image stack and **(b)** a 3D dose map of only the lung volume.

2.2 Radiomic CT Features

Prior to radiomic feature calculation, CT image preprocessing steps were applied and the volume of interest, which defined the voxels that radiomic features were calculated for, needs to be defined. The lung volume, segmented using the contours available from the RT planning process, was used as the initial volume of interest. This was then thresholded to only include voxels with Hounsfield unit (HU) values ranging from -1000 HU to -600 HU to specifically target the soft lung tissue by removing any vessels and the tumour volume. Before the radiomic features were calculated, the CT images were isotropically resampled to have voxel dimensions of 1.5x1.5x1.5mm. Quantisation was applied to the image to reduce the number of grey levels. A quantisation bin size of 25 was used which produced 16 grey levels in the defined volume of interest. Radiomic features were computed from the defined volume of interest within the CT images in Python using the PyRadiomics package (v3.1.0) [14]. All features from the following classes were included; first-order, GLCM, GLRLM, GLSZM, NGTDM and GLDM, further information on these metrics can be found in the PyRadiomics documentation. These features conform to the Image Biomarker Standardisation Initiative (IBSI) recommendations [15] and, where possible, the radiomic features were calculated in 3D. This process produced a feature table comprising 94 distinct radiomic features, each calculated from the CT lung region for every patient.

2.3 UNet CT Features

In addition to the PyRadiomics CT features, a publicly available pre-trained UNet segmentation model was leveraged as a feature extractor. The reason that a segmentation model was chosen as opposed to a classification model is that the features learned by a segmentation model trained on a varied medical dataset are likely to be more generalisable than features from a classification model trained for a specific medical task. Additionally, there are more publicly available medical image segmentation models available than classification models so there is a larger pool of previous work.
The UNet model, used as our feature extractor, was originally trained for segmenting lungs in 2D CT images, this network is referred to as "UNet (R231)" [16] and is available from *https://github.com/JoHof/lungmask*. Details of the models architecture and training can be found in the original paper [16]. Prior to extracting features from the CT image, no additional training of the model was performed. This approach of using a pre-trained model with no additional training was chosen due to the small size of our dataset, specifically the low number of positive cases, and the lack of public data available with radiation pneumonitis outcomes meaning training our own supervised classification CNN model was not possible. The activations from the convolutional layer before the final convolutional layer making the pixel-wise classifications were used as the features. This layer has a size of 256x256x64 meaning 64 features were extracted for every pixel. These pixel-wise features were converted to global features for the classification by averaging them over the spatial dimensions using a global average pooling layer added to the model so that for each slice of the 3D CT image, 64 features were calculated. The location of this global average pooling layer in the model architecture and the convolutional layer used for feature extraction is shown in figure 2. This feature averaging was done for every transverse slice of the 3D CT image containing any lung area. The features were then averaged in the z-direction producing 64 features from the 3D CT image for every patient.

2.4 Dose and Clinical Features

Dose and clinical features have previously been shown to be predictive features for radiation pneumonitis on their own as well as features that improve the performance of models based on additional features from separate sources, when included in a predictive model [1, 3, 4, 6–9]. For these

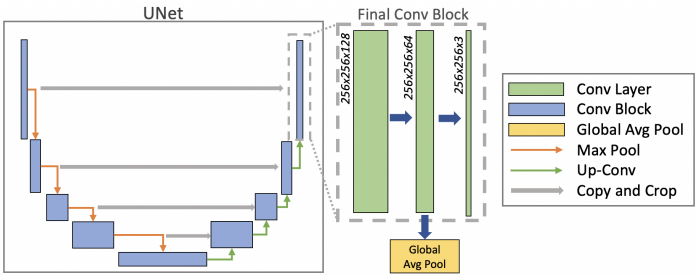


Figure 2: Location of the global average pooling layer, used to extract features from the second last convolutional layer of the UNet model. Further details of this UNet model can be found in the paper [16].

reasons, as well as fact that dose and clinical features will always be available pre-RT, they should always be included in radiation toxicity prediction studies as has been done here.

Clinical Features: Clinical features were extracted from the patients' medical information based on data that was deemed to be relevant to the classification task. The clinical features that were included were chosen based on their availability for all of the patients in the dataset as well the ability to present these features numerically. In total, 7 clinical features were used, these are; age, smoking status, COPD status, chemotherapy status, T-stage, N-stage and the primary tumour volume.

Dose Features: Dose volume histogram features were calculated from the RT planning dose map, segmented to only include the whole lung volume, excluding the tumour region defined by the gross tumour volume (GTV). The lung volume and GTV were available from the RT planning data. The dose volume histogram features calculated are; the mean lung dose (MLD), V_5, V_{10} and V_{20}. Here V_x is the percentage of the total lung volume receiving a dose above x Grays. The planned RT maximum dose and RT dose fractions were also included as features which were not calculated from the dose volume histogram but are parameters of the dose planning process giving a total of 6 dose features.

2.5 AdaBoost Classification

For the classification, boosted decision trees were used with the AdaBoost algorithm [17] in MATLAB (R2023a). Training and testing were applied using a leave-one-out cross validation (LOOCV) scheme. This involves training the model on all the data except one sample and then testing the model on that left-out sample to get a prediction. The trained model is then discarded and a different patient is selected as the one to be left out from training and the process repeats until testing has been completed on all of the patients. This is the same as k-fold cross validation where k equals the dataset size, so in this case it is 66-fold cross validation. During training, an increased misclassification cost was applied to the underrepresented class based on its proportional size in the dataset. Due to the small size of the dataset, it was not possible to perform parameter tuning of the model as this would require another hold-out set to validate the parameter tuning. The parameters used during training of the AdaBoost model are displayed in table 1.

The training and testing scheme was repeated on six different sets of features so that the contributions of the different features to the overall model performance could be observed. The six feature sets were; **1.** dose and clinical features, **2.** PyRadiomic features only, **3.** UNet features only, **4.** dose, clinical and PyRadiomic features, **5.** dose, clinical and UNet features, **6.** dose, clinical, PyRadiomic and UNet features.

Training Parameter	Value
Learning rate	1.00
Learning cycles	200
Maximal number of decision splits	10
Minimum observations per leaf	1
Minimum observations per branch node	2

Table 1: The training parameters used for the AdaBoost prediction models.

3 Results

The metrics produced during the LOOCV testing scheme are the accuracy, sensitivity, specificity and the AUC ROC with the AUC ROC being the most descriptive of the overall predictive power of the models and most commonly applied metric in the recent radiation pneumonitis prediction literature [4, 5, 7–10]. The results for the six different feature sets are shown in table 2.

Features	Acc	Sens	Spec	AUC
Dose, Clin	63.6	41.7	68.5	0.633
PyRad	72.7	66.7	74.1	0.760
UNet	66.7	58.3	68.5	0.717
Dose, Clin, PyRad	77.3	75.0	77.7	0.790
Dose, Clin, UNet	78.8	75.0	79.6	0.814
Dose, Clin, PyRad, UNet	**80.3**	**75.0**	**81.5**	**0.836**

Table 2: Results from the LOOCV over the whole dataset. The accuracy (Acc), sensitivity (Sens), specificity (Spec) are presented as percentages.

4 Discussion

The model using only the dose and clinical features achieved an AUC of 0.633. The models using only PyRadiomics and only UNet based CT features produced an AUC of 0.760 and 0.717 respectively, both an improvement on the model using only dose and clinical features. While it performed lower than the radiomic method, our technique of extracting features from the UNet model has achieved comparable results. Combining the CT features with the dose and clinical features achieved an AUC of 0.790 and 0.814 when using the PyRadiomics and UNet features respectively. This shows a

large improvement in the AUC of over 0.15 for both CT feature sources when combined with dose and clinical features compared to only using dose and clinical features. There is a large clinical relevance here as this shows that CT image features could improve the treatment planning process by allowing for the production of more personalised treatment plans. Finally, the model achieving the highest AUC used the dose, clinical, PyRadiomic and UNet features. This shows that combining CNN-based features with radiomic features can further improve radiation pneumonitis prediction model performance.

While these results are promising, it is important to highlight where future improvements can be made. The main limitation of this work is the dataset size. With a larger dataset, a validation dataset could be used for model hyperparameter tuning. Additionally, expanding this work to data from multiple centres would improve the robustness of the methods. Establishing a large benchmark dataset for radiotherapy outcome prediction would also allow for different methods to be directly compared. Model explainability is also an important future direction, when applying the UNet model for feature extraction, we have disregarded the spatial information but is possible to keep this information and isolate features to specific anatomical regions which would increase explainability.

5 Conclusion

In current clinical practice, only the dose and clinical features are considered during radiotherapy planning. This work has shown that CT images, which are collected during standard RT planning, can be a source of valuable information when predicting patient outcomes, specifically the occurrence of radiation pneumonitis. Additionally, we have shown that a pre-trained CNN for lung segmentation can be used as a feature extractor for this application and can improve the performance of radiomic based prediction models when the two CT image feature sources are combined.

Acknowledgements

This work was funded by the Engineering & Physical Sciences Research Council (EPSRC) (grant reference: EP/S022821/1) and the Beatson Cancer Charity (grant reference: 19-20-043). The work was facilitated and supported by the EPSRC, the Beatson Cancer Charity and the CDT in Applied Photonics.

References

[1] S. M. Bentzen, L. S. Constine, J. O. Deasy, et al. “Quantitative Analyses of Normal Tissue Effects in the Clinic (QUANTEC): an introduction to the scientific issues”. *International Journal of Radiation Oncology* Biology* Physics* 76.3 (2010), S3–S9.

[2] K. Boonyawan, D. R. Gomez, R. Komaki, et al. “Clinical and dosimetric factors predicting grade 2 radiation pneumonitis after postoperative radiotherapy for patients with non-small cell lung carcinoma”. *International Journal of Radiation Oncology* Biology* Physics* 101.4 (2018), pp. 919–926.

[3] D. A. Palma, S. Senan, K. Tsujino, et al. “Predicting radiation pneumonitis after chemoradiation therapy for lung cancer: an international individual patient data meta-analysis”. *International Journal of Radiation Oncology* Biology* Physics* 85.2 (2013), pp. 444–450.

[4] B. Liang, H. Yan, Y. Tian, et al. “Dosiomics: extracting 3D spatial features from dose distribution to predict incidence of radiation pneumonitis”. *Frontiers in oncology* 9 (2019), p. 269.

[5] B. Liang, Y. Tian, X. Chen, et al. “Prediction of radiation pneumonitis with dose distribution: a convolutional neural network (CNN) based model”. *Frontiers in oncology* 9 (2020), p. 1500.

[6] A. Cunliffe, S. G. Armato III, R. Castillo, et al. “Lung texture in serial thoracic computed tomography scans: correlation of radiomics-based features with radiation therapy dose and radiation pneumonitis development”. *International Journal of Radiation Oncology* Biology* Physics* 91.5 (2015), pp. 1048–1056.

[7] S. P. Krafft, A. Rao, F. Stingo, et al. “The utility of quantitative CT radiomics features for improved prediction of radiation pneumonitis”. *Medical physics* 45.11 (2018), pp. 5317–5324.

[8] C. Puttanawarut, N. Sirirutbunkajorn, N. Tawong, et al. “Radiomic and dosiomic features for the prediction of radiation pneumonitis across esophageal cancer and lung cancer”. *Frontiers in Oncology* 12 (2022), p. 197.

[9] Z. Zhang, Z. Wang, M. Yan, et al. “Radiomics and dosiomics signature from whole lung predicts radiation pneumonitis: A model development study with prospective external validation and decision-curve analysis”. *International Journal of Radiation Oncology* Biology* Physics* 115.3 (2023), pp. 746–758.

[10] Z. Zhang, Z. Wang, T. Luo, et al. “Computed tomography and radiation dose images-based deep-learning model for predicting radiation pneumonitis in lung cancer patients after radiation therapy”. *Radiotherapy and Oncology* (2023), p. 109581.

[11] R. Paul, S. H. Hawkins, L. O. Hall, et al. “Combining deep neural network and traditional image features to improve survival prediction accuracy for lung cancer patients from diagnostic CT”. *2016 IEEE international conference on systems, man, and cybernetics (SMC)*. IEEE. 2016, pp. 002570–002575.

[12] U. Lopes and J. F. Valiati. “Pre-trained convolutional neural networks as feature extractors for tuberculosis detection”. *Computers in biology and medicine* 89 (2017), pp. 135–143.

[13] A. Trotti, A. Colevas, A. Setser, et al. “CTCAE v3. 0: development of a comprehensive grading system for the adverse effects of cancer treatment”. *Seminars in radiation oncology*. Vol. 13. 3. Elsevier. 2003, pp. 176–181.

[14] J. J. Van Griethuysen, A. Fedorov, C. Parmar, et al. “Computational radiomics system to decode the radiographic phenotype”. *Cancer research* 77.21 (2017), e104–e107.

[15] A. Zwanenburg, M. Vallières, M. Abdalah, et al. “The image biomarker standardization initiative: standardized quantitative radiomics for high-throughput image-based phenotyping”. *Radiology* 295.2 (2020), pp. 328–338.

[16] J. Hofmanninger, F. Prayer, J. Pan, et al. “Automatic lung segmentation in routine imaging is primarily a data diversity problem, not a methodology problem”. *European Radiology Experimental* 4.1 (2020), pp. 1–13.

[17] Y. Freund and R. E. Schapire. “A decision-theoretic generalization of on-line learning and an application to boosting”. *Journal of computer and system sciences* 55.1 (1997), pp. 119–139.

Integrating Radiomics Harmonization and Deep Survival Analysis for Enhanced Predictive Modeling in a Multicentric Context

Nassib Abdallah[1,2,4], Jean-Marie Marion[2,3], Pierre Chauvet[2,3], Clovis Tauber[4], Thomas Carlier[5], and Mathieu Hatt[1]

[1]Laboratory of medical information processing (LaTIM), INSERM UMR 1101, Univ Brest, Brest, France
[2]LARIS, Université d'Angers, Angers, France
[3]Catholic University of the West, Angers, France
[4]INSERM, Imaging & Brain, Université de Tours, Tours, France
[5]University Hospital of Nantes, Nantes, France

Abstract This paper compares harmonization methods in the context of positron emission tomography / computed tomography (PET/CT) radiomics analysis for predicting recurrence-free survival (RFS) in head and neck cancer (HNC) patients, using the multicentre HECKTOR 2022 challenge dataset. We present harmonization techniques—HarMSTD and HarCenter—as alternatives to the popular COMBAT method. Notably, these methods do not require model retraining with new data, enhancing their practical utility. Our results demonstrate the benefit of harmonization in this context, with improved concordance index (C-index) values. The paper also compares our findings with existing literature, positioning our methods favorably in the context of the HECKTOR 2022 challenge. These novel approaches offer promising advancements in oncological predictive modeling, emphasizing the need for tailored harmonization techniques in multicenter clinical studies.

1 Introduction

The prediction of recurrence-free survival (RFS) in head and neck cancer (HNC) patients treated with radiotherapy (RT) can contribute to personalized management. In this context, large-scale datasets, like the one from the Hecktor 2022 challenge [1], play a crucial role. They serve not only as a resource for the development and validation of predictive models but also enable the identification of patients at heightened risk of local recurrence in the primary tumor volume[2]. The HECKTOR challenge presented two main tasks: automating detection and segmentation of lesions (tumors and lymph nodes) in Flurodeoxyglucose (FDG) positron tomography emission (PET)/computed tomography (CT) scans, and predicting RFS using PET/CT images and clinical variables. During the 2022 challenge, the best teams achieved C-index values between 0.66 and 0.68 [3] [4], [5], [6]either by relying on radiomics feature engineering modeling or deep learning approaches. We focus in the present work on the RFS prediction task, in which we aim to evaluate data harmonization methods, including novel and established techniques (e.g. ComBat). Significantly, none of the existing work has addressed the variability that resides between different acuiqisition centers. This is a key contribution of this paper.

2 Materials and Methods

2.1 Dataset

The HECKTOR 2022 challenge provided a dataset of FDG-PET/CT images of 828 HNC patients from 9 acquisition centers. The dataset is already splitted into 489 patients for training/validation and 339 for testing. Evaluation of RFS prediction model was carried out using the concordance index (C-index).

2.2 Methods

2.2.1 Data Collection and Preprocessing

We exploited the available PET/CT images, along with corresponding mask images and clinical information. This included extracting patient identifiers from file names and merging image paths with clinical endpoints and patient information to form a unified dataset. Clinical data were cleansed of missing values, and categorical variables like gender were numerically encoded for subsequent analysis. Special attention was given to handling errors such as dimension mismatches and mask value inconsistencies, ensuring comprehensive and accurate feature extraction.

2.2.2 Feature Extraction

Radiomic features were extracted from the PET and CT images using the corresponding masks. We computed 112 IBSI (imaging biomarker standardization initiative)-compliant radiomic features (shape, intensity and textures) for each PET and CT scan, with the *RadiomicsFeatureExtractor* from the *pyradiomics* library, configured to correct for mask inconsistencies.

2.2.3 Harmonization Techniques

Given the multicenter nature of the dataset, we implemented and compared several harmonization techniques to mitigate batch effects and inter-scanner variations and compare with the combat method. For all harmonization methods under comparison, our first step was to determine a reference center.

We selected Center 5, which had the largest patient cohort (397 patients), as our reference. This choice was based on the premise that a larger dataset would provide a more robust and representative statistical distribution for normalization purposes.

Mean, STandardDeviation Variable Harmonization (HarMSTD) : The MSTD Harmonization process is designed to align the statistical characteristics of feature datasets derived from various centers. This harmonization is achieved by adjusting the target dataset to match the mean and standard deviation of the corresponding features in the reference dataset. The transformation of each feature x in the target dataset is governed by the following equation:

$$x_{\text{new}} = \frac{\sigma_{\text{ref}}}{\sigma_{\text{target}}} \times (x - \mu_{\text{target}}) + \mu_{\text{ref}} \quad (1)$$

where:

- μ_{ref} is the mean of the reference dataset's feature.
- μ_{target} is the mean of the target dataset's corresponding feature.
- σ_{ref} is the standard deviation of the reference dataset's feature.
- σ_{target} is the standard deviation of the target dataset's corresponding feature.

This formula ensures that each feature in the target dataset is scaled in such a way that its statistical properties align with those of the reference dataset, thereby enabling a consistent and comparable analysis across datasets from different centers.

ComBat Harmonization (HarComBat) : The ComBat Harmonization method is tailored to correct for batch effects in multi-center studies. This method uses an empirical Bayes framework to adjust data for known batch effects, ensuring comparability across different datasets. The ComBat model is formalized as follows:

$$Y_{ij} = \alpha + X_i\beta + \delta_{ij} + \gamma_{ij}\varepsilon_{ij} \quad (2)$$

where:

- Y_{ij} represents the observed value for the ith subject in the jth batch.
- γ_j denotes the multiplicative batch effect for the jth batch.
- X_i represents covariates for the ith subject.
- β is the coefficient for the covariates.
- δ_{ij} is the additive batch effect for the ith subject in the jth batch.
- ε_{ij} is the residual error term.

ComBat adjusts each feature in the dataset by estimating and correcting for these batch-specific effects while preserving the biological variability. This process involves standardizing the data within each batch, estimating batch effects, and then re-scaling the data to remove these effects.

Center-Based Harmonization (HarCenter) : In our approach, titled Harmonization Based on Center, we focused on addressing the multicenter nature of the dataset by identifying a reference center and applying iterative transformations. Our primary goal was to align the data distribution of each center with that of the reference center, within a specified standard deviation range. Data were grouped by CenterID, derived from the initial three digits of each patient's ID. We combined clinical data with radiomic features, ensuring numeric consistency across all variables. Non-numeric columns were excluded to maintain a focus on quantitative analysis. The resulting dataset comprised both training (x_train) and testing (x_test) subsets. StandardScaler was initially considered for normalization; however, we opted for a custom alignment strategy. For each feature, we calculated the mean and standard deviation of the reference center. We then normalized the data from other centers, aligning each feature iteratively. Features were adjusted to fall within the mean $\pm$ 2 standard deviations of the reference distribution. This iterative approach aimed to minimize distributional discrepancies across centers, thus reducing batch effects and enhancing the comparability of multicentric data. Upon aligning the data, we constructed the harmonized dataset, combining the aligned radiomic features with the original clinical data. The final datasets, included both clinical and harmonized radiomic features, segregated into training and testing cohorts with the indexes of patients for training and testing provided by the challenge.

2.2.4 Post Harmonization Processing

Following the harmonization of features extracted from PET and CT scans, Principal Component Analysis (PCA) was employed on the concatenated feature set to both train and subsequently transform the test set based on the PCA model. The dimensionality reduction via PCA was configured to retain components that account for 95% of the explained variance, a strategic choice aimed at reducing the feature space while preserving the majority of the informative variance. This reduction was imperative to streamline the subsequent survival analysis, ensuring the model's efficiency and effectiveness by focusing on the most relevant features derived from the imaging data.

2.2.5 Survival Analysis using DeepHitSingle

In our study, we employed the DeepHitSingle model from the *pycox* library for survival analysis. This deep learning-based

approach is tailored to handle time-to-event data, aligning with our objective of predicting patient outcomes. DeepHitSingle is distinguished by its capability to simultaneously address competing risks, offering a nuanced approach to survival analysis beyond traditional methodologies. Model fitting was conducted through Optuna-based hyperparameter optimization, an automatic optimization framework that facilitates the selection of optimal model parameters. Key hyperparameters adjusted included the learning rate (1e-5, 1e-2) and the dimensions of the hidden layers (2 or 3 layers) and hidden neurons (32,64,128) within the neural network. The selection of these parameters was guided by the goal of maximizing the C-index. To evaluate the robustness of our approach, 20 different splits were used for the survival analysis for each technique.

3 Results

The results, as shown in Tables 1 and 2, evaluate the impact of various data harmonization methods on the predictive accuracy of RFS using the c-index as a measure. These evaluations were performed on the test set of the HECKTOR 2022 challenge dataset. Table 3 compares existing works related to Task 2 of the HECKTOR 2022 challenge, situating our approach within the broader research context.

Harmonization Method	c-index Value on PET/CT
HarMSTD	0.794 ± 0.102
HarCenter	0.784 ± 0.134
HarCombat	0.773 ± 0.138
No Harmonization	0.764 ± 0.132

Table 1: C-index values over 20 runs without or with each harmonization methods, not including clinical variables.

Harmonization Method	c-index Value on PET/CT
HarMSTD	0.801 ± 0.113
HarCenter	0.767 ± 0.129
HarCombat	0.777 ± 0.105
No Harmonization	0.777 ± 0.122

Table 2: C-index values over 20 runs without or with each harmonization methods, including clinical variables.

These tables comprehensively display the c-index values for different harmonization methods, with and without clinical variables, over 20 runs. They also summarize the c-index scores of various research teams in the HECKTOR 2022 challenge, highlighting the competitive position of our approach.

Reference	Method Type	C-index
VokCow Team [7]	Deep Learning	0.640
Retina Unet and Multi-Label Unet [8]	Deep Learning	0.635
DL and Radiomics Models [3]	Deep Learning	0.668
DL and Radiomics Models [3]	Feature Engineering	0.646
Team MLC [9]	Feature Engineering	0.656
AIRT Team [4]	Feature Engineering	0.670
3D nnU-Net with DiceTopK [10]	Feature Engineering	0.639
HNT-AI Model [11]	Deep Learning	0.627
Radiomics-Enhanced Deep MTS [5]	Feature Engineering	0.681
LITO team [6]	Feature Engineering	0.682
Our best approach	Feature Engineering	0.801 ± 0.113

Table 3: Summary of HECKTOR 2022 challengers results.

4 Discussion

4.1 Analysis Without Clinical Variables

The comparative evaluation presented in Table 1 emphasizes the impact of harmonization methods on predictive accuracy using only PET/CT imaging as inouts, using the c-index as a measure. The baseline performance (without harmonization) was already higher than the best results of challengers during the HECKTOR 2022, with an average C-index of 0.764, associated with a relatively high variability across the 20 runs (standard deviation of 0.132). HarMSTD demonstrated the highest efficacy in harmonizing PET/CT imaging features, achieving a c-index of 0.794 ± 0.102. Note the variability was also smaller than without harmonization. HarCenter and HarCombat led to intermediate performance between no harmonization and HarMSTD, with C-index values of 0.784 and 0.773, but similar variability as without harmonization (standard deviations 0.134 and 0.138 respectively).

5 Analysis With Clinical Variables

Integrating clinical variables in addition to the PET/CT imaging features did not significantly modify the hierarchy between the compared methods, and with higher overall performance except for HarCenter, as shown in Table 2. HarMSTD still achieved the best result (0.801 ± 0.113). On the other hand, HarCenter actually decreased the performance compared to no harmonization (0.767 ± 129 vs. 0.777 ± 0.122), whereas HarCombat led to the same average but slighlty decreased variability (0.777 ± 0.105). These results suggest that clinical variables significantly contribute to predictive modeling, potentially offsetting the lack of harmonization to some extent. The findings underscore the importance of tailored harmonization approaches, especially when clinical variables are included, with HarMSTD exhibiting potential as a versatile and effective method.

5.1 Comparing to Literature

Reflecting on the harmonization-focused analysis, our pipeline demonstrates a notable improvement over literature benchmarks (referenced in Table 3), particularly with

harmonized PET/CT imaging. The best harmonized result of 0.809 (achieved with HarMSTD in the presence of clinical variables) illustrates the substantial benefits of harmonization. This performance, significantly above the challenge's top performers, highlights the efficacy of our approach, for instance the use of PCA for dimensionality reduction in combination with the DeepHitSingle model, as the baseline without harmonization was already higher than the best challenge result. The variability observed in our results emphasizes the inherent challenges in predictive modeling, further contextualizing the comparison with literature benchmarks by the constraints faced by challengers, including limited submission opportunities and time constraints, which precluded the reporting of variability across multiple runs. Furthermore, our reliance on pre-provided lesion masks for radiomic feature extraction afforded us an advantage over the challenge participants, who did not have access to these masks in their testing set. This necessitated either the generation of lesion masks or the adoption of alternative strategies, such as feature extraction from entire bounding boxes or the deployment of convolutional neural networks without segmentation. Our approach, thus, not only underscores the effectiveness of harmonization techniques in enhancing predictive accuracy but also illustrates the potential impact of methodological advantages in comparative analyses.

6 Conclusion

Our study demonstrates the efficacy of harmonization techniques, particularly HarMSTD, in improving predictive accuracy for radiomics extracted from PET/CT imaging, achieving a notable c-index up to 0.794 in analyses without clinical variables and 0.801 with them. These results highlight alternative solutions in scenarios where ComBat may not be suitable, and underline the potential of these methods in diverse contexts within oncological predictive modeling. Our future work will include investigating the robustness of the harmonization approaches when choosing a different reference center, as well as a comparison with deep learning based harmonization using image synthesis.

7 Acknowledgments

This work was supported by the HARMONY project, funded by the Cancéropôle Grand-Ouest and the regions of Pays de la Loire, Bretagne and Centre-Val de Loire (France). We thank the HECKTOR organizers for providing access to the dataset.

References

[1] V. Andrearczyk, V. Oreiller, M. Abobakr, et al. "Overview of the HECKTOR Challenge at MICCAI 2022: Automatic Head and Neck Tumor Segmentation and Outcome Prediction in PET/CT". *Head Neck Tumor Chall* 13626 (2023), pp. 1–30. DOI: 10.1007/978-3-031-27420-6_1.

[2] H. Elhalawani, A. Mohamed, A. White, et al. "Matched computed tomography segmentation and demographic data for oropharyngeal cancer radiomics challenges". *Scientific Data* 4 (1 2017). DOI: 10.1038/sdata.2017.77.

[3] B. Ma, Y. Li, H. Chu, et al. "Deep Learning and Radiomics Based PET/CT Image Feature Extraction from Auto Segmented Tumor Volumes for Recurrence-Free Survival Prediction in Oropharyngeal Cancer Patients". *Head and Neck Tumor Segmentation and Outcome Prediction*. Ed. by V. Andrearczyk, V. Oreiller, M. Hatt, et al. Cham: Springer Nature Switzerland, 2023, pp. 240–254.

[4] K. Wang, Y. Li, M. Dohopolski, et al. *Recurrence-free Survival Prediction under the Guidance of Automatic Gross Tumor Volume Segmentation for Head and Neck Cancers*. 2022.

[5] M. Meng, L. Bi, D. Feng, et al. "Radiomics-Enhanced Deep Multitask Learning for Outcome Prediction in Head and Neck Cancer". *Head and Neck Tumor Segmentation and Outcome Prediction*. Ed. by V. Andrearczyk, V. Oreiller, M. Hatt, et al. Cham: Springer Nature Switzerland, 2023, pp. 135–143.

[6] L. Rebaud, T. Escobar, F. Khalid, et al. "Simplicity Is All You Need: Out-of-the-Box nnUNet Followed by Binary-Weighted Radiomic Model for Segmentation and Outcome Prediction in Head and Neck PET/CT". *Head and Neck Tumor Segmentation and Outcome Prediction*. Ed. by V. Andrearczyk, V. Oreiller, M. Hatt, et al. Cham: Springer Nature Switzerland, 2023, pp. 121–134.

[7] Á. V. Juanco-Müller, J. F. C. Mota, K. Goatman, et al. "Towards Tumour Graph Learning for Survival Prediction in Head amp; Neck Cancer Patients". *Head and Neck Tumor Segmentation and Outcome Prediction*. Springer Nature Switzerland, 2023, pp. 178–191. DOI: 10.1007/978-3-031-27420-6_18.

[8] Y. Wang, E. Lombardo, L. Huang, et al. "Head and Neck Cancer Localization with Retina Unet for Automated Segmentation and Time-To-Event Prognosis from PET/CT Images". *Head and Neck Tumor Segmentation and Outcome Prediction*. Ed. by V. Andrearczyk, V. Oreiller, M. Hatt, et al. Cham: Springer Nature Switzerland, 2023, pp. 202–211.

[9] V. Thambawita, A. M. Storås, S. A. Hicks, et al. *MLC at HECKTOR 2022: The Effect and Importance of Training Data when Analyzing Cases of Head and Neck Tumors using Machine Learning*. 2022.

[10] Q. Lyu. "Combining nnUNet and AutoML for Automatic Head and Neck Tumor Segmentation and Recurrence-Free Survival Prediction in PET/CT Images". *Head and Neck Tumor Segmentation and Outcome Prediction*. Ed. by V. Andrearczyk, V. Oreiller, M. Hatt, et al. Cham: Springer Nature Switzerland, 2023, pp. 192–201.

[11] Z. Salahuddin, Y. Chen, X. Zhong, et al. "HNT-AI: An Automatic Segmentation Framework for Head and Neck Primary Tumors and Lymph Nodes in FDG- PET/CT Images". *Head and Neck Tumor Segmentation and Outcome Prediction*. Ed. by V. Andrearczyk, V. Oreiller, M. Hatt, et al. Cham: Springer Nature Switzerland, 2023, pp. 212–220.

Can bootstrapping increase robustness of voxel-based-analysis?

Alan McWilliam[1,2], Marcel van Herk[1,2], Chloe Pratt[3] and Eliana Vasquez Osorio[1,2]

[1]Division of Cancer Sciences, University of Manchester, UK
[2]The Christie NHS Foundation Trust, Manchester, UK
[3]Department of Medical Physics, Leeds Cancer Centre, St James's University Hospitals, UK

Abstract: Bootstrapping is a well-established resampling method in statistics, allowing measures of accuracy to be assigned. Using bootstrapping in a voxel-based-analysis (VBA) workflow has potential to increase robustness for identifying anatomical regions where dose is significantly associated with outcome. In this work, we implement bootstrapped VBA, and assessed its impact across progressively smaller datasets.

1101 Non-Small-Cell Lung Cancer patients were available. Bootstrapping was implemented, performing 100 bootstraps of the full VBA analysis performing a t-test per voxel against survival at 12-months. Results demonstrate that bootstrapped VBA defines a smaller, more targeted region in the base of the heart. Dose extracted from the bootstrapped VBA region showed improved model performance in multivariable models compared to the standard VBA workflow. VBA showed utility in progressively smaller dataset sizes up to 250 entries. Datasets of 100 and 50 showed high uncertainty in defining significance. Bootstrapped VBA emerges as a valuable internal validation approach.

1 Introduction

Bootstrapping is a resampling method often applied to complex sampling designs and deemed to be an appropriate way to control and check the stability of the results [1]. Bootstrapping uses random sampling with replacement. This technique constructs new datasets by sampling the complete dataset randomly, allowing the same element to appear multiple times. This allows assigning a measure of accuracy to model building and provides confidence interval estimates of the parameters.

Bootstrapping is a well-established statistical technique but, as yet, its benefit in voxel-based-analysis (VBA) has not been fully explored. VBA is becoming an established spatial outcome modelling technique which incorporates the full, 3-dimensional, dose distribution within a dose-response modelling framework [2]. VBA is desirable as it (1) removes the need for prior assumptions on anatomy to consider; (2) removes the need to use summary statistics for radiotherapy dose, which often assume an organ to have a homogeneous dose-response. Instead, the full dose distribution is analysed on a voxel level, performing a statistical test per-voxel to identify anatomical regions with significant association between dose and outcome. VBA has been applied in numerous studies and results are now entering clinical practice [3,4].

In this work we investigate the potential of bootstrapping data for VBA. This work aims to show whether bootstrapping can be used as an internal validation approach to assess robustness of VBA and to explore the effect of dataset size.

2 Materials and Methods

1101 Non-Small-Cell Lung Cancer (NSCLC) patients were available from a previous data analysis [5], with institutional approval in place to analyse the data. First, bootstrapping with replacement was implemented in the data mining workflow. A binary endpoint of survival at 12-months was used with a t-test per voxel performed. To account for multiple comparisons, permutation testing was implemented: outcomes labels were permuted 100 times and the max t-statistic across the image calculated to summarise each permutation. From the distribution of the max t-statistics the 95th percentile threshold was selected as the significance value. This value is then used to threshold the real t-map to identify anatomy associated with the outcome. 100 bootstraps were performed, and for each, the full VBA workflow was performed, including permutation testing.

For each bootstrap, a t-map and threshold value is available. The significant region from each bootstrap was threshold as a binary mask and summed. This provided a map of the number of times a given voxel was included in the significant region. From this map, the 50th percentile was selected to define the anatomical region at the 95th percentile confidence interval.

Comparison with the standard workflow without bootstrapping was performed. First the location of the identified region was compared. Second, the dose to the standard and bootstrapped regions was extracted for all patients and evaluated in a multivariable Cox-proportional hazards model including known prognostic factors (sex, age, T-stage, N-Stage and performance status). We used the Akaike information coefficient (AIC) to compare model, where lower AIC indicates models with better discrimination power and a reduction of at least 2 points is considered significant [6].

Next, experiments were performed to establish if bootstrapped VBA has utility to improve performance with smaller datasets. New datasets with 500, 250, 100 and 50 entries were created by randomly sampling from the full dataset. Sampling was nested, i.e., the 100 entries were in the 250 entries which were in turn in the 500 entries dataset. Bootstrapped VBA was performed as described with the final region identification compared.

Finally, to further test the utility in small datasets a simple simulation workflow was implemented following the described workflow. A single voxel was simulated to represent a significant region of anatomy in real data. Simulations identified if there was a significant difference with outcome in the dummy data and tested for robustness across the bootstrapping, i.e., if each bootstrap identified or miss-labelled the significance. The simulation created progressively smaller datasets as described for the real data, with 1000, 500, 250, 100, 50 samples used.

To feed the simulation, data was generated from the identified region from the 1101 NSCLC patients and outcome survival at 12-months. This allowed testing with real data ranges. Simulations were run 100 times and summary statistics calculated to describe robustness of bootstrapped VBA.

3 Results

Figure 1 includes the results from the standard and bootstrapped VBA workflows, including the 95th percentile confidence intervals for the identified region using the bootstrapped data. The regions overlap, the standard workflow provides a region of 35.4cm^3 and the bootstrapped VBA a smaller region of 13.2cm^3. Across all bootstraps, the region volume showed a standard deviation of 30.1cm^3 suggesting reasonable agreement across all bootstraps. The 95th percentile confidence intervals show greater difference in the cranial-caudal direction, demonstrating the greatest uncertainty in the result is in this direction.

The survival models included in table 1 show results for both the standard and the bootstrapped VBA workflows. The variables show comparable hazards across all variables. The dose from the bootstrapped VBA workflow showed greater significance compared to the dose from the standard workflow. The AIC significantly reduced from 8033 to 8026 indicating improved model performance using the bootstrapped VBA region.

N	Mean volume (range) (cm3)	Standard Deviation of region volumes	Bootstraps with no significance
1000	31.2 (0 – 142)	31	1
500	31.7 (0 – 286)	36	6
250	24.0 (0 – 117)	28.8	16
100	22.8 (0 – 301)	47.3	38
50	72.4 (0 – 1276)	176	33

Table 2. Summary statistics from the experiments reducing the size of the datasets. Good consistency in region volumes are identified to 250 samples. 100 and 50 dataset sizes show greater instability with increasing standard deviations and greater number of bootstraps with no significance found.

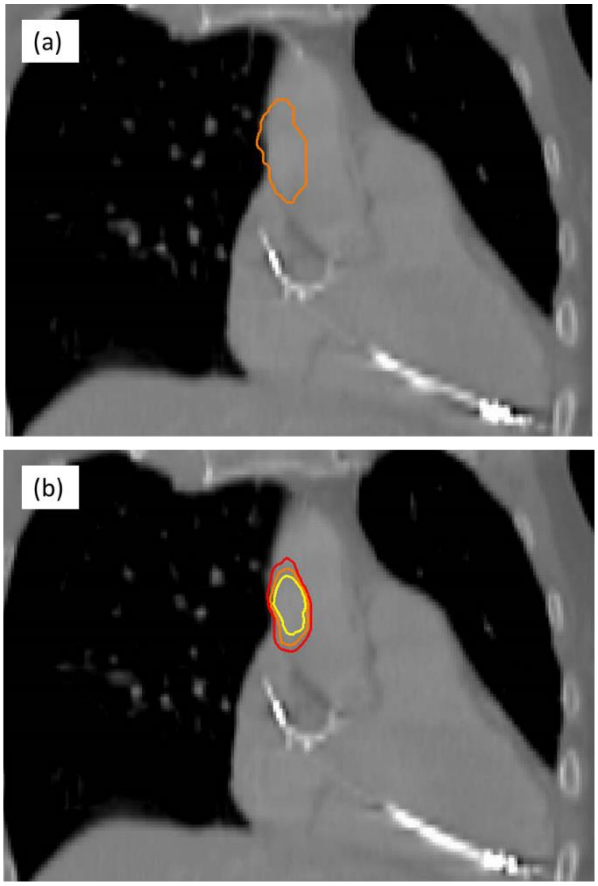


Figure 1. coronal views of the (a) the 95th percentile threshold (orange) using the standard VBA workflow and (b) the 95th percentile threshold (orange) with the 95th percentile confidence intervals (red and yellow). The bootstrapped VBA region is smaller than the standard workflow, the confidence intervals suggest greater uncertainty in the cranial – caudal direction.

Results from investigating progressively smaller datasets are summarised in table 2. Results demonstrate consistent volumes and standard deviations of the identified region volume across all bootstraps for dataset sizes of 1000, 500 and 250. For sample sizes of 100 and 50 datasets greater instability is demonstrated with increasing standard deviation and a large number of bootstraps where no significant region was identified. Visually, consistent regions in the base of the heart were found for 100, 500 and 250 samples; 100 and 50 show little consistency with a number of small regions throughout the patients anatomy.

Finally, since the VBA bootstrapping is computationally intensive, simulations were performed to ensure the results generated in table 2 were representative. The dose extracted from the defined bootstrapped region in the full dataset (figure 1b) is used. Table 3 presents the summary statistics from the 100 simulations for each dataset size. For the 1000, 500 and 250 dataset sizes, the mean number of bootstraps correctly defining significance was 100, 99 and 94 respectively across all simulations. The minimum

Variable		Standard VBA		Bootstrapped VBA	
		Hazard ratio (confidence interval)	p-value	Hazard ratio (confidence interval)	p-value
Age *(years)*		1.02 (1.01 – 1.03)	<0.001	1.02 (1.01 – 1.03)	<0.001
Sex *(comparator male)*		1.30 (1.14 – 1.52)	<0.001	1.29 (1.10 – 1.50)	<0.001
T-stage *(comparator T-stage 1)*	2	1.29 (1.01 – 1.66)	0.04	1.26 (0.98 – 1.62)	0.07
	3	1.58 (1.21 – 2.06)	<0.001	1.54 (1.18 – 2.01)	0.001
	4	1.54 (1.15 – 2.07)	0.003	1.56 (1.16 – 2.01)	0.003
N-Stage *(comparator N-stage 0)*	1	0.86 (0.67 – 1.10)	0.22	0.88 (0.69 – 1.12)	0.28
	2	1.37 (1.11 -1.69)	0.003	1.35 (1.09 – 1.67)	0.005
	3	1.47 (1.07 – 2.02)	0.02	1.47 (1.07 – 2.03)	0.02
Performance status *(comparator PS 0)*	1	1.22 (0.95 – 1.56)	0.12	1.21 (0.95 – 1.56)	0.13
	2	1.39 (1.08 – 1.79)	0.01	1.44 (1.11 – 1.86)	0.01
	3	1.60 (1.11 – 2.31)	0.01	1.59 (1.10 – 2.31)	0.01
Dose to region *(Gy)*		1.01 (1.01 – 1.02)	0.002	1.01 (1.01 – 1.02)	0.001

Table 1. Results from the Cox-proportional hazards model using the dose extracted from the anatomical region defined by the standard VBA and bootstrapped VBA workflows. All variables show comparable hazards in both models. The dose extracted from the bootstrapped VBA region shows a greater significance and the AIC of the model is significantly reduced, indicated a better model performance.

for a simulation correctly defining significance in the 250 dataset was 53 out of 100 bootstraps, therefore, there were sets of simulated data showing large uncertainty and others showing good agreement across bootstraps. For the 100 and 50 dataset sizes a high degree of uncertainty was seen across all simulation runs with many not identifying any significance. The minimum and maximum numbers of bootstraps correctly identifying significance within each simulations run covers the full range, demonstrating a high sensitivity to patients sampled. These simulation results agree with the analysis of the clinical data (figure 1 and table 2) for our exemplar. Dataset sizes of 250 provide good region stability with low variation on the identified region volume and small number of bootstraps not correctly defining significance. Dataset sizes of 100 and 50 showed higher instability to patients sampled and bootstrapping the data does not improve stability.

4 Discussion

In this abstract, we have investigated the role of bootstrapping for region identification in VBA, assessing its impact in small datasets. In our exemplar dataset, patients with NSCLC treated with radiotherapy, bootstrapped VBA defined a smaller region in the base of the heart where dose was significantly associated with survival. Dose extracted from this region demonstrated improved model performance, suggesting this is a more targeted and clinically important anatomical location. Bootstrapping provided robust region detection for datasets of 250 or more entries, both in the clinical data and in simulation experiments.

Bootstrapping is well used across statistics and recommended in TRIPOD guidelines as an internal validation approach [7]. The implementation with VBA also acts as internal validation that the region detected is reproducible. Additionally, it can provide confidence intervals or the region allowing visualisation of the modelling uncertainties.

N	Correct classification Mean (SD) range	No significance found Mean (SD) range	Incorrect classification Mean (SD) range
1000	100 (0) 100 - 100	0	0
500	99 (0.7) 0 - 5	0.3 (0.7) 0 - 5	0
250	94 (8.2) 53 - 100	6 (8.1) 0 - 45	0.1 (0.2) 0 - 2
100	66.5 (24.7) 3 - 100	32.5 (23.3) 0 - 79	1 (2.7) 0 - 22
50	50.2 (30.8) 0 - 100	45.2 (25.6) 0 - 100	4.6 (10.91) 0 - 60

Table 3. Summary statistics from the simulation experiments. For each dataset size, 100 simulations were performed. The mean number of bootstraps showing correct significance, no significance or incorrect significance across the simulations are included with the standard deviation. The range is also included showing the maximum and minimum times each category was selected across the simulations.

Further work is needed to fully explore the role of bootstrapped VBA across different treatment sites, clinical endpoints and statistical analysis. However, our exemplar shows promise of benefit. Here, we analysed a t-test with a binary outcome variable. In the future, this work should be extended to explore utility in other statistical tests, for example, continuous outcomes analysed voxel-wise with a Pearson or Spearman correlation [8], or multivariable approaches like linear regressions or cox-proportional hazards per voxel [9].

The simulation approach implemented showed agreement with the clinical dataset providing confidence in the results. The simulation can be further improved to allow user defined dose functions. These would allow the dose separation between outcome groups to be controlled and allow detailed exploration of bootstrapped VBA's power to resolve and identify these effects. Simulations can be extended to include correlations and multivariable models.

5 Conclusion

In conclusion, bootstrapping has shown utility as a form of internal validation within VBA. In our exemplar application, the defined significance region across all bootstraps showed improved model performance, indicating a more clinically relevant anatomical region was identified. Further clinical sites and endpoints should be investigated. If results hold, bootstrapped VBA should be recommended as best practice.

References

[1] Thomas J. DiCiccio, Bradley Efron "Bootstrap confidence intervals". Statist. Sci. 11.3 (1996) DOI: 10.1214/ss/1032280214.; [2] Palma G, Monti S, Cella L. "Voxel-based analysis in radiation oncology: A methodological cookbook". Phys Med. 69:192-204 (2020) DOI: 10.1016/j.ejmp.2019.12.013.; [3] McWilliam A, Palma G, Abravan A, et al. "Voxel-based analysis: Roadmap for clinical translation". Radiother Oncol. 188 (2023). DOI: 10.1016/j.radonc.2023.109868.; [4] Price G, Devaney S, French DP, et al. "Can Real-world Data and Rapid Learning Drive Improvements in Lung Cancer Survival? The RAPID-RT Study". Clin Oncol.34(6) (2022). DOI: 10.1016/j.clon.2021.12.017.; [5] McWilliam A, Kennedy J, Hodgson C, et al. "Radiation dose to heart base linked with poorer survival in lung cancer patients". Eur J Cancer. 85 (2017). DOI: 10.1016/j.ejca.2017.07.053.; [6] Sutherland C, Hare D, Johnson PJ, et al. "Practical advice on variable selection and reporting using Akaike information criterion". Proc Biol Sci. 290 (2007). DOI: 10.1098/rspb.2023.1261.; [7] Collins GS, Reitsma JB, Altman DG, Moons KG. "Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement". BMJ. 350 (2015). DOI: 10.1136/bmj.g7594.; [8] Beasley W, Thor M, McWilliam A, et al. "Image-based Data Mining to Probe Dosimetric Correlates of Radiation-induced Trismus". Int J Radiat Oncol Biol Phys. 102.4 (2018). DOI: 10.1016/j.ijrobp.2018.05.054.; [9] Green A, Vasquez Osorio E, Aznar MC, et al. "Image Based Data Mining Using Per-voxel Cox Regression". Front Oncol. (2020) DOI: 10.3389/fonc.2020.01178.

Trading Robustness? Introducing total variance objectives to Pareto surface approximation techniques for radiotherapy

Remo Cristoforetti[1,2,3], Philipp Süss[4], and Niklas Wahl[1,2]

[1]Department of Medical Physics in Radiation Oncology, German Cancer Research Center (DKFZ), Heidelberg, Germany
[2]National Center for Radiation Research in Oncology (NCRO), Heidelberg Institute for Radiation Oncology (HIRO), Heidelberg, Germany
[3]Department of Physics and Astronomy, Heidelberg University, Heidelberg, Germany
[4]Optimization division, Fraunhofer Institute for Industrial Mathematics (ITWM), Kaiserslautern, Germany

Abstract This work introduces a novel approach to introduce robustness into multicriteria optimization (MCO) in proton therapy. Our approach exploits a formulation of expected value optimization relying on precomputed total variance and expected dose influence. This facilitates merging robust optimization with Pareto front approximation while avoiding explicit calculation of scenarios during optimization. The method also introduces robustness as an objective, enabling interactive management of robustness trade-offs.
We demonstrate MCO including robustness trading with a single field proton plan on the TG119 phantom and a two field proton plan on a liver case. Trade-offs between robustness and dose metrics can be effectively approximated and visualized. Our approach allows planners more explicit control over robustness, and is directly translatable to other MCO approaches like lexicographic optimization.

1 Introduction

Radiotherapy treatment planning is most commonly modeled as a multicriteria optimization (MCO) problem, as trade-offs between multiple organs' toxicities and target coverage need to be managed. Approaches to solve this MCO problem have evolved from the manual weighted-sum approach, which involves repetitive manual tweaking of objective weights in consecutive optimization runs, to Pareto surface approximation techniques or automated lexicographic solvers.
In parallel, techniques like robust optimization were introduced to personalize the management of treatment uncertainties, particularly in particle therapy. In robust optimization, uncertainties are explicitly modeled through computation of error scenarios. These are then aggregated during optimization to minimize a robustness metric of the objective function, like its maximum over all scenarios or its expected value.
While the combination of robust optimization and modern MCO approaches has been suggested before [1], merging both is often limited by incompatible problem formulations or combination of both methods' increased computational demands. For MCO, computational complexity depends on the number of objectives. Performance of robust optimization depends on the number of error scenarios used during optimization. Introducing new robustness objectives computed from a large number of error scenarios to MCO is therefore limited by computational resources.
To address this challenge, we introduce a novel approach that incorporates total variance objectives into Pareto surface approximation techniques. Our method stands out as it formulates robustness objectives whose evaluation cost during optimization is independent of the number of the calculated error scenarios, thereby significantly enhancing computational efficiency without compromising on the robustness of the treatment plans. This independence is achieved by precomputing influence matrices of expected dose and total variance. Then, during approximation of the Pareto front, scenarios can be discarded or stored on disk and only these lightweight matrices are needed. For visualization or navigation of the solutions, the scenarios can be reloaded and the robustness trade-off can be visualized.
We presents the theoretical foundation of our method, the resulting algorithm and their implementation, and demonstrate the possible trade-offs in planning problems.

2 Materials and Methods

Our approach exploits a specific form of robust treatment planning, called probabilistic optimization or expected value optimization [2, 3], and combines it with Pareto surface approximation and navigation techniques.

2.1 Probabilistic Optimization

A common formulation of a probabilistic optimization problem would minimize the expected value $\mathbb{E}[f(d(x)]$ of an objective function f depending on beamlet intensities x [3]:

$$x^* = \arg\min_x \mathbb{E}[f(d(x))]\,. \tag{1}$$

The objective function f is usually formulated depending on dose d and chained to the beamlet intensities x by precomputation of a dose influence matrix D such that $d = Dx$.
For practical purposes, $\mathbb{E}[f]$ in eq. (1) is computed by averaging over a limited number of S samples from the assigned probability distributions. This, in turn, requires computation of S dose influence matrix samples D_s to be able to compute the dose during each iteration of an optimization run. This means that during optimization, the S samples of the dose influence matrix need to be kept in memory while keeping S as small as possible to guarantee computationally efficient treatment plan optimization.

2.2 Total Variance and expected dose influence

For the penalized least-squares objective function $f^{lsq}(d) = (d-d^{ref})^T P(d-d^{ref})$, where d^{ref} is the prescription dose and P a diagonal matrix encoding voxel penalty factors, it was previously shown [4] that its expectation can be separated into a total variance term (written as the trace over the doses' covariance matrix Σ) and an expected value term:

$$\mathbb{E}[f^{lsq}(d)] = \mathrm{tr}(P\Sigma) + (\mathbb{E}[d]-d^{ref})^T P(\mathbb{E}[d]-d^{ref}) \quad (2)$$
$$= x^T\Omega x + \mathbb{E}[D]x - d^{ref})^T P(\mathbb{E}[D]x - d^{ref}) \; . \quad (3)$$

Equation (3) introduces the "total variance" influence matrix Ω together with the expected dose influence matrix $\mathbb{E}[D]$. While the latter can easily be estimated by taking the elementwise mean over all D_s, computing Ω requires estimating the expected value of elementwise combinations of D via

$$\mathrm{tr}(P\Sigma) = \sum_{jk} x_j x_k \underbrace{\sum_i p_i \left(\mathbb{E}\left[D_{ij}D_{ik}\right] - \mathbb{E}\left[D_{ij}\right]\mathbb{E}\left[D_{ik}\right]\right)}_{\Omega_{jk}} \; , \quad (4)$$

with voxel penalties p_i in voxel i and intensity vector x.
Under the assumption that penalties are equal in certain voxels, i. e., voxels i belonging to the same volume of interest (VOI) V, corresponding Ω_V can be computed after pulling the penalty out of the sum, simplifying to

$$\Omega_V = \mathbb{E}[D_V^T D_V] - \mathbb{E}[D_V]^T\mathbb{E}[D_V] \; , \quad (5)$$

where D_V are the dose influence matrices for V.
Since Ω_V and $\mathbb{E}[D]$ can be precomputed during scenario dose calculation, we can now build objective functions depending on the total variance $x^T\Omega_V x$ or the expected dose $\mathbb{E}[D]x$ without requiring storage of scenarios during optimization. As $\mathbb{E}[D]$ and D have similar structure and Ω is a symmetric square matrix with number of rows/columns equal to the number of beamlets, optimization performance exhibits minimal overhead and is independent to the number of error scenarios used in the precomputation of Ω_V's and $\mathbb{E}[D]x$.

2.3 Multicriteria formulation and optimization

In general, an MCO problem is of the form

$$\text{"min"}_{x\in\mathscr{X}} \left[f_1(x), \ldots, f_k(x)\right] \quad (6)$$

where f_1 to f_k are objective functions for the treatment planning problem which should ideally all be minimized simultaneously by x. Hard constraints that must all hold (e.g. non-negativity of x) define the set $\mathscr{X}$. It is typically true that there exists no $x \in \mathscr{X}$ such that all objective function attain their individual minimum. Interesting solutions to (6) have the property that if one of the individual objectives should be further improved, at least one other objectives would necessarily have to degrade. This is often referred to as "efficiency" or "Pareto-efficiency" (after Vilfredo Pareto, an Italian economist). We label the set of all such efficient treatment plans the *Pareto front* (in some publications, this set is also referred to as the *Pareto surface*).
We formulate the MCO problem in this work by defining the following objectives for the planning structures. We distinguish between risk structures for which we want to analyze the trade-offs against all other objectives, and those for which we just want the dose to be within certain bounds.

- for a target structure $\mathscr{T}$, we define
  1. an objective $f^{lsq}_{\mathscr{T}}(\mathbb{E}[d_{\mathscr{T}}(x)])$ as defined above to minimize the squared deviation of expected dose from a reference dose (normalized by $\frac{1}{|\mathscr{T}|}$), and
  2. the total variance objective $f^{\Omega}_{\mathscr{T}}(x) := \frac{1}{|\mathscr{T}|}x^T\Omega_{\mathscr{T}}x$
- for risk structures $\mathscr{R}$ which we want to include in a MCO trade-off analysis, we define
  1. the 2-EUD as objective: $f^{EUD}_{\mathscr{R}}(\mathbb{E}[d(x)]) = \frac{1}{|\mathscr{R}|}\mathbb{E}[d]^T_{\mathscr{R}}\mathbb{E}[d]_{\mathscr{R}}$, and
  2. the total variance objective $f^{\Omega}_{\mathscr{R}}(x) := \frac{1}{|\mathscr{R}|}x^T\Omega_{\mathscr{R}}x$

Aside from the physical requirement $x \geq 0$, $\mathscr{X}$ also contains hard constraints on the over- and underdosing of target structures. We found that the constraints $g^{UD}_{\mathscr{T}}(\mathbb{E}[d_{\mathscr{T}}(x)]) := \frac{1}{|\mathscr{T}|}(d^{ref}_{\mathscr{T}} - \mathbb{E}[d_{\mathscr{T}}(x)])_+ \leq 1$ and $g^{OD}_{\mathscr{T}}(\mathbb{E}[d_{\mathscr{T}}(x)]) := \frac{1}{|\mathscr{T}|}(\mathbb{E}[d_{\mathscr{T}}(x)] - d^{ref}_{\mathscr{T}})_+ \leq 0.5$ work well for each involved target T. Here, $(\cdot)_+ := max(0,\cdot)$. Finally, for each risk structure not included in the MCO analysis, we add an overdose constraint $g^{OD}_{\mathscr{R}}(\mathbb{E}[d_{\mathscr{R}}(x)]) \leq 2$ with d^{ref} at half of the main target prescription dose.
Since all our objectives and constraints are convex functions, the Pareto front will be the surface of a k-dimensional convex body and we may directly apply well established methods *sandwiching methods* [5–8] to approximate this set. These methods iteratively produce efficient treatment plans and stop when the approximation quality is good enough. Details can be found in the references given above.

2.4 Implementation and Software

Computation of $\mathbb{E}[D]$ and Ω as well as the robustness analyses were done in matRad [9]. We use an in-house sandwiching method implemented by the Fraunhofer ITWM to approximate the Pareto fronts. Objectives are scaled to the [0,1] interval to prevent numerical distortions due to differing value domains. We set the termination criterion to stop the approximation when the ε-indicator [5–8] for the scaled objectives is below 0.05. For each case, we state the number of computed solutions to reach the required quality. This number corresponds to the number of plan optimizations. We use Ipopt [10] and KNITRO [11] to do the numerical optimization of the resulting scalar optimizations.

3 Results

3.1 TG-119 C-Shape

For a first low-dimensional demonstration, we used the TG119 phantom (directly available in matRad), a single proton field was set to cover the C-shaped Target and expose the distal Core structure to range uncertainties.
Setup uncertainties of $\pm 2.25\,\text{mm}$ and range uncertainties of $\pm 3.5\,\% \pm 1\,\text{mm}$ were assumed as standard deviation to sample 10 error scenarios from the respective multivariate normal distribution. These 10 scenarios were used to estimate $\mathbb{E}[D]$ and Ω before optimization.
Objectives and constraints were chosen according to section 2.3, resulting in a 5D MCO problem requiring 40 solutions for an acceptable Pareto surface approximation. For all solutions, robustness analysis was performed with DVHs and Standard-deviation-volume-histograms (SDVHs). SDVHs relate dose standard deviation to percentage volume for a given VOI to quantify the VOI-specific robustness. We illustrate the observed trade-off within the Core OAR in Figure 1.

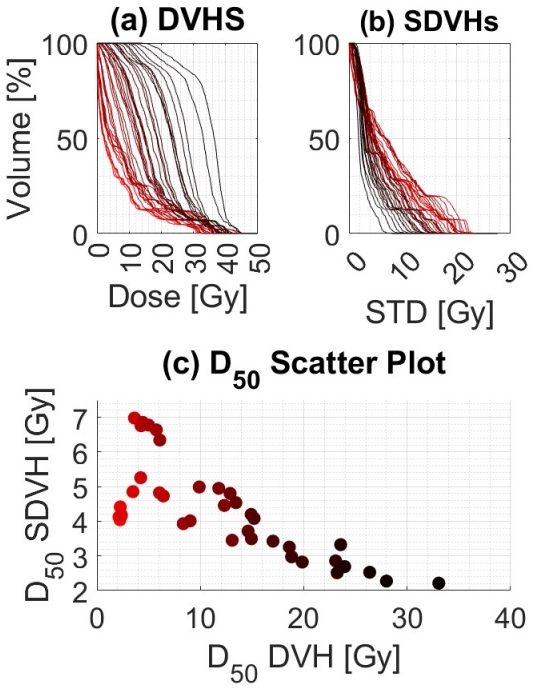


Figure 1: DVHs (a) and SDVHs (b) for all 40 solutions in the Core. (c) Scatter plot relates median dose and standard deviation (D_{50}). Similar colors correspond to the same solution in (a-c).

Figure 1 shows that the exposition of the Core to range uncertainties creates a direct trade-off between the Core robustness (total variance objective) and the dosage (EUD objective). With MCO, this trade-off can now be directly explored with Pareto surface navigation.

3.2 Liver case

To demonstrate our approach on a high-dimensional patient case, we planned a two-field proton plan for the Liver case available in matRad. One of the fields was chosen to expose the heart to range uncertainties. Similar uncertainty assumptions as in section 3.1 were applied using the CTV as target and liver and heart as OAR. A larger set of 100 scenarios was sampled to construct $\mathbb{E}[D]$ and Ω to demonstrate that our approach works with commonly unsuitably large numbers of scenarios. For optimization, again the protocol defined in section 2.3 was applied, resulting in a 7D MCO problem requiring 60 Pareto-optimal solutions. For all solutions, robustness analysis was again performed reloading the 100 scenarios initially used to construct $\mathbb{E}[D]$ and Ω.
Figure 2 illustrates the 7D MCO solution with navigation interface allowing trading robustness and dosimetric obejctives.

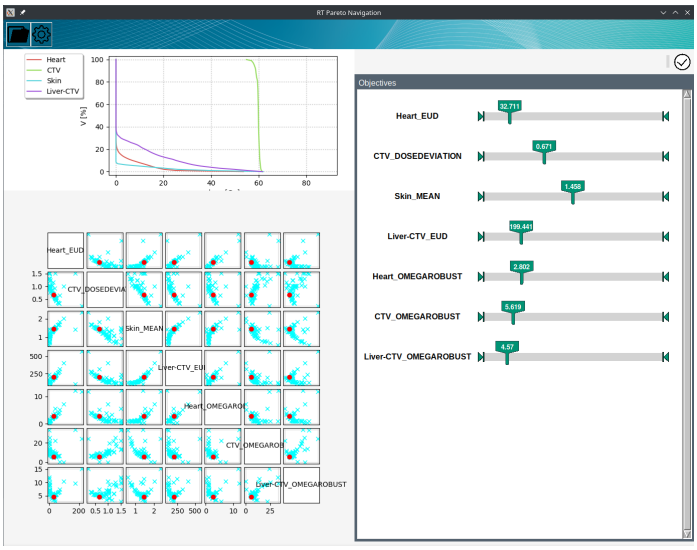


Figure 2: Pareto navigation for the liver case. Each slider on the right corresponds to an objective function. To improve one objective, the slider has to be pulled to the left. The bottom left is a scatterplot matrix of all pairs of objectives, showing the 2D projection of the Pareto front (in turquoise) as well as the currently navigated point (red). All elements then update in real time.

Figure 3 shows two exemplary solutions illustrating how heart robustness can be traded against CTV robustness and coverage; improved heart robustness will downregulate the proton beam exposing the heart to range uncertainties.
Figure 4 adds DVHs for the CTV and SDVHs for the Heart to the analysis and illustrates the correlation of target coverage with heart robustness, indicating a clear trade-off between the two that can be explored by Pareto navigation.

4 Discussion

One of the main clinical benefits of the presented methods is difficult to convey on paper: robustness can now be easily included in the interactive planning process. A common stage in modern multicriteria radiotherapy treatment planning is the interactive *Pareto Navigation* [12, 13] as implemented in Varian's Eclipse Treatment System and in RaySearch Laboratories' RayStation. This decision support conveys the shape

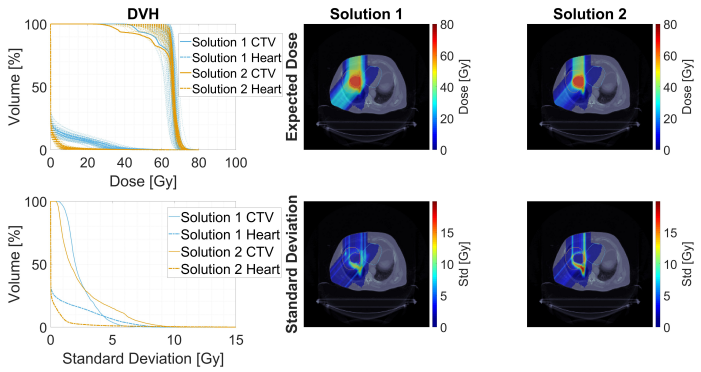


Figure 3: On the left, DVHs and SDVH for the CTV and Heart structures. Histograms are reported for all the 100 error scenarios nad for two solutions on the Pareto-front. On the right, expected dose and SD distributions for the two solutions are compared.

of a high-dimensional Pareto front and its implied trade-offs between the objectives to the decision-maker. It lets the planner explore the chances and limitations of the treatment by interactively manipulating the plan quality.
By adding the total variance objectives to the multicriteria formulation, the planner can interactively trade off the expected plan quality against the robustness on a structure level. We hope that this interactive analysis will improve the planner's and treating physician's confidence to apply robustness to their treatment plans.

5 Conclusion

We successfully introduced robustness as an objective into MCO for proton therapy. The approach limits computational overhead by precomputing expected dose and total variance influence before optimization, avoiding the need for explicit scenario evaluation. Trade-offs between robustness and dose metrics could be effectively calculated and visualized using Pareto surface approximation techniques.

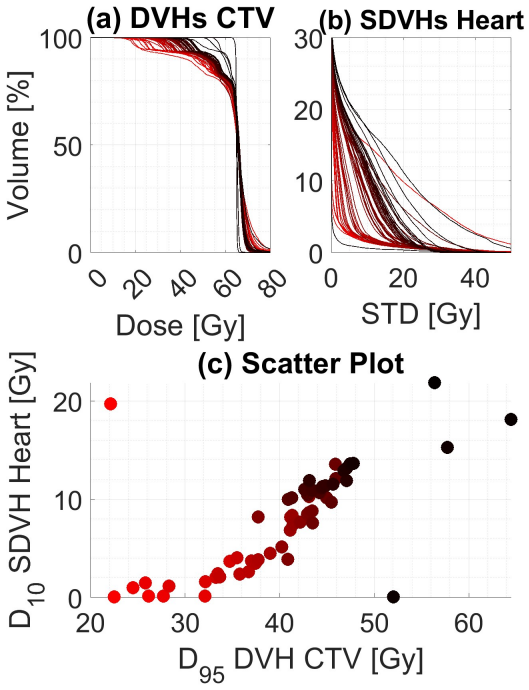


Figure 4: (a) CTV's DVHs and (b) Heart's SDVHs for all the obtained solution. (c) Scatter plot of the CTV's DVHs D_{95} and the Heart's SDVHs D_{10} points.

References

[1] W. Chen, J. Unkelbach, A. Trofimov, et al. "Including robustness in multi-criteria optimization for intensity-modulated proton therapy." *Physics in Medicine & Biology* 57.3 (2012), p. 591. DOI: 10.1088/0031-9155/57/3/591.

[2] J. Unkelbach, M. Alber, M. Bangert, et al. "Robust radiotherapy planning." *Physics in Medicine & Biology* 63.22 (2018), 22TR02. DOI: 10.1088/1361-6560/aae659.

[3] A. Fredriksson. "A characterization of robust radiation therapy treatment planning methods—from expected value to worst case optimization." *Medical Physics* 39.8 (2012), pp. 5169–5181. DOI: https://doi.org/10.1118/1.4737113.

[4] M. Bangert, P. Hennig, and U. Oelfke. "Analytical probabilistic modeling for radiation therapy treatment planning." *Physics in Medicine & Biology* 58.16 (2013), p. 5401. DOI: 10.1088/0031-9155/58/16/5401.

[5] K. Klamroth, J. Tind, and M. M. Wiecek. "Unbiased approximation in multicriteria optimization." *Mathematical Methods of Operations Research* 56 (2003), pp. 413–437. DOI: 10.1007/s001860200217.

[6] J. I. S. Hernandez. "Multi-objective optimization in Mixed Integer Problems with application to the Beam Selection Optimization Problem in IMRT." PhD thesis. Technical University of Kaiserslautern, 2011.

[7] R. Bokrantz and A. Forsgren. "An algorithm for approximating convex pareto surfaces based on dual techniques." *INFORMS Journal on Computing* 25 (2013), pp. 377–393. DOI: 10.1287/ijoc.1120.0508.

[8] I. Lammel. "Approximation of Pareto Surfaces in Multi-Criteria Optimization." PhD thesis. RPTU Kaiserslautern, 2023.

[9] H.-P. Wieser, E. Cisternas, N. Wahl, et al. "Development of the Open-Source Dose Calculation and Optimization Toolkit matRad." *Medical Physics* 44.6 (2017), pp. 2556–2568. DOI: 10.1002/mp.12251.

[10] A. Wächter and L. T. Biegler. "On the Implementation of a Primal-Dual Interior Point Filter Line Search Algorithm for Large-Scale Nonlinear Programming." *Mathematical Programming* 106 (2006), pp. 25–57. DOI: 10.1007/s10107-004-0559-y.

[11] R. H. Byrd, J. Nocedal, and R. Waltz. "KNITRO: An integrated package for nonlinear optimization." *Large-Scale Nonlinear Optimization*. Springer, 2006, pp. 35–59. DOI: 10.1007/0-387-30065-1_4.

[12] M. Monz, K.-H. Küfer, T. R. Bortfeld, et al. "Pareto navigation—algorithmic foundation of interactive multi-criteria IMRT planning." *Physics in Medicine and Biology* 53 (2008), pp. 985–998. DOI: 10.1088/0031-9155/53/4/011.

[13] C. Collicott, E. Bonacker, I. Lammel, et al. "Interactive navigation of multiple convex patches." *Journal of Multi-Criteria Decision Analysis* 28.5-6 (2021), pp. 311–321. DOI: 10.1002/mcda.1768.

Mid-Range Proton Arc Therapy

Qingying Wang[1], Mingli Chen[1], Mahdieh Kazemimoghadam[1], Xuejun Gu[*1, 2] and Weiguo Lu[*1]

[1]Department of Radiation Oncology, The University of Texas Southwestern Medical Center, Dallas, TX 75390, United States of America
[2]Department of Radiation Oncology, Stanford University, Stanford, CA 94305, United States of America
[*]Co-corresponding authors: xuejungu@stanford.edu, weiguo.lu@utsouthwestern.edu

Abstract Proton arc therapy (PAT) holds promise for enhancing target coverage and sparing normal tissues by incorporating all beam angles in treatment plan optimization. However, the conventional use of all possible energy layers for each angle significantly impacts delivery efficiency and increases planning complexity as treatment delivery time is heavily extended by energy switch and planning parameters are greatly amplified by including energy selection optimization. In this study, we propose a novel Mid-Range Proton Arc Therapy (MRPAT) framework that utilizes the mid-range beam, employing a single pre-selected energy layer for delivering mid-range spots at each beam angle in PAT and adopts a novel spot optimization strategy to efficiently solve large-scale PAT optimization problem. As the gantry rotates, the MRPAT enforces the scanning spots filling the target with the Bragg peaks well inwards away from the target's distal edge and effectively utilizes the high linear energy transfer (LET) of Bragg peaks to enhance Relative Biological Effectiveness (RBE) in the target while mitigating adjacent organs at risk (OARs) damage caused by range uncertainty. The mid-range energy smoothly varies with the beam angle without increasing delivery time as energy switching takes place during gantry rotation.

1 Introduction

The dosimetric advantage of proton arc therapy (PAT) has long been recognized. This is attributed to cross-firing radiation in rotational delivery, providing opportunities to enhance target coverage and spare normal tissues[1, 2]. However, PAT encounters three major challenges: 1) optimization complexity is substantial, given the numerous degrees of freedom involving spot position and intensity for each beam angle, 2) intra-beam energy switching extends the delivery time, and 3) the impact of range uncertainty grows with the number of beams. The conventional plan optimization for PAT relies on range modulation to generate spread-out Bragg peaks (SOBP), like the single field, where each beam angle can deliver spots to cover the entire 3D grid in the target. However, this approach is highly redundant for arc delivery, as arc therapy can deliver a complete 3D grid of spots without SOBP. Various strategies have been proposed to alleviate redundancies in PAT. Techniques such as distal edge tracking (DET) [3], mono- or binary energy delivery [4,5], and geometry-based layers [6] preselect energies before spot intensity optimization [7]. DET has been evaluated against the single field and was found to be less robust for range uncertainty [8]. Other energy layer preselection often aims to minimize energy switching but may encounter challenges in achieving both the efficiency and plan quality aims [7]. Integrated approaches which consider optimizing energy selection simultaneously with spot position and intensity, lead to increased optimization complexity and potential challenges with local minima from imposed energy switching constraints [9]. Hybrid methods involve sequential energy and spot intensity optimization, such as progressively adding energy layers [10] or redistributing them to neighboring beam angles by starting from full energy optimizations on a subset of angles [11]. Despite these efforts, the significant challenges associated with PAT remain unresolved. In this work, we propose using the mid-range beam [12] for each angle in proton arc therapy (MRPAT) combined with the low complexity global optimization framework [13] in tackling these challenges.

2 Materials and Methods

2.1 The mid-range proton arc therapy (MRAPT)

We start outlining MRPAT framework by explaining the mid-range layer definition and its efficacy. First, assume the target center is aligned with the machine isocenter. For each beam angle, we calculate the beam range at the isocenter using water equivalent path length (WEPL) obtained through ray tracing of patient CT volume. This beam range is selected to define the mid-range layer for this beam angle. The mid-range layer goes through the isocenter but may not be the same as the isocenter plane, as illustrated in Figure 1 The isocenter plane is a flat plane perpendicular to the beam direction and marked by the black cross, while the mid-range layer formed by scanning spots, as marked by the red cross, may lie ahead or behind the isocenter plane due to range variations along individual beam paths. A mid-range layer is selected for every beam angle, representing the sole energy utilized in arc treatment plan optimization for that angle.

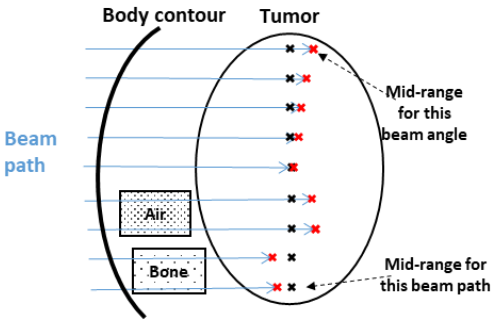


Figure 1: Illustration of the mid-range layer (red cross). The beam is shooting from the left side. The energy for the mid-range layer (red cross) is determined by the beam range at the isocenter point. The spots of the mid-range layer may be ahead or behind the isocenter plane (black cross). This is due to beam range variations along different beam paths.

Despite furnishing only a single energy layer per beam angle, the mid-range layers have the capacity to sweep across the entire 3D grid of spots as the full arc beams provide ample spot coverage for the target. This is illustrated in Figure 2 with small circles representing the spot (Bragg Peak) positions of mid-range layer, one color for one angle.

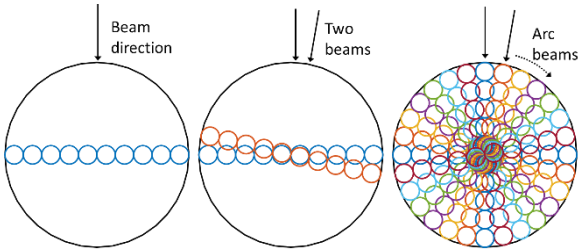


Figure 2: Illustration of mid-range layers that cover the target with spots. From left to right, a single beam delivers a mid-range layer of spots indicated by the small circles; with two beams, the second layer adds more spots. With arc beams, other layers sweep in the target and could fill the targets with spots.

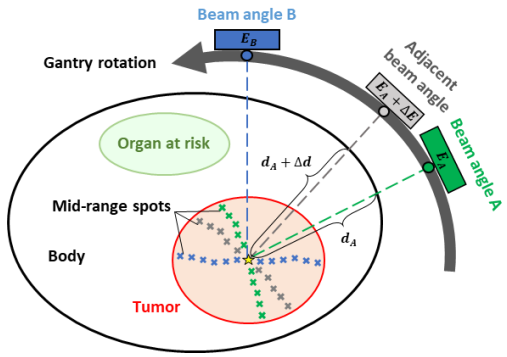


Figure 3: Illustration of proton arc delivery with mid-range energy layers and smooth energy variations with respect to gantry rotation. As the angle change from A to its adjacent angle, the energy changes from E_A to $E_A + \Delta E$, accounting for a small path length change Δd.

In addition to providing ample coverage to the target, the mid-range energy sequence has several features. First, the energy varies smoothly with respect to the beam angle because the WEPL from the patient surface to the isocenter (Figure 3) is a smooth continuous function of beam angles. Such smooth energy variation would not increase the delivery time as energy switching can take place during gantry rotation (typically 1 revolution per minute gantry rotation [11]). Second, the dose delivered by mid-range spots is mainly attributed to the Bragg peaks, which provide the maximum linear energy transfer (LET) and potential relative biological effectiveness (RBE) [14]. In contrast, the commonly utilized SOBP technique relies on a significant portion of dose from the entrance dose with low LET. Third, each time the delivered spots are the mid-range spots. This helps mitigate the range uncertainty issues, and the plan is more robust, as the mid-range spots are likely to remain in the target despite range uncertainty.

A MRPAT plan is created by optimizing intensity of all spots on the mid-range layers from all beam angles, we adopt the global optimization framework, in which [13], the spot dose is modeled by kernel decomposition [15], and optimization can be implemented through fast Fourier transformation (FFT), leading to complexity reduction by several orders of magnitude [13]. Figure 4 illustrate the entire workflow of MRPAT, and the optimization problem is formulated as Equation (1).

$$\begin{aligned} &\underset{w}{\text{minimize}}\, f(D) \\ s.t.\quad &D = K_0 \otimes w \\ &w \geq 0 \end{aligned} \tag{1}$$

where $f(\cdot)$ is a penalty function using medical constraints to evaluate the quality of the dose distribution D which is modeled via the convolution between reference kernel K_0 and corrected spot intensities w. And w is iteratively updated by the derivative shown as Equation (2) in which $K_0'(x) = K_0(-x)$.

$$\frac{\partial f}{\partial w} = \frac{\partial f}{\partial D} \otimes K_0' \tag{2}$$

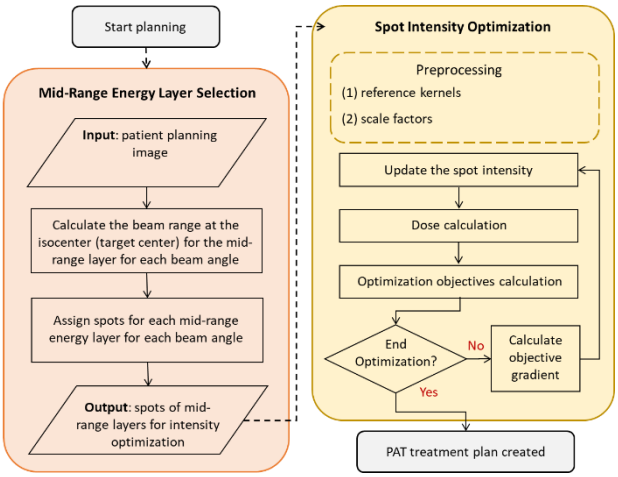


Figure 4: Workflow of MRPAT. The workflow starts with mid-range energy layer selection (orange box) followed by spot intensity optimization (yellow box).

We further elucidate the low-complexity advantage of MRPAT by listing the processes and corresponding computation complexity by steps in Table 1. Assuming that the dose grid has a dimension of $M \times M \times M$, spot sampling in dimension $N \times N \times N$ and the discretized beam angle number is K. For conventional fluence map optimization (FMO) using all candidate spots, the spatial and computational complexity both could grow up to $\mathcal{O}(K \cdot N^3 \cdot M^3)$ contributed from $K \cdot N^3$ spot candidates and M^3 dose voxels. In contrast, the MRPAT-FFT framework only need to store the reference kernels, thus the spatial complexity is $O(K \cdot M^3)$. Additionally, with the FFT, the highest computational complexity could be reduced to $O(K \cdot M^3 \cdot logM)$. For a numerical case, if we set $M = 2^7$, $N = 2^5$ and $K = 2^8$, both spatial and temporal complexity of full FMO problem are in the order of 16 T ($1\ \text{T} = 2^{40}$),

while for MRPAT, the spatial complexity reduced to 0.5 G ($1\ \mathrm{G} = 2^{30}$) and temporal complexity reduced to 4 G, which can be easily operated with a state of art computer workstation.

Spot Intensity Optimization in MRPAT	
Pre-processing	
K_0, the reference kernel	$\mathcal{O}(K \cdot M^3)$
$\widehat{K_0}$, FFT of the reference kernel	$\mathcal{O}(K \cdot M^3 logM)$
Calculation of dose and optimization objective	
$\hat{w}$, FFT of the corrected spot intensity	$\mathcal{O}(N^3 logN)$
$\hat{D} = \widehat{K_0} \cdot \hat{w}$, inner product	$\mathcal{O}(K \cdot M^3)$
$D = \mathcal{F}^{-1}(\hat{D})$, inverse FFT of dose	$\mathcal{O}(M^3 logM)$
$f(D)$, the objective function	$\mathcal{O}(M^3)$
Objective gradient calculation ∇f	
$y = \partial f / \partial D$, derivatives with respect to D	$\mathcal{O}(M^3)$
$\hat{y}$, FFT of the derivatives	$\mathcal{O}(M^3 logM)$
$\widehat{\nabla f} = \hat{y} \cdot \widehat{K_0'}$, inner product	$\mathcal{O}(K \cdot M^3)$
$\nabla f = \mathcal{F}^{-1}(\widehat{\nabla f})$, inverse FFT of the derivate	$\mathcal{O}(M^3 logM)$

Table 1: Computation complexities for spot intensity optimization in the MRPAD framework.

2.2 Simulation study

To demonstrate the feasibility of MRPAT, we compared three proton plans: single field full energy (denoted by full-1), conventional arc plan (full energy per angle for 128 angles, denoted by full-arc PAT), and MRPAT (the mid-range single energy per angle for 128 angles). Figure 5 illustrates the studied phantom, which is a cylindrical phantom (2D circle) with a centrally located target surrounded by four organs at risk (OARs).

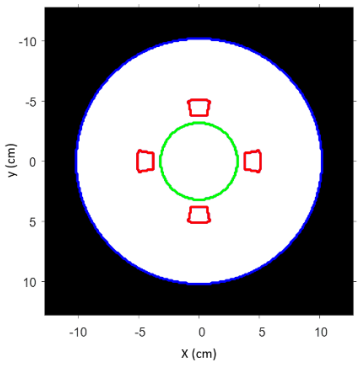


Figure 5: Phantom for our simulation study. It is a cylindrical phantom of density 1 (2D circle, radius of 10 cm) outlined in blue, with the target located at the isocenter with a radius of 3 cm outlined in green and surrounded by four organs at risk (OARs) outlined in red.

3 Results

The dose distributions of the three plans are shown in Figure 6. For the single projection plan, the entrance dose was higher. The other two plans, full-arc PAT and MRPAT, exhibit similar target coverage and OAR sparing, but their dose patterns are very different. MRPAT used more angles and spread out the low dose over the whole region. As noted previously, spot utilization is very redundant in arc beams with full energy. Therefore, though the MRPAT plan only utilized one energy layer per angle, it can achieve similar dose objectives as the full-arc PAT plan. The dose volume histograms (DVH) in Figure 7 also indicate that these two plans, MRPAT and full-arc PAT, had similar performance. The optimization results are summarized in Table 2. The total intensity utilized from lowest to highest was with full-1 beam, full-arc PAT, and MRPAT. This is because the proportion of high energy spots employed from highest to lowest was by full-1, full-arc PAT, and MRPAT, and the integral dose of higher energy spot is higher and thus requiring lower intensity to achieve the desired dose. The DVH points (D_2 and D_{98}) and homogeneity index indicate the two arc plans had similar plan quality and much improved OAR sparing compared to the single field plan with full energy layer utilization. However, the calculation time was substantially different. For arc beams, intensity optimization for full-arc PAT was around tenfold longer than that for MRPAT in this 2D study.

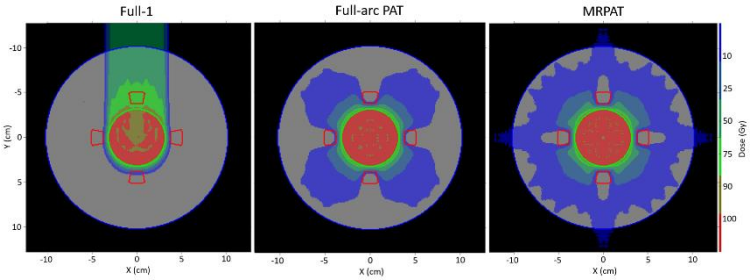


Figure 6: Dose distribution comparison. Left to right: single projection with full energy layers, arc beams (128 angles) with full energy for each angle, arc beams (128 angles) with a single midrange energy layer for each angle. The left side caption describes the structures, and the right-side color bar describes the dose levels.

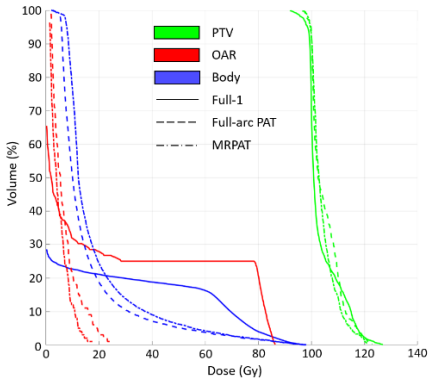


Figure 7: DVH comparison. The DVHs for the target PTV (green), OAR (red), and body (blue) are shown for the three plans: single angle with full energy (solid), arc beams with full energy (dash), arc beams with midrange energy (dash-dot).

Spot Mode	# Angle	# Energy	# Spots	Intensity	PTV D_2 (Gy)	PTV D_{98} (Gy)	OAR D_2 (Gy)	HI	CalcTime (s)
full	1	21	330	647.99	120.42	96.30	85.38	0.24	24.17
full	128	26	24626	670.42	120.64	97.90	21.69	0.23	7425.78
mid-range	128	1	1756	794.59	119.69	98.64	14.45	0.21	689.79

Table 2: Summary of the optimization results of the three compared plans, with the following column (number) labels: (1) the energy mode, (2) the number of utilized angles, (3) the number of maximum energy layers utilized per angle, (4) the total number of spots, (5) the total intensity of spots, (5) D_2 of PTV, (6) D_{98} of PTV, (7) D_2 of OAR, (8) homogeneity index, (9) and calculation time in second.

4 Discussion

We propose a novel MRPAT framework that employs a single pre-selected energy layer—the mid-range energy layer—for each angle to address the challenges of optimization, delivery time, and range uncertainty in proton arc therapy. Our phantom simulation shows MRPAT is feasible and produces plan quality comparable to that of conventional arc beams, as full energy utilization in arcs is redundant. In this investigation, we employed the WEPL to the target center to determine the mid-range layer. Alternative methods for mid-range layer selection exist, such as using the median or mean energy of midpoints along the beam paths that intersect the target. The choice between methods may depend on the specific characteristics of the target shape. With MRPAT, optimization is simplified without considering energy selection and optimization, reducing the problem size by an order of magnitude compared to a full-arc plan's optimization. Implementing MRPAT in the global optimization framework [13] drastically reduces computation time, making it suitable within clinical time frames.

Arc beam delivery time, which depends on the number of energy switches, can be extensive. With MRPAT, there is no intra-beam energy switching, and inter-beam energy switching can overlap with gantry rotation without additional time as energy varies smoothly with beam angle. The MRPAT enhances plan robustness, as each beam delivers spots primarily in the middle region of the target, likely remaining within the target despite range uncertainty.

An additional potential benefit of MRPAT is that the dose in the target is mainly attributed to the Bragg peaks, unlike in SOBP, where a large portion of the dose is from entrance dose. The region near Bragg peaks have maximum LET, potentially improving RBE. Future studies will include LET distributions for RBE-based MRPAT.

5 Conclusion

We introduced a novel MRPAT framework for proton therapy and demonstrated its ability to generate plan quality comparable to full-energy PAT delivery. MRPAT reduces optimization complexity, facilitates efficient arc delivery, and establishes robust plans by leveraging mid-range spots. Key advantages of MRPAT include harnessing Bragg peaks for dose distribution, offering the potential to maximize LET for enhanced RBE.

References

[1] G. A. Sandison, E. Papiez, C. Bloch, and J. Morphis, "Phantom assessment of lung dose from proton arc therapy," *International Journal of Radiation Oncology*Biology*Physics*, vol. 38, no. 4, pp. 891–897, Jul. 1997, doi: 10.1016/S0360-3016(97)00059-X.

[2] A. N. Schreuder and J. Shamblin, "Proton therapy delivery: what is needed in the next ten years?," *Br J Radiol*, vol. 93, no. 1107, p. 20190359, Mar. 2020, doi: 10.1259/bjr.20190359.

[3] J. O. Deasy, D. M. Shepard, T. R. Mackie. Distal edge tracking: A proposed delivery method for conformal proton therapy using intensity modulation. Paper presented at: XIIth International Conference on the Use of Computers in Radiation Therapy1997.

[4] J. P. Blanco Kiely and B. M. White, "Dosimetric feasibility of single-energy proton modulated arc therapy for treatment of chordoma at the skull base," *Acta Oncol*, vol. 55, no. 9–10, pp. 1243–1245, 2016, doi: 10.3109/0284186X.2016.1170199.

[5] D. Sanchez-Parcerisa *et al.*, "Range optimization for mono- and bi-energetic proton modulated arc therapy with pencil beam scanning," *Phys Med Biol*, vol. 61, no. 21, pp. N565–N574, Nov. 2016, doi: 10.1088/0031-9155/61/21/N565.

[6] W. Cao *et al.*, "Intensity modulated proton arc therapy via geometry-based energy selection for ependymoma," *Journal of Applied Clinical Medical Physics*, vol. 24, no. 7, p. e13954, 2023, doi: 10.1002/acm2.13954.

[7] E. Engwall *et al.*, "Fast robust optimization of proton PBS arc therapy plans using early energy layer selection and spot assignment," *Phys. Med. Biol.*, vol. 67, no. 6, p. 065010, Mar. 2022, doi: 10.1088/1361-6560/ac55a6.

[8] U. Oelfke and T. Bortfeld, "Intensity modulated radiotherapy with charged particle beams: studies of inverse treatment planning for rotation therapy," *Med Phys*, vol. 27, no. 6, pp. 1246–1257, Jun. 2000, doi: 10.1118/1.599002.

[9] W. Gu, D. Ruan, Q. Lyu, W. Zou, L. Dong, and K. Sheng, "A novel energy layer optimization framework for spot-scanning proton arc therapy," *Medical Physics*, vol. 47, no. 5, pp. 2072–2084, May 2020, doi: 10.1002/mp.14083.

[10] M. Chen, X. Gu, W. Lu. Efficient proton arc therapy plan with minimal energy layer switches. PTCOG 2020. Taipei.

[11] X. Ding, X. Li, J. M. Zhang, P. Kabolizadeh, C. Stevens, and D. Yan, "Spot-Scanning Proton Arc (SPArc) Therapy: The First Robust and Delivery-Efficient Spot-Scanning Proton Arc Therapy," *International Journal of Radiation Oncology*Biology*Physics*, vol. 96, no. 5, pp. 1107–1116, Dec. 2016, doi: 10.1016/j.ijrobp.2016.08.049.

[12] M. Chen, Y. Zhong, Y. Shao, S. Jiang, and W. Lu, "Mid-Range Probing – Towards Range-Guided Particle Therapy," *Physics in medicine and biology*, vol. 63, no. 13, p. 13NT01, Jun. 2018, doi: 10.1088/1361-6560/aaca1b.

[13] M. Chen, X. Gu, and W. Lu, "Global optimization for spot-based treatment planning," *Medical Physics*, vol. 49, no. 12, pp. 7648–7660, Dec. 2022, doi: 10.1002/mp.15890.

[14] H. Paganetti, "Relative biological effectiveness (RBE) values for proton beam therapy. Variations as a function of biological endpoint, dose, and linear energy transfer," *Phys. Med. Biol.*, vol. 59, no. 22, p. R419, Oct. 2014, doi: 10.1088/0031-9155/59/22/R419.

[15] M. Chen, Z. Yang, Z. Wardak, S. Stojadinovic, X. Gu, and W. Lu, "Dose kernel decomposition for spot-based radiotherapy treatment planning," *Medical Physics*, vol. 49, no. 2, pp. 1196–1208, 2022, doi: 10.1002/mp.15415.

From Concept to Clinic: How Our In-house Automated Planning System Augmented the Commercial TPS for Treating Over 10,000 Patients

Masoud Zarepisheh*, Gourav Jhanwar, Ying Zhou, Qijie Huang, Hai D. Pham, Jie Yang, Laura Cervino, Jean M Moran, Joseph O. Deasy, and Linda Hong

Department of Medical Physics, Memorial Sloan Kettering Cancer Center, New York, NY, US
*Corresponding author: zarepism@mskcc.org

Abstract Most modern Treatment Planning Systems (TPS) come with Application Programming Interface (API) scripting capabilities, allowing for the development of custom applications that enhance the functionalities of commercial TPS. In our clinic, we have developed and integrated ECHO (Expedited Constrained Hierarchical Optimization), an in-house automated planning system, with the Eclipse TPS. ECHO has played a pivotal role in the treatment of over 10,000 patients. It streamlines the planning process: after preparing the contours and beams (IMRT fields or VMAT arcs), users launch the ECHO plug-in within the TPS, select optimization structures, and run the program. ECHO automatically retrieves necessary data from the TPS, constructs and solves optimization problems, and imports the final results (optimal fluence for IMRT, optimal control points for VMAT) back into the TPS. It then runs final dose calculation in TPS and notifies the user via email for plan review, all within 15-120 minutes. ECHO completely bypasses TPS optimization engine, however, it utilizes TPS's final dose calculation and, for IMRT, leaf sequencing. In this work, we discuss our development journey and clinical experiences, outlining both past and current technical challenges, along with the solutions we have devised. We have also recently introduced an open-source project, PortPy (Planning and Optimization for Radiation Therapy in Python), which aims to share our experience with the wider research community.

1 Introduction

Despite advancements, radiotherapy treatment planning remains a complex, time-consuming, and labor-intensive process, with the quality of plans often depending on the experience and skills of the planners. Automated planning has been a significant focus in the scientific community for approximately two decades, leading to many advancements[1–3]. However, the transition of research ideas into clinical practice by commercial vendors frequently takes years, and even when these innovations are introduced, clinics may not have access to the specific vendors or the products might not meet the unique needs of their criteria. For instance, a significant proportion of our patients (over 30% of our spinal patients) have undergone prior radiation, necessitating a customized plan that accounts for their previous treatments. Class-solution techniques, like the knowledge-based method found in the Eclipse system, is not applicable for these patients. Moreover, our clinical criteria undergo periodic revisions, potentially necessitating retraining for systems dependent on previous practices.

Most modern Treatment Planning Systems (TPSs) are equipped with API scripting capabilities, which allow researchers to enhance the functionalities of TPSs. Our group has developed a fully automated treatment planning system, named Expedited Constrained Hierarchical Optimization (ECHO), that integrates seamlessly with the FDA-approved Varian Eclipse system through API scripting and facilitated the treatment of over 10,000 patients to date. In this work, we will briefly explore some of the technical challenges we have faced, along with the solutions we have devised. To share our experience with the broader research community, our group has recently launched an open-source initiative, Planning and Optimization for Radiation Therapy in Python (PortPy). PortPy has seen over 7,000 downloads in the past four months, indicating its widespread adoption and use within the scientific community.

2 Materials and Methods

In the subsequent sections, we will briefly cover some of the hurdles encountered in creating a self-contained optimization engine designed to augment the commercial TPS. Table (1) provides a summary of these challenges, our technical solutions, their current implementation status within our clinic, and associated publications.

Table 1. Challenges, Solutions, and Their Implementation Status in our Clinic.

Implement. Status	Challenge	Solution	Pub.
Deployed in clinic	Multi-criteria challenge	Hierarchical Opt.	[2,4]
R&D		Deep Learning	[5]
Deployed in clinic	Dose discrepancy between opt. and final dose calc.	Lagrangian Method	[2,6]
R&D		Compression	Ongoing
Deployed in clinic	Non-convex DVH constraints	Convex Approximation	[7]
Deployed in clinic	Non-convex VMAT	Sequential Convex Prog.	[6,8]
R&D	Reducing plan complexity	Wavelets	[9]
R&D	Proton beam selection	Bayesian optimization	[10]
R&D	Proton uncertainty management	Robust/distributed optimization	[11,12]

2.1 Non-convexity and Local Sub-optimality. It is well-known that some treatment planning optimization problems lead to non-convexity, causing the optimization algorithm to potentially trap in a "local" sub-optimal solution and fails to reach the best "global" optimal solution. However, most of these non-convex problems can be solved to global optimality through the use of computationally demanding Mixed Integer Programming (MIP). While solving MIP

problems might require days or weeks for a single patient, making it impractical for regular use, they can be solved in the absence of time constraint and serve as valuable benchmarks for validating more computationally efficient algorithms. Figure 1 showcases the comparison between the global optimal solution (solid lines) and the results of our algorithms (dashed lines) for two distinct non-convex optimization problems: 1) Intensity Modulated Radiation Therapy (IMRT) optimization with Dose-Volume Histogram (DVH) constraints[7] (left figure), and 2) Volumetric Modulated Arc Therapy (VMAT) optimization including direct machine parameter optimization[6,8] (right figure). These comparisons verify that our algorithms can generate solutions near the global optimal. Our open-source package, PortPy, enables researchers to achieve global optimal solutions for complex non-convex optimization problems, including those with DVH constraints, VMAT, and beam orientation optimization, further facilitating the development and validation of new techniques by the research community[13].

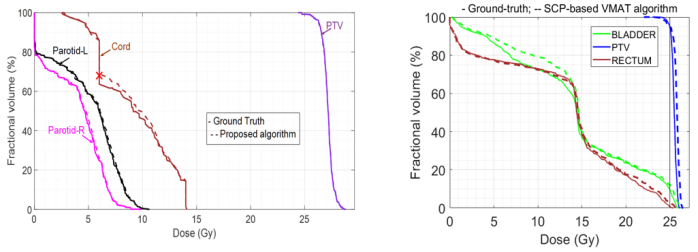


Figure 1. The global optimal solutions (solid lines) for DVH constraints (left) and VMAT (right) are compared against the results of our algorithms (dashed lines).

2.2 Dose Discrepancy Between Optimization and Final Dose Calculation. Treatment planning optimization necessitates the pre-calculation of the radiation dose delivered to the patient's body from thousands of small "beamlets", typically stored in a matrix known as the dose influence matrix or dose deposition matrix. This matrix represents the main computational challenge in optimization problems and is often truncated into a sparse matrix to enhance computational efficiency. However, this sparsification results in a discrepancy between the optimized dose and the finally calculated dose. While there are other sources of dose discrepancy, such as the tongue-and-groove effect and leaf transmission, their impact is usually much less significant. To mitigate this issue, a "correction loop" has been proposed in the literature[14], which periodically calculates a more accurate dose and reintegrates this information into the optimization problem for adjustments. This approach necessitates interruptions in the optimization algorithm, making it feasible only with simpler optimization algorithms like gradient descents, which are suitable for solving unconstrained optimization problems. Unfortunately, this technique is not applicable to more complex algorithms, such as interior point methods, which are utilized for solving constrained optimization problems. To overcome this limitation, we have proposed the reconstruction of an equivalent unconstrained version of a constrained problem using Lagrangian multipliers[2]. This approach allows for the correction of dose discrepancies without disrupting the optimization process. Although the "correction loop" strategy helps to reduce the dose discrepancy issue, it does not fully resolve it, and relying on such an *ad hoc* technique could potentially compromise the quality of the treatment plan. Upon a more detailed examination of the dose influence matrix, we have recently discovered that it is highly compressible, thanks to the correlations among machine parameters such as adjacent beams and beamlets. Our preliminary findings show an impressive compression rate of over 98% with a minimal error margin (approximately 2%). These results will be discussed in detail in a separate study. It is worth mentioning that our open-source initiative, PortPy, encompasses all necessary data for optimization, including the dose influence matrix, which has been pre-calculated for many patients (currently 50 lung patients) and extracted from the Eclipse system.

2.3 Optimally Balancing Plan Complexity and Dose Conformity. IMRT/VMAT effectiveness relies on conforming radiation dose to the tumor by modulating radiotherapy machine parameters (e.g., fluence, multi-leaf collimator). Although some modulation is necessary, increased modulation does not always enhance dose conformity, and excessive modulation can increase plan complexity and hinder the delivery accuracy. The importance of optimally balancing dose conformity and plan complexity has been long recognized, yet it remains a challenge as also noted in a 2020 Red Journal Editorial Note[15] and highlighted by ESTRO[16,17]. In IMRT, we have recently introduced a novel wavelet-based planning framework[9], which treats each beam's 2-dimensional modulated fluence as an image and utilizes advanced wavelet tools from imaging science to induce local and global fluence smoothness. This framework also allows for the proactive elimination of overly complex and clinically irrelevant plans by employing only low-frequency wavelets. For VMAT, we have recently developed a new VMAT planning framework, named Sequential Convex Programming[6,8] (SCP). The SCP-based VMAT framework is the first, to our knowledge, that fully integrates plan complexity into treatment planning and identifies a VMAT plan close to the “global” optimal plan.

2.4 Safety of In-house Developed Software. Integrating with a commercial TPS and conducting the final dose calculation and evaluation within it offers a level of reassurance. However, implementing a strict quality assurance (QA) protocol for in-house software development is crucial, particularly for applications like ours that entirely bypass the TPS optimization engine. Our team adheres to the guidelines outlined by Moran et al. (2022)[18] for the final evaluation and documentation of our in-house software projects. Moreover, we conduct thorough retrospective studies with randomly chosen patients, comparing the manually generated treatment plans used for treatment with those generated by our ECHO system, before introducing new disease sites

or fractionation regimens (see Figure 4). Film and portal dosimetry, in alignment with departmental policies, are also utilized for QA purposes.

3 Results

Figure 2 illustrates the ECHO's system design and workflow. Once the structure contours and beams/arcs are set up, users launch ECHO as a plugin from the Eclipse TPS, selecting structures for optimization. The API then retrieves necessary patient data (e.g., beam parameters, influence matrix) for optimization. The system solves the optimization problems, importing optimal machine settings (fluence for IMRT, control points for VMAT) back into Eclipse for the final dose calculation. There is a correction-loop to align optimized and calculated doses. Lastly, an automated email notifies the user when the plan is ready for review. This workflow is fully automated, requiring user input only for initial contour and beam setup.

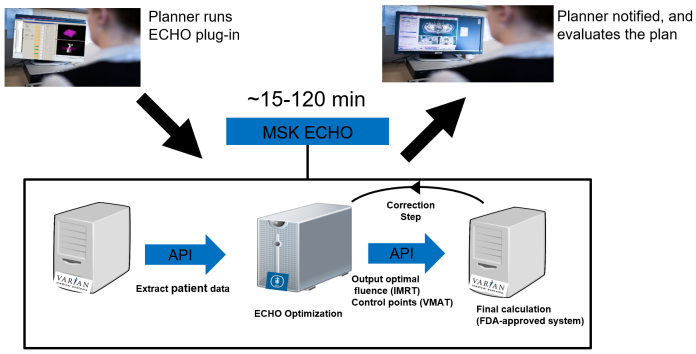


Figure 2. ECHO workflow is fully automated using API scripting. ECHO completely bypasses the TPS optimization engine, but it relies on the TPS for final dose calculation and, in case of IMRT, for the final leaf sequencing as well.

ECHO has played a pivotal role in the treatment of over 10,000 patients across a range of disease sites such as spine, prostate, lung, and oligometastases at our center. Figure 3 displays the monthly plan generation statistics through ECHO and the distribution across different disease sites.

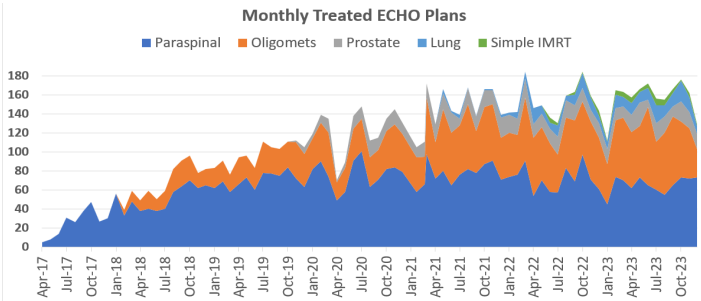


Figure 3. About 180 plans are generated monthly using ECHO.

As previously mentioned, we conduct a retrospective analysis of ECHO's performance prior to introducing a new site or prescription regimen. This analysis encompasses both quantitative comparisons using clinically relevant dose metrics, akin to those depicted in Figure 4, and qualitative evaluations of 3D dose distribution and DVH by experienced physicists and radiation oncologists.

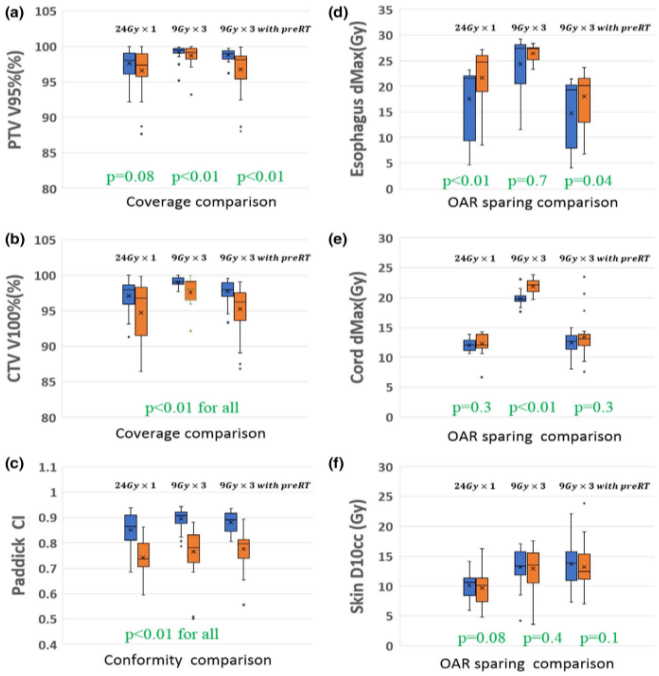


Figure 4. Pre-clinical validation, conducted prior to clinical deployment of new site or prescription, includes simulating ECHO plans for previously treated patients and performing a comparative analysis.

4 Discussion

API scripting, now accessible in major Treatment Planning Systems (TPS), serves as a robust mechanism to enhance clinical workflows through increased automation. This work demonstrates the integration of home-grown optimization techniques into clinical operations by extracting data from the TPS, executing optimization processes externally, and seamlessly reintegrating the results back into the TPS without manual intervention. However, it is important to acknowledge that API scripting, while transformative, is still an evolving technology and might not encompass all necessary functionalities. For instance, due to the limited API capabilities in the clinical version of Eclipse, we had to transfer patient data to a non-clinical version of Eclipse for processing. This process, though automated via Varian's DB Daemon transfer tools, introduces additional complexity.

The criticality of safety and quality assurance in the development of in-house software for critical applications like radiotherapy cannot be overstated. In addition to performing comprehensive and objective retrospective evaluations to determine the impact of the software on the quality of care, it is recommended to follow established guidelines, like those suggested by Moran et al.[18], to ensure the highest levels of safety and efficacy.

Finally, we hope our open-source project, PortPy, accelerates research in treatment planning optimization and

encourages the development of new treatment planning techniques. These could potentially be applied clinically using API scripting, bridging the gap between research and clinical use.

5 Conclusion

We have developed and successfully implemented ECHO, a fully automated treatment planning system. Thanks to the integration capabilities provided by Eclipse API scripting, ECHO is now fully integrated with commercial TPS and seamlessly incorporated into our clinical workflow. ECHO has been instrumental in the treatment of over 10,000 patients, reducing the average planning time by half—from a full day to just half a day—while significantly improving the quality and consistency of treatment plans.

References

[1] Breedveld, S., Storchi, P. R. M. & Heijmen, B. J. M. The equivalence of multi-criteria methods for radiotherapy plan optimization. *Phys. Med. Biol.* **54**, 7199–7209 (2009).

[2] Zarepisheh, M. *et al.* Automated intensity modulated treatment planning: The expedited constrained hierarchical optimization (ECHO) system. *Med. Phys.* **46**, 2944–2954 (2019).

[3] Nguyen, D. *et al.* Advances in Automated Treatment Planning. *Semin. Radiat. Oncol.* **32**, 343–350 (2022).

[4] Hong, L. *et al.* Clinical Experience of Automated SBRT Paraspinal and Other Metastatic Tumor Planning With Constrained Hierarchical Optimization. *Adv Radiat Oncol* **5**, 1042–1050 (2020).

[5] Jhanwar, G., Dahiya, N., Ghahremani, P., Zarepisheh, M. & Nadeem, S. Domain knowledge driven 3D dose prediction using moment-based loss function. *Phys. Med. Biol.* **67**, (2022).

[6] Dursun, P. *et al.* Automated VMAT treatment planning using sequential convex programming: algorithm development and clinical implementation. *Phys. Med. Biol.* **68**, (2023).

[7] Mukherjee, S., Hong, L., Deasy, J. O. & Zarepisheh, M. Integrating soft and hard dose-volume constraints into hierarchical constrained IMRT optimization. *Med. Phys.* **47**, 414–421 (2020).

[8] Dursun, P., Zarepisheh, M., Jhanwar, G. & Deasy, J. O. Solving the volumetric modulated arc therapy (VMAT) problem using a sequential convex programming method. *Phys. Med. Biol.* **66**, (2021).

[9] Tefagh, M. & Zarepisheh, M. Built-in wavelet-induced smoothness to reduce plan complexity in intensity modulated radiation therapy (IMRT). *Phys. Med. Biol.* **68**, (2023).

[10] Taasti, V. T., Hong, L., Shim, J. S. A., Deasy, J. O. & Zarepisheh, M. Automating proton treatment planning with beam angle selection using Bayesian optimization. *Med. Phys.* **47**, 3286–3296 (2020).

[11] Taasti, V. T., Hong, L., Deasy, J. O. & Zarepisheh, M. Automated proton treatment planning with robust optimization using constrained hierarchical optimization. *Med. Phys.* **47**, 2779–2790 (2020).

[12] Fu, A., Taasti, V. T. & Zarepisheh, M. Distributed and scalable optimization for robust proton treatment planning. *Med. Phys.* **50**, 633–642 (2023).

[13] Jhanwar, G. *et al.* Portpy: An Open-Source Python Package for Planning and Optimization in Radiation Therapy Including Benchmark Data and Algorithms. *AAPM 65th Annual Meeting & Exhibition*.

[14] Siebers, J. V., Lauterbach, M., Tong, S., Wu, Q. & Mohan, R. Reducing dose calculation time for accurate iterative IMRT planning. *Med. Phys.* **29**, 231–237 (2002).

[15] Kamperis, E., Kodona, C., Hatziioannou, K. & Giannouzakos, V. Complexity in Radiation Therapy: It's Complicated. *International Journal of Radiation Oncology*Biology*Physics* **106**, 182–184 (2020).

[16] Kaplan, L. P. *et al.* Plan quality assessment in clinical practice: Results of the 2020 ESTRO survey on plan complexity and robustness. *Radiother. Oncol.* **173**, 254–261 (2022).

[17] Hernandez, V. *et al.* What is plan quality in radiotherapy? The importance of evaluating dose metrics, complexity, and robustness of treatment plans. *Radiother. Oncol.* **153**, 26–33 (2020).

[18] Moran, J. M. *et al.* A Safe and Practical Cycle for Team-Based Development and Implementation of In-House Clinical Software. *Adv Radiat Oncol* **7**, 100768 (2022).

Fully-automated integrated multi-criterial optimisation of transmission pencil beams in IMPT for penumbra sharpening

W. Kong[1], M. Huiskes[2], S. J. M. Habraken [2,3], C.R.N. Rasch[2,3], E. Astreinidou[2], B.J.M. Heijmen[1], and S. Breedveld[1]

[1]Department of Radiotherapy, Erasmus MC Cancer Institute, Erasmus University Medical Center, Rotterdam, The Netherlands
[2]Department of Radiation Oncology, Leiden University Medical Center, Leiden, The Netherlands
[3]HollandPTC, Delft, The Netherlands

Abstract

Objective: In intensity-modulated proton therapy (IMPT), Bragg peaks (BPs) result in a steep dose fall-off at the distal target end. Typically, the lateral dose fall-off is less steep. High-energy pristine transmission pencil beams (TPBs) with BPs outside the patient exhibit a sharp lateral penumbra at the target level. In this study, we hypothesised that by combining IMPT BPs with TPBs (denoted 'IMPT&TPB'), dose distributions in head-and-neck cancer patients could be improved in terms of OAR dose compared to conventional IMPT with only BPs. *Approach:* Our spot selection algorithm (SISS-MCO) for fully-automated IMPT treatmentplanning[1–3], embedded in Erasmus- iCycle, was extended for the combined optimisation of IMPT BPs and TPBs. It employs automated multi-criteria optimisation to generate for each patient a single Pareto-optimal plan using prioritised planning objectives and hard constraints, as specified in a generic wish-list[2]. For 8 nasopharynx cancer patients and 8 oropharynx cancer patients, the IMPT&TPB plan was compared to the competing IMPT plan. *Main results:* Clinical OAR and target constraints were met in all plans. On average, 11 of the 19 investigated OAR plan parameters improved significantly by allowing TPBs in the optimisations, while one parameter showed small but significant deterioration. *Conclusion:* For head and neck cancer patients, IMPT&TPB resulted in comparable target coverage with overall superior OAR sparing, the latter originating from sharper penumbras close to OARs.

1 Introduction

External beam radiation therapy with protons instead of photons can reduce radiation exposure to critical organs through the Bragg peaks (BP). However, the lateral dose fall-off may be relatively shallow, as proton pencil beams have a considerable penumbra at the patient's entrance level. This problem is especially relevant in intensity modulated proton therapy (IMPT) for head and neck cancer (HNC), where the applied range shifters, used to treat more superficial target regions, result in increased penumbras of pencil beams entering the patient. Inside patients, penumbras further increase due to unavoidable multiple Coulomb scattering, resulting in the largest penumbras close to the BPs.

Transmission pencil beams (TPBs) consist of high energy protons with BPs that are effectively outside the patient; tumour dose is deposited by the relatively flat entrance dose plateau of the pencil beams. Consequently, proton energies at the target level are still relatively high, and therefore, the lateral penumbras are relatively steep. The omission of range shifters in TPBs further contributes to relatively steep lateral penumbras in high-dose regions close to the targets. However, a notable drawback of using only TPBs is that it leads to increased irradiation of normal tissues located beyond the distal edge of the target.

In this study, we hypothesised that by combining IMPT BPs with TPBs (denoted 'IMPT&TPB'), dose distributions in HNC patients could be improved in terms of OAR dose compared to conventional IMPT with BPs only.

2 Materials and Methods

2.1 Patient data

This study used anonymised planning CT scans and delineations of eight patients with nasopharyngeal cancer (NPC) and eight patients with oropharyngeal cancer (OPC) who had previously been treated with IMPT. Informed consent was obtained from all patients for the retrospective use of the data.

For all patients, a simultaneous integrated boost scheme was applied, prescribing 70 Gy_{RBE} to the primary tumour (CTV70) and 54.25 Gy_{RBE} to bilateral neck volumes (CTV54.25). For all selected patients, the applied IMPT plan achieved for both CTVs the clinically requested coverage. Robust optimisation was performed for all targets, brainstem, spinal cord, cochlea, optic nerves, optic chiasm, eyes and mandible bone, while non-robust optimisation was performed for parotids, submandibular glands, oral cavity, oesophagus, larynx, pharyngeal constrictor muscles (PCM) (superior, middle, and inferior) and cricopharyngeus.

2.2 Treatment planning

Our algorithm for spot selection (SISS-MCO) for fully-automated multi-criterial IMPT treatment planning[1–3], embedded in Erasmus- iCycle, was extended for combined optimisation of IMPT with BPs and TPBs. For each patient, it generates a single Pareto-optimal plan using predefined prioritised objectives and hard constraints as specified in a generic wish-list[2]. The same wish-list was used for both IMPT and IMPT&TPB.

The SISS-MCO optimisation includes large-scale sparsity-induced BPs/TPBs selection in the first phase and multi-criteria optimisation with a subset of candidate pencil beams from phase 1 in Erasmus-iCycle [2, 4] in the second phase. The subset was chosen such that it constituted 99% of the total candidate pencil beam weight after phase 1. Pencil beams with monitor units (MU) below the clinically minimum deliverable MU were removed in the third phase. For IMPT plans, the optimisations started with 20,000 candidate BPs that were uniformly distributed in the CTVs with a 5 mm expansion, while in IMPT&TPB, 20.000 BPs and 18.000 TPBs were similarly distributed. TPBs were obtained using 244 MeV shoot through beams. The best combination of BPs and TPBs was chosen by the optimiser.

IMPT&TPB and IMPT plans were robustly generated using minimax optimisation with 19 robustness scenarios. For plan comparisons, obtained D98% in the voxel-wise minimum dose distributions for the high-dose CTVs (CTV70) were scaled to 95.0% of the prescription dose.

All plans were generated on a node on our high computing cluster equipped with two Intel Xeon Gold 6248 @ 2.5 GHz CPUs and a NVIDIA Quadro RTX 8000 GPU. Only six of the available 40 CPU cores were used for the calculation of the dose deposition matrices for the candidate pencil beams and for final dose calculations due to the limited multiprocessing capability of our implementation of the applied ASTROID dose engine [5].

2.3 Plan evaluations

For all robustly optimised target and OAR dose parameters, plan evaluations and comparisons were performed accordingly. For robustly optimised OARs, the voxelwise maximum D2% (VWmax D2%) or worst-case scenario mean dose (WC Dmean) were applied, in agreement with the applied wishlist. For non-robustly optimised OARs, the mean dose of the nominal scenario (Dmean) or the nominal scenario D2% was determined depending on the OAR. Differences in dosimetric metrics were tested for significance ($p < 0.05$) using two-sided Wilcoxon signed-rank tests.

3 Results

3.1 Dose evaluations

All IMPT and IMPT&TPB plans met the clinical OAR and target hard constraints applied in our institution. All D2% of the target VWmax dose distributions were below 107% of the prescribed dose. For both NPC and OPC, IMPT&TPB showed small increases in D2% of the target VWmax dose, albeit all values were below clinical constraints.

IMPT&TPB was superior for the majority of OAR plan parameters in both tumour localisations (Figure 1). Depending on the NPC patient, considerable dose differences favouring IMPT&TPB were observed in the WC Dmean in cochleae and, VWmax D2% of the brainstem, optic nerves, optic chiasm and lenses. Furthermore, for both NPC and OPC, reductions were also observed in the mean dose of the submandibular glands (SMG), parotids and the middle pharyngeal constrictor muscle inferior.

3.2 Dose distribution comparisons

The impact of TPBs on IMPT&TPB dose distributions is shown in Figure 2a for an NPC case and in Figure 2b for an OPC case. Often, the optimiser chose TPBs on the edges of the CTV that are proximal to OARs. For both tumour localisations, TPBs were often chosen to shoot between a CTV and a parotid, SMG, or cochleae, and along the edges of the CTVs. For nasopharynx cases, TPBs were also often chosen between the brainstem or optic nerves and a CTV.

3.3 Calculation times, numbers of spots and energy layers

On average, IMPT&TPB and IMPT plan generations took 3.7 and 2.7 hours, respectively. Dose calculations contributed significantly to plan generation times, i.e., 1.8 hours for IMPT&TPB and 1.1 hours for IMPT.

On average, IMPT&TPB plans constituted of 1761 BPs and 792 TPBs, while IMPT plans had 2179 BPs. The average number of selected energy layers was the same for both (39).

4 Discussion

This study has demonstrated that incorporating TPBs into regular IMPT with only BPs could improve the quality of treatment plans, by exploiting the sharper lateral penumbras of TPBs. IMPT&TPB and IMPT showed comparable target coverage, while OAR sparing was overall superior for IMPT&TPB. In nations that employ a model-based selection approach for proton therapy, lower OAR doses can result in more patients treated with proton therapy.

In this work, we have assumed that TPBs were delivered without a range shifter in the beam to maximise the steepness of lateral TPB penumbras. In our institution, two out of the six applied IMPT beams in HNC are already delivered without a range shifter. For the other beams, clinical delivery of simulated IMPT&TPB treatments would require that part of the total dose be delivered without a range shifter, resulting in longer delivery times. Possibly, using the range shifter in these four beams also for TPBs only has a minor negative impact on the resulting IMPT&TPB dose distributions. This needs further investigation. Avoiding the

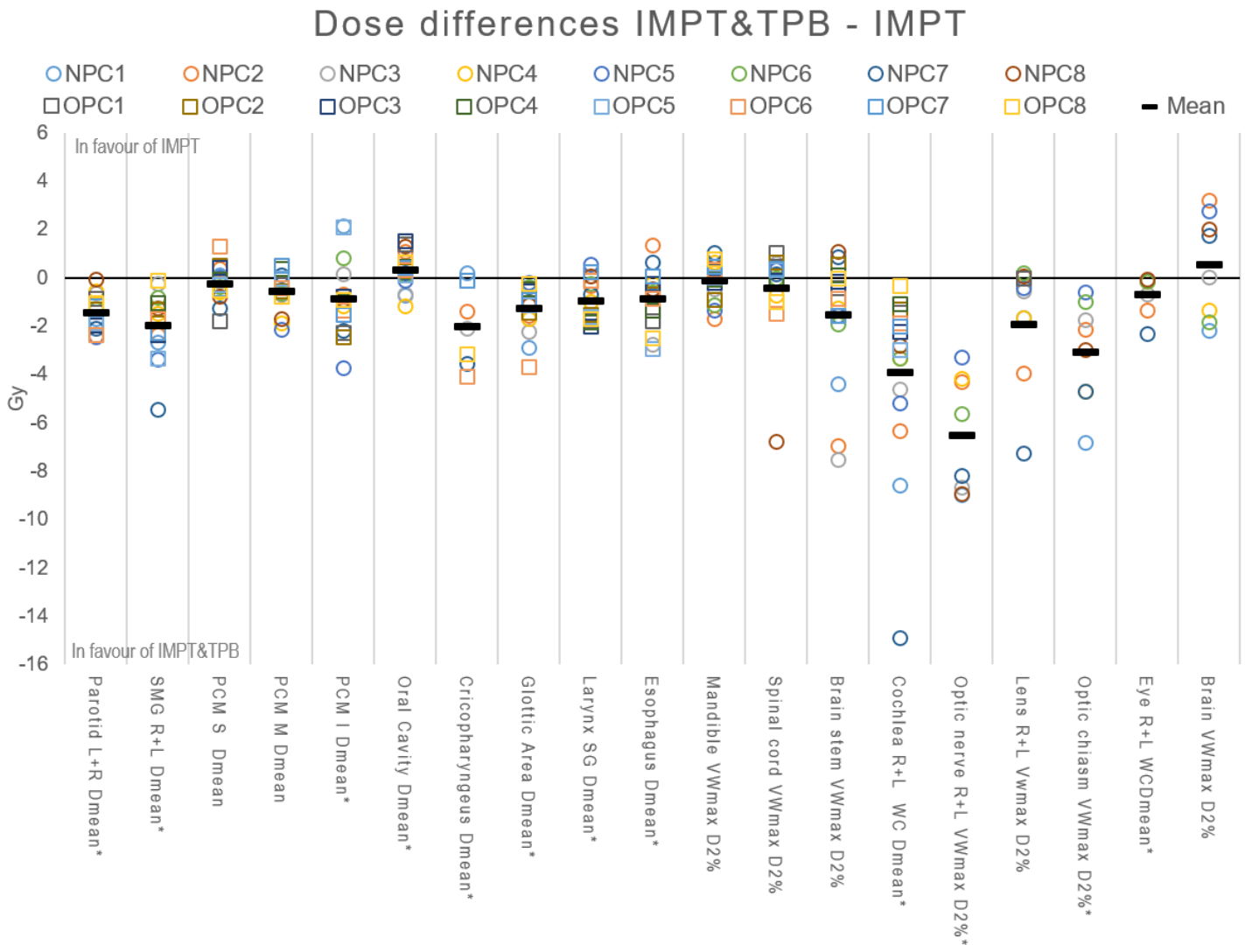


Figure 1: Differences in dosimetric plan parameters of IMPT&TPB and IMPT treatment plans. NPC cases are represented by circles and OPC cases by squares. Statistical significance is denoted by asterisks ($p<0.05$). Abbreviations: R=right, L=left, SMG=submandibular gland, PCM=pharyngeal constrictor muscle, S=superior,I=inferior, M=middle

use of range shifters for TPBs in all beams with only minor delivery time increases could be realised with a system for automatic range shifter insertion.

The added value of using TPBs is apparent but varies between patients. In this study, the planning wish-list that was used for both IMPT&TPB and IMPT was optimised for IMPT in a previous study. Possibly, tailoring of the wish-list to maximally favour penumbra sharpening with TPBs could further enhance OAR sparing for more patients. Also (automated) beam-angle optimisation for IMPT&TPB could maybe enhance the positive impact of TPBs.

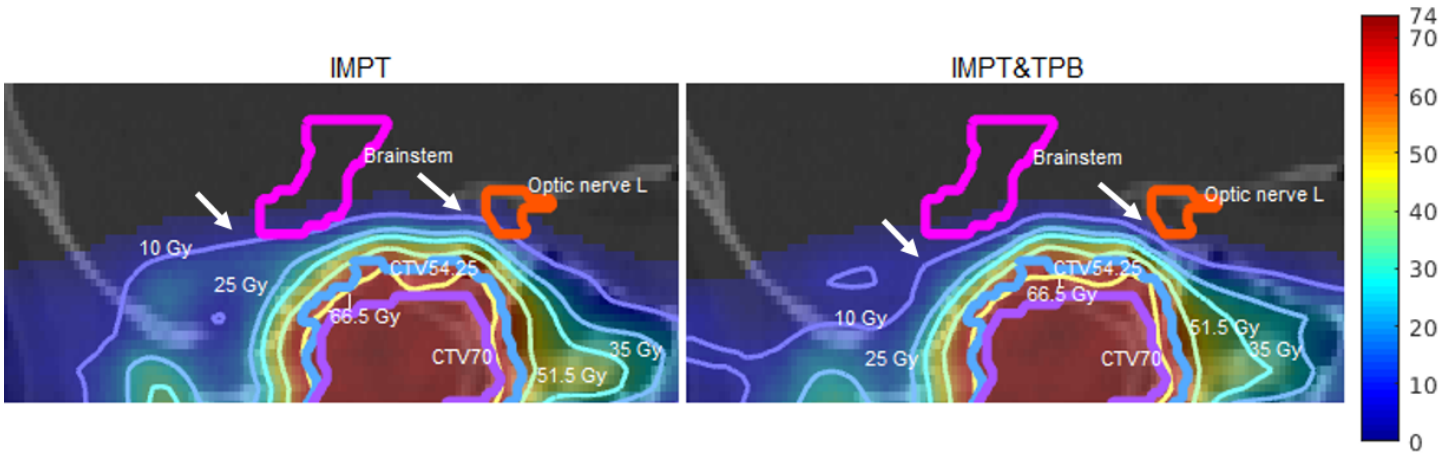
For all included study patients, full coverage of CTV70 was achieved, while in general clinical practice, a substantial part of NPC patients can suffer from insufficient coverage of the high-dose CTV. Possibly, for this patient group, IMPT&TPB could enhance the attainable coverage compared to regular IMPT. These are all topics for further research.

5 Conclusion

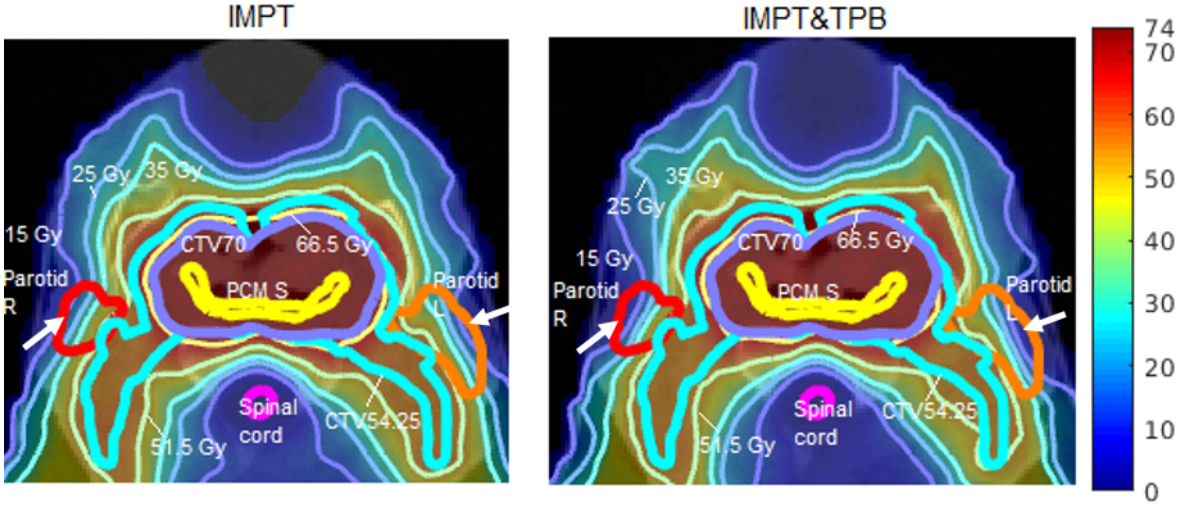
The developed optimiser for fully-automated combined optimisation of IMPT Bragg peaks and transmission pencil beams significantly reduced OAR doses in head-and-neck cancer by using transmission pencil beams to enhance lateral target dose fall-off.

Acknowledgements

This study presented in this manuscript is partly funded by Varian a Siemens Healthineers Company.



(a) NPC



(b) OPC

Figure 2: IMPT&TPB dose distributions for an NPC case **(a)**, and for an OPC case in the axial view in **(b)**. Contours are depicted using thick lines, whereas isodoses are represented with narrow lines. The white arrows indicate regions with sharpened penumbras.

References

[1] W. Kong, M. Oud, S. Habraken, et al. "SISS-MCO: large scale sparsity-induced spot selection for fast and fully-automated robust multi-criteria optimisation of proton plans". *Physics in Medicine and Biology* (2024). DOI: 10.1088/1361-6560/ad1e7a.

[2] S. Breedveld, P. R. Storchi, P. W. Voet, et al. "iCycle: Integrated, multicriterial beam angle, and profile optimization for generation of coplanar and noncoplanar IMRT plans". *Medical physics* 39.2 (2012), pp. 951–963. DOI: 10.1118/1.3676689.

[3] M. Huiskes, W. Kong, M. Oud, et al. "Validation of Fully Automated Robust Multicriterial Treatment Planning for Head and Neck Cancer IMPT". *International Journal of Radiation Oncology*Biology*Physics* (2024). DOI: 10.1016/j.ijrobp.2023.12.034.

[4] S. Breedveld, P. R. Storchi, and B. J. Heijmen. "The equivalence of multi-criteria methods for radiotherapy plan optimization". *Physics in Medicine & Biology* 54.23 (2009), p. 7199. DOI: 10.1088/0031-9155/54/23/011.

[5] H. M. Kooy, B. M. Clasie, H.-M. Lu, et al. "A case study in proton pencil-beam scanning delivery". *International Journal of Radiation Oncology* Biology* Physics* 76.2 (2010), pp. 624–630. DOI: 10.1016/j.ijrobp.2009.06.065.

Interactive dose modification: a novel approach to proton beam therapy treatment planning

Matthew Lowe[1]

[1]Christie Medical Physics & Engineering, The Christie NHS Foundation Trust, Manchester, UK

Abstract A novel approach to treatment planning for pencil beam scanning proton therapy is presented by which the user can interact intuitively with a dose distribution to modify a treatment plan. In this proof-of-principle implementation, the dose distribution is displayed using standalone software independent of the treatment planning system (TPS). The user can mouse click or scroll to increase or decrease dose to any voxel. This is achieved by modifying spot weights in proportion to their contribution to the selected voxel using an influence matrix describing the contribution of each spot to each voxel. The dose distribution is updated in <180 ms on average, appearing real-time to the user. Minimum spot weights are accounted for, ensuring plan deliverability. Modified spot weights are returned to the TPS from which the normal pathway to patient treatment can continue. The proposed approach allows the user to produce a desired dose distribution without knowing how to describe it using optimisation objectives, reducing the expertise needed to modify plans.

1 Introduction

Proton treatment planning using pencil beam scanning (PBS) enables the production of highly conformal dose distributions through the delivery of proton pencil beams (spots) to different locations throughout a patient. Spots are assigned different weights (proportional to the number of protons delivered to each location) which are optimised to provide an individualised patient treatment plan. Typically, several thousand spots from multiple fixed beam directions are used and as such there are many degrees of freedom in producing a clinically desirable treatment plan.

While treatment planning is a multi-disciplinary process, particularly for complex cases, the creation of the plan is done by an expert planner. Spot weight optimisation is done through a process of inverse optimisation: the optimum dose distribution is derived using, e.g., steepest descent, conjugate gradient, or quasi-Newton methods, by minimising an objective function comprising multiple planning objectives, weighted by priority. e.g., the minimum dose to the target volume should be 100% of the prescription dose; the maximum dose to *x*% of a given organ at risk (OAR) should not exceed a dose, *y*. Such objectives do not necessarily replicate the exact clinical aims of the plan but may instead be chosen to drive the optimisation algorithm to the required outcome.

In complex cases, treatment planning can be an art of compromise, requiring a balance of the conflicting aims of target volume coverage and minimising dose to nearby or overlapping OARs. It is a resource intensive task, and the quality and timeliness of plan production is dependent on planner experience. Predicting the minimal achievable dose to OARs or the compromise to target coverage from meeting given aims can be difficult. The optimisation cycle may be initiated multiple times, with the planner responding to the resulting dose distribution by adjusting planning objectives or priorities. Desired changes in one part of the dose distribution can result in changes throughout the dose distribution as a result of the optimisation process.

Once the planner is satisfied with their treatment plan, it is reviewed by the prescribing oncologist who, in light of their own expertise, may request changes to the plan, providing feedback to the planner and requiring further investigation and optimisation before it is reviewed again.

The tool presented here aims to reduce both the time taken to produce a desired dose distribution and the barrier to entry for improving treatment plans. It aims to benefit planning staff as well as empowering oncologists who review treatment plans, but are not trained to produce them, to make changes to meet their requirements.

This tool exists as standalone software taking information from an existing PBS treatment plan and enabling the user to interact with the dose distribution using a mouse to add or remove planned dose from selected regions. This interaction changes the contributing spot weights and displays an updated, deliverable dose distribution in real-time, demonstrating the result of such a change. Once the desired dose distribution has been attained, spot weights are returned to the treatment planning system (TPS).

2 Materials and Methods

2.1 Implementation

The software is implemented in Python v3.10.9 and the results presented here are shown from running on an Intel Xeon 2.50 GHz CPU E5-2680 v3 with 64Gb RAM.

In this proof of principle implementation, initial treatment plans and structure sets are taken from Eclipse v16.1 (Varian Medical Systems). However, the proposed approach is independent of TPS, requiring only that an influence matrix can be exported and that the new spot weights can be imported. In this case, this has been done using the Eclipse scripting application programming interface (ESAPI).

2.2 Functionality

CT (computed tomography) scan, structure set and dose distributions are displayed co-registered along with the dose volume histograms (DVH) of selected structures. Figure 1 shows a representative image of the software.

The user can scroll through the CT scan using the keyboard up and down arrows. Left clicking or scrolling up with the mouse cursor over a given point will modify spot weights to increase the dose to this location. Right clicking, or scrolling down, will reduce the dose. Individual beam doses can be displayed allowing modification of spots from the

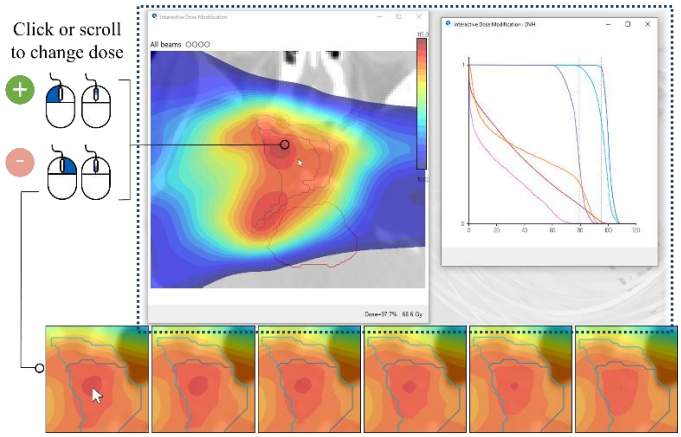


Figure 1. Graphical user interface (within the dashed box) with an example for case 1. Inset shows the effect of successive right clicks on reducing the dose to the selected point. In this example the updated dose distribution is shown within 120 ms of each click.

displayed beam alone.
The available functionality is shown in table 1.

2.3 Influence matrix

An influence matrix, D_{ij}, is extracted from the TPS describing the unit contribution of each spot, j, to each voxel, i, in the dose distribution. This is a large, but sparse array, allowing for sparse data structures to be used to minimise memory requirements.
Using the influence matrix, the dose, d, to a given voxel is given by:

$$d_i = \sum_j D_{ij} w_j \,,$$

where, w_j, is the weight of spot, j.
The influence matrix, and resultant dose distribution has the same resolution as the CT scan (e.g., for case 1, shown in figure 1, this is 0.97×0.97×1 mm).
In this implementation the influence matrix has been reconstructed using ESAPI to calculate the dose distribution for each spot. In some TPS this may be more directly accessible.

2.4 Spot-weight modification

On interaction with the dose distribution, all spots contributing to the interaction site are modified in proportion to their contribution to the interaction site—the weights of spots that have a large influence on the interaction site will change more than those of spots with a low influence on the interaction site. This is done by evaluating the relevant column of the influence matrix and multiplying by a factor that is positive when dose is to be increased and negative when dose is to be decreased to give a modification factor, α_j, for each spot (typically between -0.015 and 0.03). Updated weights for each spot, w_j' are given by:

$$w_j' = (1 + \alpha_j) w_j \,.$$

By default, all spots are considered when interacting with the dose distribution. This behaviour can be adjusted in two

Table 1. Interaction options and functionality

I	**Show isodose labels** displays isodose labels between 10% and 120% of the prescription dose.
S	**Show spot positions** displays spot positions as a circle with a size reflecting the spot weight and centred on the voxel with the highest dose from each spot. Locked or pinned spots are shown with a cross through them.
P	**Place pin** turns the cursor into a crosshair with which left clicking on the dose distribution places a pin which prevents modification of any spot contributing more than a threshold value to this location. The pin position is shown by the symbol 📌 and appears transparent on other slices. Multiple pins may be placed throughout the dose distribution.
[	**Remove all pins**
L	**Lock/unlock beam** when an individual beam is shown, prevents modification of spots from that beam when clicking on the dose distribution with all beams shown. Locked and unlocked beams are indicated on the display.
R	**Renormalise** scales all spot weights such that the mean dose to the normalisation structure matches the prescription dose.
X	**Export spot weights to TPS** spot weights are exported to a file which can be imported in to the TPS using a script.
← →	**Cycle through individual beam doses**
↑ ↓	**Scroll through CT slices**
0	**Restore initial spot weights** assigns the spot weights that were exported from the treatment planning system
Ctrl + Z	**Undo** reverts changes to the spot weights one interaction at a time
Ctrl + 1	**Save to bookmark** saves the current spot weights to a bookmark. Can be used for any number from 1 to 9. Existing bookmarks are shown in the top right of the screen.
1 / 2 ...	**Restore bookmark** assigns spot weights saved to a given bookmark
On dose display:	**Increase dose** when clicking or scrolling within the dose display the dose at the position of the cursor is increased.
On dose display:	**Decrease dose** when clicking or scrolling within the dose display the dose at the position of the cursor is decreased.
On DVH display:	**Reselect ROIs** reselect ROIs for display on the dose display and DVH display and reselect normalisation structure.
On colour bar:	**Set dose display** set the minimum and maximum dose display. Different values are set for the display of all beams and individual beams.
On colour bar:	**Switch between absolute and relative dose**
Ctrl	**Increase effect ×5** holding ctrl while clicking will increase the magnitude of spot weight changes by a factor of 5.
Shift	**Increase effect ×2** holding shift while clicking will increase the magnitude of spot weight changes by a factor of 2. This can be combined with holding ctrl to increase the magnitude of the change by a factor of 10.

ways: by locking individual beams (pressing 'L' when an individual beam is displayed) or by placing a pin in the dose distribution (pressing 'P' and selecting a point in the dose distribution). Locking individual beams prevents changes to spots belonging to those beams when interacting with the total dose distribution. Placing a pin on the dose distribution prevents the modification of any spot that contributes more than a threshold value to the pin position. This minimises the change in dose at the pinned position, allowing the dose at one location to be modified without affecting some other critical part of the dose distribution.

A record of spot weights on each click is stored so that changes can be reverted, and bookmarks can be stored for the current spot weights and restored later.

2.5 Ensuring deliverability

PBS enables different combinations of spot positions and weights to be combined to produce a treatment plan. Spot weights can take any value but are subject to a minimum deliverable weight, $w_{\min}$ which is dependent on the delivery system. In conventional optimisation approaches, this may be accounted for during spot weight optimisation or via post processing to ensure deliverability.

Here, the deliverability of the displayed dose distribution is guaranteed by considering two spot lists: a raw spot list, W_{raw}, and a post-processed spot list W_{proc}. When interacting with the dose distribution, W_{raw}, is used to determine the contribution of each spot to the selected location, and it is W_{raw}, which is updated in response. This spot list is then copied to W_{proc} in which spots below $w_{\min}$ are rounded up to the minimum deliverable spot weight (where they are $> 0.5\, w_{\min}$) or otherwise set to zero. The displayed dose is based on W_{proc}, such that the displayed dose distribution is deliverable.

In W_{raw}, spot weights are limited to $> 0.1\, w_{\min}$. These spots are set to zero during the post processing step (as they are $<0.5\, w_{\min}$) and as such do not contribute to the dose distribution but are not irreversibly removed from W_{raw}.

2.6 Performance optimisation

To ensure that user interaction with the tool is responsive a number of actions are taken. Firstly, dose calculation using the updated spot list is only performed for the displayed CT slice, rather than the whole dose distribution. When the user scrolls to another slice the displayed dose is updated based on the current values. One consequence of this is that although DVHs are displayed, DVH calculation is not performed in the same time frame, as it requires calculation of the dose to each voxel in the selected ROIs. The DVH display is updated whenever the user hovers over the display and is displayed with dashed lines whenever they are outdated. Secondly, a smaller calculation volume, cropped to the relevant volume is used to reduce the size of the influence matrix by reducing the number of voxels considered (see e.g., Figure 1 compared to the dose distribution from the TPS shown for case 1 in table 2). Finally, where possible, operations are parallelised, and the code uses just-in-time compilation using Numba [1] to speed up the dose calculation.

2.7 Integration with existing planning pathways

Once a satisfactory dose distribution is achieved, the modified spot weights can be exported to a file. This file is then read by a script in the TPS which sets the spot weights in the plan to the exported values. Plan specific header information is provided to ensure that the data is correctly associated. From this point, the plan in the TPS is indistinguishable from one produced through conventional optimisation approaches and can proceed to treatment using existing planning pathways.

2.8 Clinical example cases

Proof of principal has been demonstrated using 6 patient cases covering a range of treatment indications. Details are shown in table 2. Patient data were anonymised and used with consent.

3 Results

The response time, measured as the time from registering a click on a point in the dose distribution to displaying the updated dose distribution for each plan is shown in table 2. For the 6 cases evaluated, response times of between 118-364 ms were measured. In all but one case (case 5) the response time is <175 ms such that the interaction to the user is perceived as being near instantaneous.

DVH calculation, which is implemented with the calculation of dose to all voxels that are contained within the selected structures takes longer and is dependent on the total volume of the selected structures.

RAM usage is considerably reduced using sparse arrays. This additionally enables larger dose calculation volumes to be considered with fixed resources.

The software supported fast and intuitive interaction for all cases and modified spot weights were successfully reimported back to the TPS.

4 Discussion

The tool presented here proposes a method of improving PBS proton treatment plans that is both fast and intuitive.

There remain a number of options to improve calculation times. In this proof of principle, calculation is done using CPU and does not exploit GPU acceleration. Implementation in an alternative programming language may also provide some performance benefits.

In this implementation, the dose resolution is matched to the CT resolution, but this could be lowered, reducing the size of the influence matrix, and decreasing calculation times.

The influence matrix used here was derived from an analytical dose calculation in Eclipse. However, this could be provided from alternative sources. Direct access to the influence matrix would also provide important workflow

benefits as its reconstruction through scripting can be a time-consuming process. The dose calculation used for the influence matrix determines the accuracy of the resultant dose distribution. For example, if Monte Carlo was used, the interactive dose modification would operate with equivalent accuracy.

The use of sparse arrays has an impact on response time but significantly reduces memory and loading time. The largest impact was seen in case 5 where the response time increased from 200±28 ms to 364±4 ms with the use of sparse arrays.

Some optimisation approaches, such as the use of deep learning models, can produce treatment plans without the need for objective functions [2]. However, if clinical requirements are not fully met additional postprocessing is required to modify them. Further optimisation approaches that do not use objective functions could also be envisioned.

Other interactive approaches to planning have been demonstrated in clinical practice. Graphical optimisation has been demonstrated in brachytherapy treatment planning where users can drag isodose lines and the required dwell times for brachytherapy sources is calculated [3]. Compared to PBS, where there may be several thousand spots, there are typically a more limited number of dwell positions in this context as well as there being a simpler description of dose distribution making the problem for brachytherapy more tractable than in proton therapy.

Multi-criteria optimisation (MCO) approaches have also been demonstrated [4] with the use of sliders to adjust plans and explore planning trade-offs. In contrast to MCO, the approach proposed here does not require specific structures or objectives to be defined up front and does not require calculation of a Pareto surface. It allows for more direct control over how and where a plan is changed.

By returning the modified spot weights back to the treatment planning system, the existing treatment planning framework including critical safety checks, plan evaluation, robustness analysis, and treatment preparation can be used.

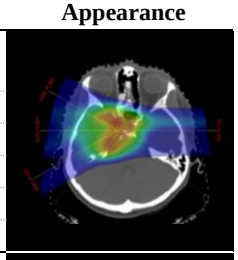
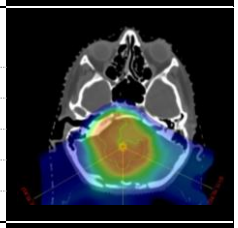
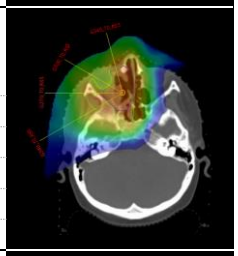
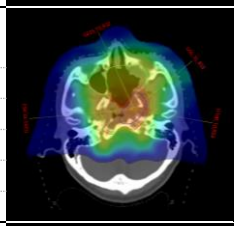
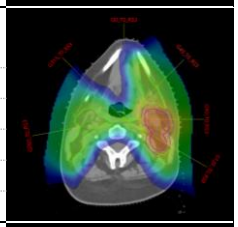
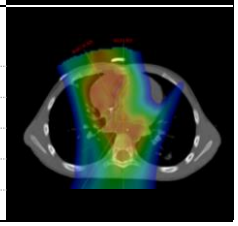
Future development will focus on: evaluating and quantifying potential time savings or improvements in plan quality; assessing user experience; exploring at what stage in the planning process it may be most effective to employ such a tool; and combination with automated planning approaches to enable such technologies to be clinically implemented more widely.

5 Conclusion

A proof of principle implementation of an interactive dose modification tool has been presented. It is capable of changing existing PBS proton treatment plans both quickly and intuitively.

This approach may complement existing automated planning approaches to simplify treatment planning as well as enabling the introduction of automation approaches that do not yet exceed the capabilities of human generated plans and require some modification.

Table 2. Patient cases. Response time reflects the time from registering a click in the dose distribution to displaying the updated dose. Transverse slices from the treatment planning system are shown. RAM usage reflects the peak value during use of the software.

ID			Appearance
1	Indication	Base of skull chondrosarcoma	
	Number of spots	1225	
	Response time	118 ± 1 ms	
	Sparse array size	250 Mb	
	RAM usage	1720 Mb	
	No. of voxels	111(x)×96(y)×76(slices)	
2	Indication	Posterior fossa ependymoma	
	Number of spots	2335	
	Response time	134 ± 4 ms	
	Sparse array size	791 Mb	
	RAM usage	4600 Mb	
	No. of voxels	131×115×120	
3	Indication	Embryonal rhabdomyosarcoma of the maxillary sinus	
	Number of spots	4076	
	Response time	157 ± 3 ms	
	Sparse array size	793 Mb	
	RAM usage	5320 Mb	
	No. of voxels	132×137×57	
4	Indication	Adenocarcinoma of the sphenoid sinus	
	Number of spots	2444	
	Response time	122 ± 1 ms	
	Sparse array size	526 Mb	
	RAM usage	2210 Mb	
	No. of voxels	106×106×38	
5	Indication	Nasopharyngeal carcinoma	
	Number of spots	16942	
	Response time	364 ± 4 ms	
	Sparse array size	2070 Mb	
	RAM usage	21770 Mb	
	No. of voxels	168×129×88	
6	Indication	Mediastinal Rhabdomyosarcoma	
	Number of spots	6109	
	Response time	172 ± 2 ms	
	Sparse array size	1210 Mb	
	RAM usage	8620 Mb	
	No. of voxels	121×139×64	

References

[1] S.K. Lam, A. Pitrou, S. Seibert, et al. “Numba: a LLVM-based Python JIT compiler”. Proceedings of the 2nd workshop on the LLVM compiler infrastructure in HPC (2015), pp. 1-6. DOI:10.1145/2833157.2833162

[2] M. Wang, Q. Zhang, S. Lam, et al. “A review on application of deep learning algorithms in external beam radiotherapy automated treatment planning” Front Oncol 10 (2023) DOI:10.3389/fonc.2020.580919

[3] B. Tang, X. Liu, X. Wang, et al. “Dosimetric comparison of graphical optimization and inverse planning simulated annealing for brachytherapy of cervical cancer”. J Contemp Brachytherapy. 11(4) (2019) pp. 379-383. DOI:10.5114/jcb.2019.87145

[4] W. Chen, D. Craft, T.M. Madden, et al. “A fast optimization algorithm for multicriteria intensity modulated proton therapy planning”. Med Phys 37(9) (2010) pp. 4938-4945. DOI:10.1118/1.3481566

Simultaneous Reduction of Number of Spots and Energy Layers in Intensity Modulated Proton Therapy for Rapid Spot Scanning Delivery

Anqi Fu[1], Vicki T. Taasti[2], and Masoud Zarepisheh[1]

[1]Department of Medical Physics, Memorial Sloan Kettering, New York, NY, USA
[2]Department of Radiation Oncology, Maastricht University Medical Center, Maastricht, NL

Abstract Reducing proton therapy delivery time can improve patient comfort, but must be balanced against clinical goals and treatment plan quality. We propose a method to improve the delivery efficiency of proton therapy by simultaneously reducing the number of spots and energy layers using reweighted l_1 regularization. We formulate the treatment planning problem as a convex optimization problem with an objective consisting of a dosimetric plan quality term plus a weighted l_1 regularization term. We then iteratively solve this problem and adaptively update the regularization weights to promote sparsity of the spots and layers. To assess our method's effectiveness, we compare its performance on four head-and-neck cancer patients against standard l_1 and group l_2 regularization methods. Reweighted l_1 regularization reduces the number of spots and energy layers by 40% and 35%, respectively, with minimal cost to dosimetric plan quality. In addition, it provides a better trade-off between plan delivery efficiency and plan quality than standard l_1 or group l_2 regularization, requiring the lowest cost to quality to achieve a given level of efficiency.

1 Introduction

During intensity modulated proton therapy (IMPT), protons are delivered spot-by-spot within each energy layer and layer-by-layer with each beam. The total plan delivery time is approximately equal to the total energy layer switching time, spot travel time, and dose delivery time [1, 2]. We seek to reduce IMPT delivery time by reducing the number of spots and energy layers. A shorter treatment time improves patient comfort, increases patient throughput, and decreases the risk of error due to intra-fractional motion. At the same time, we want to ensure the quality of the treatment plan remains uncompromised.

One approach to this problem is to solve the proton treatment planning problem with a regularization function that encourages sparsity (i.e., zero values) in the spot intensities and energy layers. Common sparsity-inducing regularizers include the l_1 norm [3–5] and the group l_2 norm [1, 6–8]. However, l_1 regularization only promotes sparsity of the spots, but not the energy layers, while group l_2 regularization promotes sparsity of the layers, but results in a denser spot vector. In this study, we propose a reweighted l_1 regularization method that *simultaneously* reduces the number of nonzero spots and energy layers, while upholding dosimetric plan quality and providing a better trade-off between plan quality and delivery efficiency than either l_1 or group l_2 regularization.

2 Materials and Methods

2.1 Problem Formulation

Let m be the number of voxels in the patient's body, n the number of spots, and G the number of energy layers. For a given spot intensity vector $x \in \mathbf{R}^n$, the dose delivered per voxel is $d = Ax$, where $A \in \mathbf{R}_+^{m\times n}$ is the dose influence matrix. We want to choose x so the deviation between d and a prescribed dose $p \in \mathbf{R}_+^m$ is low; this deviation is represented by the plan *cost function*

$$f(\overline{d},\underline{d}) = \overline{w}\|\overline{d}\|_2^2 + \underline{w}\|\underline{d}\|_2^2, \tag{1}$$

where $\overline{d} := (d-p)_+$ is the overdose, $\underline{d} = (d-p)_-$ is the underdose, and $\overline{w}, \underline{w} \in \mathbf{R}_+^n$ are weight parameters.
Simultaneously, we wish to reduce the plan delivery time (i.e., increase delivery efficiency), which is positively correlated with the number of nonzero spots and energy layers in the treatment plan [1, 9]. Thus, we introduce a *regularization function* $r : \mathbf{R}^n \to \mathbf{R}$ that penalizes x in a way that decreases the number of nonzero spots/layers, and define the proton treatment planning problem to be

$$\begin{aligned} \text{minimize} \quad & f(\overline{d},\underline{d}) + \lambda r(x) \\ \text{subject to} \quad & Ax - \overline{d} + \underline{d} = p, \quad Bx \le c \\ & x \ge 0, \quad \overline{d} \ge 0, \quad \underline{d} \ge 0 \end{aligned} \tag{2}$$

with respect to $x, \overline{d}$, and $\underline{d}$, where B and c are constants used to express linear dose constraints (e.g., mean/max dose upper bounds). The regularization parameter $\lambda \ge 0$ controls the trade-off between dosimetric plan quality, represented by plan cost $f(\overline{d},\underline{d})$, and delivery efficiency, captured by the regularization term $r(x)$. A larger value of λ places more importance on efficiency.

2.2 Reweighted l_1 Method

An ideal choice of r would be the l_0 norm, or cardinality function, which directly penalizes the number of nonzero spots or energy layers. Unfortunately, the cardinality is nonconvex and computationally expensive to evaluate. Popular convex approximations also have flaws: the l_1 norm promotes sparsity of the spots, but not the energy layers, while the group l_2 norm promotes sparsity of the layers, but encourages even denser spots.

In this section, we introduce the reweighted l_1 regularization method [10], which simultaneously promotes sparsity in both the spots and the energy layers. Let $\mathscr{I}_g$ be the set of indices of the spots in layer g. Define the *weighted group l_1 regularizer*

$$r_3(x;\beta) = \sum_{g=1}^{G} \beta_g \sum_{i\in\mathscr{I}_g} |x_i| \tag{3}$$

with weight parameters $\beta_g \in \mathbf{R}_+ \cup \{+\infty\}$ for $g = 1,\ldots,G$. An intuitive way to set β_g is to make it inversely proportional to the optimal intensity of energy layer g, so that each spot contributes about equally to the total regularization penalty, similar to the cardinality regularizer. However, since we do not know the optimal layer intensity, we will approximate this weighting scheme via an iterative method.
Let the initial weights $\beta^{(1)} = \mathbf{1}$. At each iteration $k = 1,2,\ldots,$

1. Set $x^{(k)}$ equal to a solution of Problem 2 with $r(x) := r_3(x;\beta^{(k)})$.

2. Calculate the total intensity of each energy layer $e_g^{(k)} = \sum_{i\in\mathscr{I}_g} x_i^{(k)}$, where $\mathscr{I}_g$ is the set of indices of the spots in layer g.

3. Apply a lower threshold to the energy layer intensities to get

$$\tilde{e}_g^{(k)} = \max(e_g^{(k)}, \varepsilon^{(k)}), \quad g = 1,\ldots,G,$$

where $\varepsilon^{(k)} = \delta \cdot \max_{g'} e_{g'}^{(k)}$ for some small $\delta \in (0,1)$.

4. Update the weights. First, compute the standardized reciprocals

$$\alpha_g^{(k)} = \left(\frac{1}{\tilde{e}_g^{(k)}}\right) \Bigg/ \left(\sum_{g'=1}^{G} \frac{1}{\tilde{e}_{g'}^{(k)}}\right), \quad g = 1,\ldots,G. \tag{4}$$

Then, calculate the scaling term

$$\lambda^{(k)} = \sum_{g=1}^{G} \tilde{e}_g^{(k)} \Bigg/ \left(\sum_{g=1}^{G} \alpha_g^{(k)} \tilde{e}_g^{(k)}\right). \tag{5}$$

The new weights are $\beta_g^{(k+1)} = \lambda^{(k)} \alpha_g^{(k)}$.

5. Terminate on convergence of the objective, or when k reaches a maximum number of iterations.

In practice, the number of reweighting iterations needed is typically very low, with most of the improvement coming from the first $k = 2$ to 3 iterations.

2.3 Computational Framework

We implemented the l_1, group l_2, and reweighted l_1 treatment planning methods in Python using CVXPY [11–13] and solved the optimization problems with MOSEK [14]. Then, we compared the performance of the methods on four head-and-neck patients from The Cancer Imaging Archive (TCIA) [15, 16]. For the sake of brevity, we present detailed figures for a single representative patient (2) below, and summarize results for all patients in Tables 1 and 2 at the end of the paper.

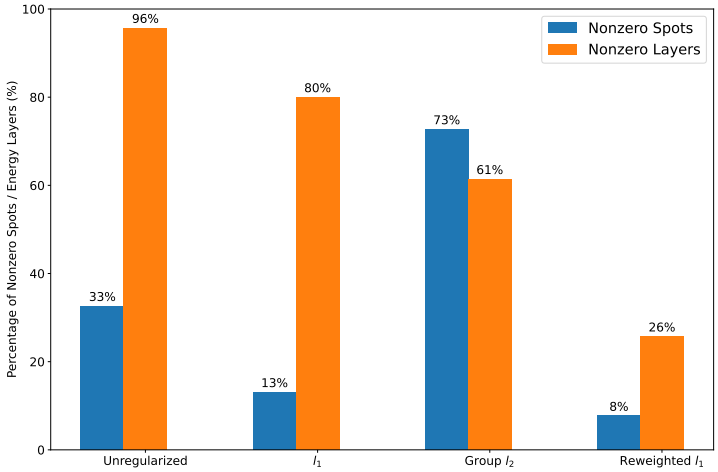


Figure 1: Percentage of nonzero spots/energy layers (relative to the total) for the unregularized model and the l_1, group l_2, and reweighted l_1 regularized models ($\lambda = 5$).

3 Results

3.1 Reduction in Spots and Energy Layers

Figure 1 summarizes the results from the different regularization methods using $\lambda = 5$. Without regularization, one-third of the total 7011 spots and nearly all of the 70 energy layers are nonzero. Standard l_1 regularization drastically sparsifies the spot intensity vector, but the percentage of nonzero layers remains high at 80%, while group l_2 regularization does the opposite, decreasing the number of nonzero energy layers, while significantly *increasing* the number of nonzero spots to 73%. Only reweighted l_1 regularization reduces both the number of nonzero spots and energy layers, resulting in a treatment plan with just $541(8\%)$ nonzero spots and $18(26\%)$ nonzero layers – significantly lower than all the other methods.

3.2 Delivery Efficiency versus Plan Quality

To analyze the trade-off between delivery efficiency and treatment plan quality, we plot the relative (percentage-wise) change in the number of nonzero spots/energy layers versus the relative change in plan cost (Equation 1) with respect to the unregularized solution. These Pareto optimal trade-off curves are shown in Figures 2 and 3 for l_1, group l_2, and reweighted l_1 regularization. In Figure 2, both l_1 and reweighted l_1 curves drop sharply from the origin, attaining a $> 45\%$ decrease in nonzero spots for a $< 10\%$ increase in plan cost, with reweighted l_1 outperforming l_1 by about 5 percentage points. Group l_2 regularization performs worst

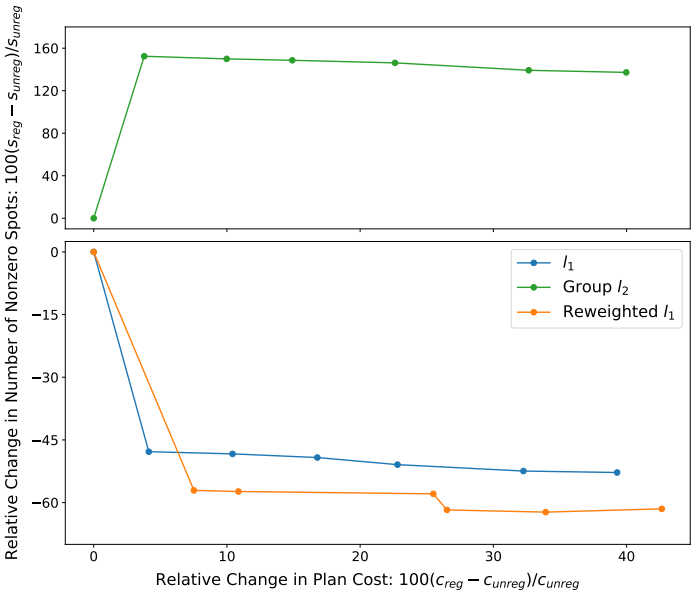


Figure 2: Relative change in number of nonzero spots versus relative cost to plan quality with respect to the unregularized plan. Each point corresponds to a different value of λ for the specified regularization method; the origin is $\lambda = 0$.

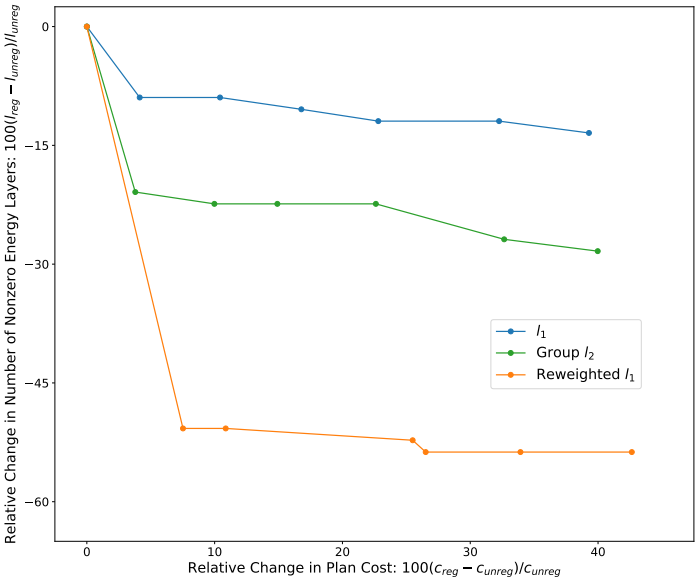


Figure 3: Relative change in number of nonzero energy layers versus relative cost to plan quality with respect to the unregularized plan.

of all, increasing both the number of nonzero spots and plan cost, as indicated by the inverted trade-off curve.

All three regularization methods reduce the number of nonzero energy layers with modest cost to plan quality, as depicted in Figure 3. For an 8% to 10% increase in plan cost, l_1 obtains a roughly 8% drop and group l_2 a 20% drop in nonzero layers. However, reweighted l_1 significantly outperforms both these regularizers, achieving a $> 50\%$ decrease in the number of nonzero layers for a $< 10\%$ increase in plan cost. Overall, it is clear the reweighted l_1 method provides the best trade-off between delivery efficiency and treatment plan quality. The resulting dose-volume histogram (DVH) curves (Figure 4) are similar to the unregularized plan, obtained with only a fraction of the number of nonzero spots and energy layers.

4 Discussion

In this brief paper, we have not presented calculations of the plan delivery time because it requires machine-specific parameters such as dose rate, energy layer switching time, and spot travel time. We have also not enforced the minimum-monitor-unit constraint, although this can be done by combining reweighted l_1 regularization with the two-step optimization method described in [8]. In general, our reweighted l_1 method is agnostic to the optimization framework and can be used in conjunction with an automation tool (e.g., hierarchical optimization [17]). It can also be integrated into existing techniques in the literature to handle DVH constraints [18–20] and plan robustness requirements [21, 22], making it a highly flexible algorithm.

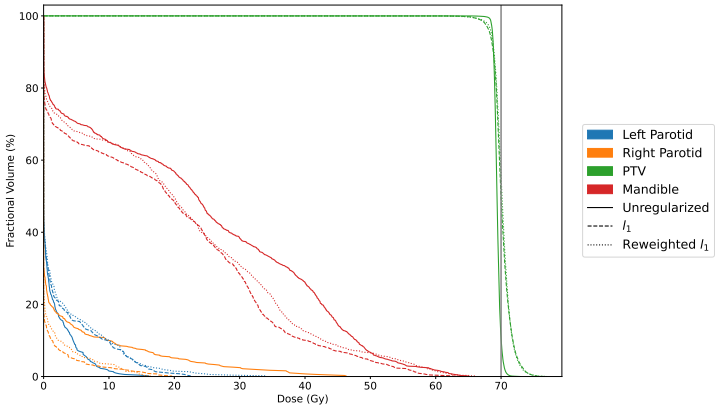


Figure 4: DVH curves obtained from the unregularized model (solid), and the standard l_1 (dashed) and reweighted l_1 (dotted) models regularized to 10% relative cost to plan quality (with $\lambda = 1.325$ and $\lambda = 0.3$, respectively). The vertical gray line indicates the prescription $p = 70$ Gy. Group l_2 curves were omitted because they largely overlap with l_1, but the corresponding plan demands a $> 55\%$ increase in the number of nonzero spots.

5 Conclusion

The reweighted l_1 regularization method is capable of simultaneously reducing the number of spots and energy layers in a proton treatment plan, while imposing minimal cost to dosimetric plan quality and achieving a better trade-off between delivery efficiency and plan quality than standard l_1 and group l_2 regularization.

Supplemental Data

Patient	l_1	Group l_2	Reweighted l_1
1	−38%	40%	−41%
2	−49%	150%	−57%
3	−30%	80%	−33%
4	−24%	91%	−27%

Table 1: Relative change in number of nonzero spots at a relative change in plan cost of 10% with respect to the unregularized plan.

Patient	l_1	Group l_2	Reweighted l_1
1	−5%	−5%	−37%
2	−8%	−23%	−50%
3	−6%	−9%	−31%
4	−7%	−5%	−25%

Table 2: Relative change in number of nonzero energy layers at a relative change in plan cost of 10% with respect to the unregularized plan.

References

[1] H. Gao, B. Clasie, M. McDonald, et al. "Technical Note: Plan-Delivery-Time Constrained Inverse Optimization Method with Minimum-MU-per-Energy-Layer (MMPEL) for Efficient Pencil Beam Scanning Proton Therapy". *Medical Physics* 47.9 (Sept. 2020), pp. 3892–3897.

[2] G. Zhang, H. Shen, Y. Lin, et al. "Energy Layer Optimization via Energy Matrix Regularization for Proton Spot-Scanning Arc Therapy". *Medical Physics* 49.9 (Sept. 2022), pp. 5752–5762. DOI: 10.1002/mp.15836.

[3] R. Tibshirani. "Regression Shrinkage and Selection via the Lasso". *Journal of the Royal Statistical Society: Series B (Methodology)* 58.1 (1996), pp. 267–288.

[4] D. Vidaurre, C. Bielza, and P. Larrañaga. "A Survey of L_1 Regression". *International Statistical Review* 81.3 (Dec. 2013), pp. 361–381.

[5] T. Hastie, R. Tibshirani, and M. Wainwright. *Statistical Learning with Sparsity: The Lasso and Generalizations*. Monographs on Statistics and Applied Probability. Taylor and Francis, 2015.

[6] M. Yuan and Y. Lin. "Model Selection and Estimation in Regression with Grouped Variables". *Journal of the Royal Statistical Society: Series B (Methodology)* 68.1 (Feb. 2006), pp. 49–67. DOI: 10.1111/j.1467-9868.2005.00532.x.

[7] M. F. Jensen, L. Hoffmann, J. B. B. Petersen, et al. "Energy Layer Optimization Strategies for Intensity-Modulated Proton Therapy of Lung Cancer Patients". *Medical Physics* 45.10 (Oct. 2018), pp. 4355–4363. DOI: 10.1002/mp.13139.

[8] Y. Lin, B. Clasie, T. Liu, et al. "Minimum-MU and Sparse-Energy-Layer (MMSEL) Constrained Inverse Optimization Method for Efficiently Deliverable PBS, Plans". *Physics in Medicine and Biology* 64.20 (2019), p. 205001. DOI: 10.1088/1361-6560/ab4529.

[9] P. R. Poulsen, J. Eley, U. Langner, et al. "Efficient Interplay Effect Mitigation for Proton Pencil Beam Scanning by Spot-Adapted Layered Repainting Evenly Spread out Over the Full Breathing Cycle". *International Journal of Radiation Oncology, Biology, Physics* 100.1 (Jan. 2018), pp. 226–234.

[10] E. J. Candés, M. B. Wakin, and S. P. Boyd. "Enhancing Sparsity by Reweighted l_1 Minimization". *Journal of Fourier Analysis and Applications* 14.1 (2008), pp. 877–905. DOI: 10.1007/s00041-008-9045-x.

[11] S. Diamond and S. Boyd. "CVXPY: A Python-embedded modeling language for convex optimization". *Journal of Machine Learning Research* 17.83 (2016), pp. 1–5.

[12] A. Agrawal, R. Verschueren, S. Diamond, et al. "A Rewriting System for Convex Optimization Problems". *Journal of Control and Decision* 5.1 (2018), pp. 42–60.

[13] S. Diamond, A. Agrawal, R. Murray, et al. *CVXPY: Disciplined Convex Programming in Python, Version 1.3*. 2022.

[14] M. ApS. *The MOSEK Optimizer API for Python, Version 10.0*. 2022.

[15] K. Clark, B. Vendt, K. Smith, et al. "The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository". *Journal of Digital Imaging* 26.6 (2013), pp. 1045–1057.

[16] T. Nolan. *Head-and-Neck Squamous Cell Carcinoma Patients with CT Taken During Pre-Treatment, Mid-Treatment, and Post-Treatment (HNSCC-3DCT-RT)*. DOI: 10.7937/K9/TCIA.2018.13upr2xf.

[17] M. Zarepisheh, L. Hong, Y. Zhou, et al. "Automated and Clinically Optimal Treatment Planning for Cancer Radiotherapy". *INFORMS Journal on Applied Analytics* 52.1 (2022), pp. 69–89.

[18] A. Fu, B. Ungun, L. Xing, et al. "A Convex Optimization Approach to Radiation Treatment Planning with Dose Constraints". *Optimization and Engineering* 20 (2019), pp. 277–300.

[19] M. Zarepisheh, M. Shakourifar, G. Trigila, et al. "A Moment-Based Approach for DVH-Guided Radiotherapy Treatment Plan Optimization". *Physics in Medicine and Biology* 58.6 (2013), pp. 1869–1887.

[20] S. Mukherjee, L. Hong, J. O. Deasy, et al. "Integrating Soft and Hard Dose-Volume Constraints into Hierarchical Constrained IMRT Optimization". *Medical Physics* 47.2 (2020), pp. 414–421.

[21] A. Fu, V. T. Taasti, and M. Zarepisheh. "Distributed and Scalable Optimization for Robust Proton Treatment Planning". *Medical Physics* 50.1 (Jan. 2023), pp. 633–642. DOI: 10.1002/mp.15897.

[22] V. T. Taasti, L. Hong, J. O. Deasy, et al. "Automated Proton Treatment Planning with Robust Optimization Using Constrained Hierarchical Optimization". *Medical Physics* 47.7 (2020), pp. 2779–2790.

Ameliorate immune suppression during radiation therapy to advanced stage lung cancer via a predictive algorithm

C Nguyen[1], G Williams[1], M Baessler[1], C E Hodge[2], J T Morgan[2], R Gentzler[2], N Liyanage[1], and K Wijesooriya [1,2]

[1]Department of Physics, University of Virginia, Charlottesville, USA
[2]Department of Radiation Oncology, University of Virginia, Charlottesville, USA

Abstract

Purpose

Radiation Therapy (RT) is known to modulate the immune system and contribute to the generation of anti-tumor T cells stimulating T cell infiltration into tumors. However, this anti-tumor activity is offset by radiation-induced immunosuppression (RIIS). Lymphocytes, (80% are T cells) are highly radiosensitive and RIIS means destroying existing as well as newly created T lymphocytes. Optimizing RT treatment planning by considering circulating blood and lymphatics as a critical organ at risk may mitigate RIIS and lead to the preservation of cytotoxic T lymphocytes converting the tumor from an immunologically cold to a hot environment increasing the efficacy of immunotherapy. In this paper, we present an algorithm to predict RIIS based on a model of circulating blood and lymphatics using advanced stage lung cancer patients treated with radiation therapy and its utilization in ameliorating immune suppression.

Methods

We have developed a Python-based algorithm that uses DICOM data for radiation therapy treatment plans, dose maps, patient CT data sets, and organ delineations to stochastically simulate blood flow and lymphatic flow and predict the doses absorbed by circulating lymphocytes. These absorbed doses are used to predict the fraction of lymphocytes killed by a given treatment plan. This predictive model was used on a cohort of 35 retrospective advanced-stage NSCLC patients to reduce immune suppression via treatment planning. Clinical Goals were created per LU 001.

Results

The RIIS predictive model had 78% of the patients having a immune cell kill difference between the prediction and the measurement less than $0.2x10^9$ cells/L.

Reduction of immune suppression via re-planning with the aid of the predictive algorithm showed varying amounts of immune system sparing and as much as 21.4% (kill % decreases from 78.2% to 56.8%).

Conclusions

Our model performed well in predicting RIIS in patients treated using standard fractionation to lung. It has the capability to be interfaced with treatment planning systems to prospectively reduce immune cell toxicity while maintaining national LU001 conformity and plan quality criteria.

1 Introduction

Approximately one third of patients with non–small-cell lung cancer (NSCLC) have stage III, locally advanced disease at diagnosis (1). The standard of care for patients with a good performance status and unresectable stage III NSCLC is platinum-based doublet chemotherapy concurrent with radiotherapy (CRT)(2). However, the median progression-free survival among patients who have received CRT is poor (approximately 6 months), and only 15% of patients are alive at 5 years (1)(3). PACIFIC trial (4)(5) was a phase 3 study which looked at the efficacy of immunotherapy (Durvalumab), an anti-PDL1 antibody, after CRT in Stage III non-small-cell lung cancer. With 709 patients enrolled in the trial, they observed a markedly increase in progression-free survival with Durvalumab compared to placebo with a median survival of (16.8 months (95% confidence interval [CI], 13.0 to 18.1) with Durvalumab versus 5.6 months (95% CI, 4.6 to 7.8) with placebo) changing the standard of care from CRT alone to immunotherapy following CRT. The PACIFIC trial demonstrated an OS advantage for all patients, regardless of PD-L1 status, with the addition of durvalumab (6).

Ionizing radiation (IR) has a variety of acute and long-lasting adverse effects on the immune system. Lymphocytes are among the cell types most sensitive to radiation-induced injury. It is well known that T cell death occurs within hours after irradiation (interphase death) (7)(8), and that IR also has long-term adverse effects on T cell immunity (9)(10)(11)(12).

The purpose of this work was to develop a model that predicts RIIS for advanced stage lung cancer treated with radiation therapy by estimating the dose delivered to circulating blood lymphocytes. We aimed to develop this model to interface with treatment planning systems (based on DICOM format) and generate fast predictions to augment decision making during treatment planning. Reduction of immune suppression via a predictive algorithm will benefit the treatments of immunotherapy in combination with radiotherapy.

2 Materials and Methods

Simulation Study: We have developed a comprehensive model, using Python software, to predict the immune suppression from lung SBRT on a patient specific manner. This predictive model is dynamic with the blood circulation intertwined with the lymphatic flow for all organs encompassing a region of interest up to a very low threshold dose of 40 cGy per fraction. It has been reported that 50 cGy is a threshold for lymphocyte kill (13) The interplay between time dependent RT delivery and the movement of blood and lymphatics, and lymphocyte osmosis between the blood flow, primary (bone marrow, thymus)/secondary lymphoid organs (lymph nodes, spleen, lymphatic vessels) as well as non-lymphoid organs (liver, lung, etc) were modeled as discussed below based on literature (14 – 19). After each radiation fraction, dose accumulation to lymphocytes in all above organs including circulating blood is modeled as follows. For each patient

included in the study, radiation therapy treatment plans, CT planning image sets, dose maps per each treatment field, delivery times, and the structure sets were accessed and imported into the simulation model via DICOM import to obtain the organ specific doses for each voxel of that organ. The dose delivered to the organs was obtained by applying a logical mask to the delivered dose map, the dose for each of these structures was applied while avoiding leakage into other organs. Blood circulation in the human body is a much faster process than lymphatic flow with very different cardiac outputs (20) and flow velocities for organs such as heart, aorta, vena cava, pulmonary artery, lung, liver, and other organs. This was simulated using the following: (i) varying blood flow levels to organs (20), great vessels, and heart (21), total blood volume of 5L (22) were taken from published literature. (ii) each treatment field was delivered within a variable time depending on the total delivered beam on time for that field and the dose rate of the energy used. In the model the dose was applied in time steps to the blood matrix and the blood matrix was rotated to simulate blood flow through the body. Following each treatment field blood was randomly mixed with the remaining blood. The circular shifting of the blood matrix using cross sectional area provided a method to simulate blood flow through complex organ shapes while ensuring accuracy over the course of an entire treatment. We have modeled the lymphocyte osmosis from organs to blood in the non-equilibrium state by a set of equations that is updated every 6 hrs. The flow rates (k_(blood,) k_s, k_BM,k_LN) are fractional flow rates of lymphocytes going away from the organs: blood, spleen, bone marrow, other organs, and lymph nodes. Since these are fractional flow rates the actual flow rates depend on the number of lymphocytes residing in a given organ at any given time. Therefore, this defines the time dependence on lymphocytes following RT. Depending on the organs that receive lymphotoxic doses, other organ's flow rates will try to establish equilibrium conditions. We used a linear-quadratic model (LQ model), of the form: K(D)=1-exp (-α D-β D^2) for lymphocyte kill with radiation. The two parameters, α and β, were fixed using the invitro measurements by Nakamura (23). This model predicts 15% kill for a dose of 0.5 Gy, 67% kill for a dose of 2 Gy, and 90% kill for a dose of 3 Gy. Taking into account all the radiation losses of lymphocytes, and the rate of addition of new lymphocytes into the blood stream, the model calculates a prediction for the number of lymphocytes left in the blood stream at any given time. Finally, a X^2 minimization was performed on the difference between the simulated and measured percent lymph kill to tune the model.

Replanning to Reduce RIIS Study: For the above advanced-stage NSCLC patients, new radiation therapy plans were created to spare the dose to aorta, thoracic spine, vena cava, pulmonary arteries, heart, and lymph nodes with the aid of the predictive algorithm. Dosimetric comparisons were made between the 3D, IMRT, VMAT plans that were delivered without consideration to the blood rich and immune rich organs, and the plans we designed to intentionally reduce dose to organs at risk that are important in reducing radiation induced immune suppression. Clinical Goals were created per NRG LU 001. RIIS reduction comparisons were quantified using the predictive algorithm.

3 Results

Simulation Study: Two examples of RIIS prediction to measurement in advanced stage lung cancer are given below. i) Top Figure 1 is a PTV of 274 cc with 200 cGy x 33 fractions leading to a post treatment absolute lymphocyte count (ALC) of 0. 37x10^9 cells/L, while the prediction was for a 0.53x10^9 cells/L. ii) Bottom Figure 1 is a PTV of 818 cc with 250 cGy x 15 fractions leading to a post treatment absolute lymphocyte count (ALC) of 0. 93x10^9 cells/L, while the prediction was for a 0.81x10^9 cells/L.

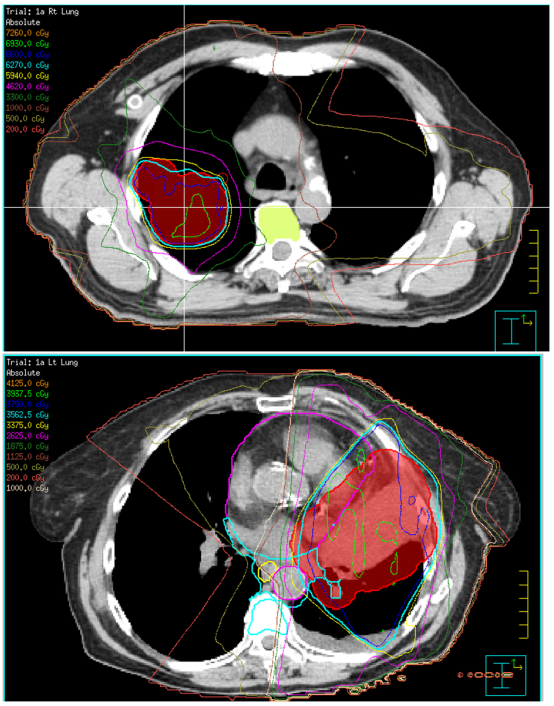


Figure 1: PTV comparison between two advanced-stage lung cancer patients.

Replanning to Reduce RIIS Study: In all cases, whether it is 3D dosimetry or IMRT, there is significant sparing of blood rich organs such as heart (in pink), lymph rich organ such as lymph nodes (in purple), and bone marrow rich thoracic spine (in green). In Figure 2 below, new optimized plan is in the left hand side (and squares in DVH), and the original plan is in the right hand side (triangles in DVH).

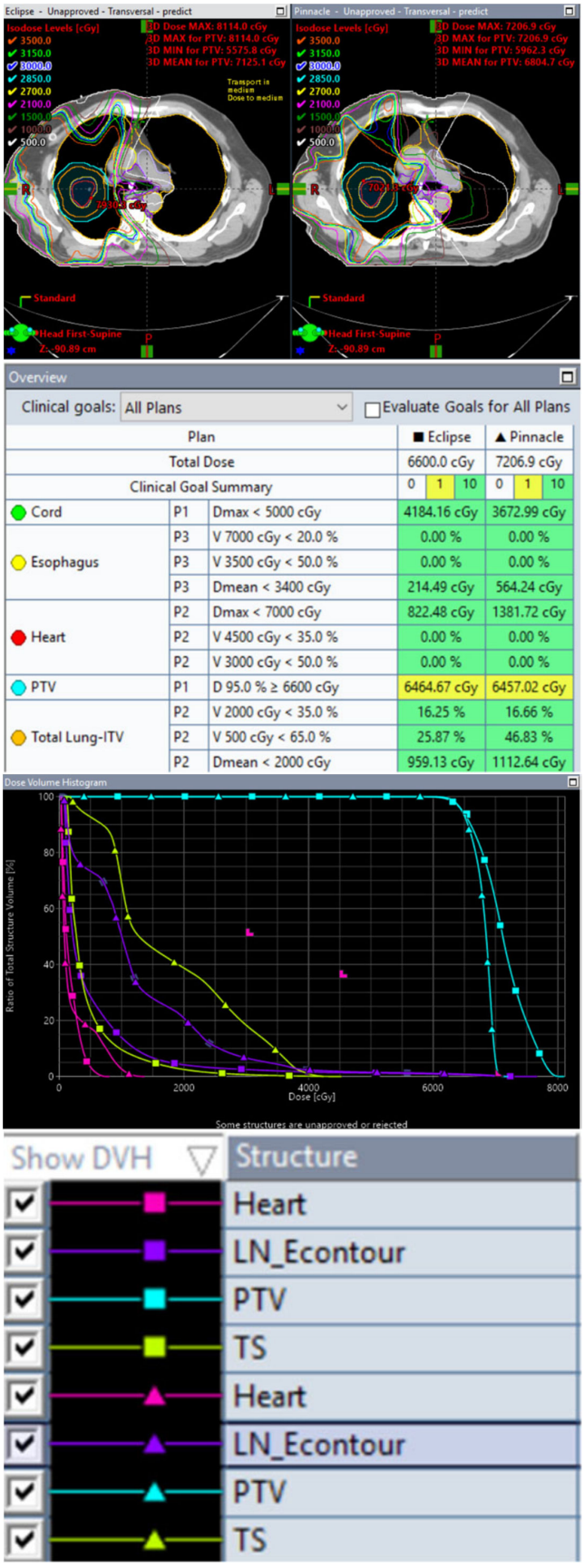


Figure 2: Dosimetric sparing of immune rich OAR in advanced stage lung cancer

Reduction of immune suppression via re-planning with OAR optimization criteria lead to varying amounts of immune system sparing according to the predictive algorithm. In some cases as much as 21.4% decrease was observed (kill % decrease from 78.2% to 56.8%) (Figure 3).

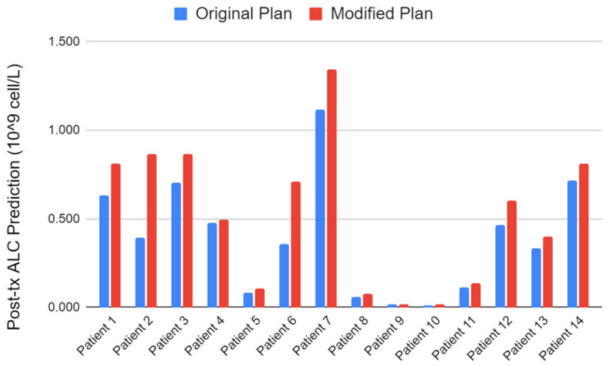


Figure 3: Reduction of immune suppression via re-planning

4 Discussion

In this work, we have presented a model which predicts radiation-induced lymphocyte reduction following radiation therapy to lung by simulating the dose delivered to circulating lymphocytes in the bloodstream and lymphatics. This approach is patient-specific, using three-dimensional dose maps from treatment planning software, varied treatment times, and organ-specific blood flows. Our model predicts initial lymphocyte death using a linear-quadratic survival curve from literature The model could accurately predict post-treatment ALC with 78% of patients having an error below 0.2×10^9 cells/L.

5 Conclusion

We present a model that accurately predicts lymphocyte depletion following RT to Lung as well as the onset of lymphopenia in a cohort of 35 patients under ten-fold cross validation. This model has the capability to be interfaced with treatment planning systems to prospectively reduce immune cell toxicity without compromising treatment efficacy during treatment planning.

References

[1] 1Aupérin A, Le Péchoux C, Rolland E, Curran WJ, Furuse K, Fournel P, et al. Meta-analysis of concomitant versus sequential radiochemotherapy in locally advanced non - small-cell lung cancer. J Clin Oncol. 2010;

[2] Yoon SM, Shaikh T, Hallman M. Therapeutic management options for stage III non-small cell lung cancer. World Journal of Clinical Oncology. 2017.
[3] Ahn JS, Ahn YC, Kim JH, Lee CG, Cho EK, Lee KC, et al. Multinational randomized phase III trial with or without consolidation chemotherapy using docetaxel and cisplatin after concurrent chemoradiation in inoperable stage III non-small-cell lung cancer: KCSG-LU05-04. J Clin Oncol.2015;
[4] Antonia SJ, Villegas A, Daniel D, Vicente D, Murakami S, Hui R, et al. Overall survival with durvalumab after chemoradiotherapy in stage III NSCLC. N Engl J Med. 2018;
[5] Antonia SJ, Villegas A, Daniel D, Vicente D, Murakami S, Hui R, et al. Durvalumab after chemoradiotherapy in stage III non–small-cell lung cancer. N Engl J Med. 2017;
[6] Antonia SJ, Villegas A, Daniel D, et al: Overall survival with durvalumab after chemoradiotherapy in stage III NSCLC. New Engl J Med 379:2342-2350, 2018.
[7] Radford IR. Radiation response of mouse lymphoid and myeloid cell lines. Part I. Sensitivity to killing by ionizing radiation, rate of loss of viability, and cell type of origin. Int J Radiat Biol.1994;
[8] Palayoor ST, Macklis RM, Bump EA, Coleman CN. Modulation of Radiation-Induced Apoptosis and G 2 /M Block in Murine T-Lymphoma Cells. Radiat Res. 1995;
[9] Kusunoki Y, Hayashi T. Long-lasting alterations of the immune system by ionizing radiation exposure: Implications for disease development among atomic bomb survivors. International Journal of Radiation Biology. 2008.
[10] Hayashi T, Kusunoki Y, Hakoda M, Morishita Y, Kubo Y, Maki M, et al. Radiation dosedependent increases in inflammatory response markers in A-bomb survivors. Int J Radiat Biol.2003;
[11] Neriishi K, Nakashima E, Delongchamps RR. Persistent subclinical inflammation among A-bomb survivors. Int J Radiat Biol. 2001;
[12] Neriishi K, Akiba S, Amano T, Ogino T, Kodama K. Prevalence of Hepatitis B Surface Antigen, Hepatitis B e Antigen and Antibody, and Antigen Subtypes in Atomic Bomb Survivors. Radiat Res. 1995;
[13] L. Basler, N. Andratschke, S. Ehrbar MG and ST-L. Modelling the immunosuppressive effect of liver SBRT by simulating the dose to circulating lymphocytes: an in-silico planning study. Radiat Oncol. 2018;1–8.
[14] W. L. Ford, “Distribution of lymphocytes in health”, J Clin Pathol Suppl (R Coll Pathol). 1979; 13: 63–69.
[15] Ganusov V.V., Auerbach J. Mathematical modeling reveals kinetics of lymphocyte recirculation in the whole organism. PLoS computational biology 2014; 10 (5).
[16] Blum K.S., Reinhartd P. Lymphocyte numbers and subsets in the human blood: do they mirror the situation in all organs?. Immunology letters 2007; 108 (1): 45-51.
[17] Ganusov V. V., De Boer R.J. Do most lymphocytes in humans really reside in the gut?." Trends in immunology 2007; 28 (12): 514-518.
[18] McDaniel M.M., Ganusov V. V., Estimating residence times of lymphocytes in ovine lymph nodes." Frontiers in immunology 2019; 10; 1492.
[19] Ganusov V.V., Tomura M., Experimental and mathematical approaches to quantify recirculation kinetics of lymphocytes. In Mathematical, Computational and Experimental T Cell Immunology Springer, Cham, 2021; 151-169.
[20] Williams LR, Leggett RW. Reference values for resting blood flow to organs of man. Clinical Physics and Physiological Measurement. 1989; 10: 187–217.
[21] Sharp PF. The measurement of blood flow in humans using radioactive tracers. Physiological Measurement. 1994; 15: 339–79.
[22] Leggett RW, Williams LR. A proposed blood circulation model for reference man. Health Phys. 1995;69(2):187–201.
[23] Nakamura N, Kusunoki Y, Akiyama M. Radiosensitivity of CD4 or CD8 positive human T-lymphocytes by an in vitro colony formation assay. Radiat Res. 1990;123(2):224–7.

Comparing two methods of using toxicity outcomes to study dose-response mapping of bladder and rectum in prostate cancer patients undergoing radiotherapy

Tanuj Puri[1], Marcel van Herk[1], Eliana Vasquez osorio[1], Tiziana Rancati[2], Petra Seibold[3], Eliana Gioscio[2], Andrew Green[4], Andrew P. Morris[5], Jane Shortall[1], Sarah Kerns[6], David Azria[7], Marie-Pierre Farcy-Jacquet[8], Jenny Chang-Claude[3], Alison Dunning[9], Maarten Lambrecht[10], Barbara Avuzzi[11], Dirk de Ruysscher[12], Elena Sperk[13], Christopher Talbot[14], Adam Webb[14], Ana Vega[15], Liv Veldeman[16], Barry Rosenstein[17], Catharine West[1], Ananya Choudhury[1], Peter Hoskin[1], and Alan McWilliam[1] Correspondence: tanuj.puri@manchester.ac.uk

[1]Division of Cancer Sciences, The Christie NHS Foundation Trust, The University of Manchester, UK. [2]Data Science Unit, Fondazione IRCCS Istituto Nazionale dei Tumori, Milan, Italy. [3]Division of Cancer Epidemiology, German Cancer Research Center (DKFZ), Heidelberg, Germany. [4]EMBL's European Bioinformatics Institute, Cambridge, UK. [5]Division of Musculoskeletal & Dermatological Sciences, The University of Manchester, UK. [6]Medical College of Wisconsin, Milwaukee, USA. [7]L'Institut du Cancer de Montpellier, Department of Radiation Oncology, Montpellier, France. [8]Centre Hospitalier Régional Universitaire de Nîmes, Nimes, France. [9]University of Cambridge, UK. [10]Department of Oncology, KU Leuven, Belgium. [11]Unit of radiotherapy 1, Fondazione IRCCS Istituto Nazionale dei Tumori, Milan, Italy. [12]Maastricht University Medical Center, School for Oncology and Developmental Biology Grow, Maastricht, Netherlands. [13]Department of Radiation Oncology, Universitätsmedizin Mannheim, Medical Faculty Mannheim, Germany. [14]University of Leicester, UK. [15]Universidade de Santiago de Compostela, Santiago de Compostela, Spain. [16]Ghent University Hospital, Ghent, Belgium. [17]Mount Sinai Hospital, New York, USA.

Abstract: Voxel-based analysis (VBA), such as radiotherapy dose-response maps (DRM) derived by combining dose-surface maps (DSM) and toxicity outcomes, aims to identify high-risk sub-regions in organs at risk (OARs). The study aims to assess the differences between two different methods of using late toxicities, namely, with and without baseline normalization, for identifying high-risk sub-regions on DRM of the bladder and rectum in prostate cancer (PCa) patients undergoing radiotherapy. The selection of the toxicity analysis method, whether with or without baseline normalization, influences the identification of high-risk sub-regions in DRMs of OARs. Subtracting baseline from post-treatment toxicity grades reduces the effectiveness of the dose-response relationship of bladder and rectum in prostate cancer patients undergoing radiotherapy. This influence is primarily attributed to variations in the number of events available for the analysis of each toxicity as other variables were fixed within the DRM methodology. Further efforts are required to enhance the robustness of the methodology, ensuring the generation of more reliable and reproducible results. These results may have implications for other OARs in other cancer types.

1. Introduction

Different spatial dose patterns are associated with different toxicity outcomes for organs at risk (OARs) post-radiotherapy. Various methods exist to study the dose-toxicity relationship in patients undergoing radiotherapy. Dose constraints based on dose-volume histograms (DVH) **[1]** are commonly employed to strike a balance between effectively treating the tumor while preserving surrounding healthy organs. The DVH analysis is constrained by several assumptions, including but not limited to, the following: consistent dose delivery (implying that the dose is uniformly distributed throughout the target volume), homogeneity of tissue composition (implying uniform radiation absorption within the specified volume), uniform tissue response (implying consistent sensitivity and uniform reaction across all regions within the treated volume to the delivered dose), dose linearity (implying larger doses result in greater tissue response, and vice versa) and volume-effect independence (implying the effects of radiation on distinct tissue volumes are independent).

Voxel-based techniques, such as dose-response maps (DRM) derived from associating dose-surface maps (DSM) and toxicity outcomes **[2]**, aim to identify high-risk sub-regions in OARs, with the future possibility of minimizing dose to such sub-regions during treatment plan optimization **[3-4]**. However, we know that DRM results are affected by the choice of parameter value within the DRM pipeline and the number of events included within the analysis for each toxicity **[5-6]**. However, it is unclear if using different methods of toxicity analysis will change the number of events included in the analysis, affecting the identification of high-risk sub-regions in OARs using DRMs. Therefore, this study aimed to assess the differences between two different methods of analyzing late toxicities, namely, with and without baseline normalization, for identifying high-risk sub-regions on the DRM of the bladder and rectum in prostate cancer (PCa) patients undergoing radiotherapy.

2. Materials and Methods

We analyzed data from 1808 PCa patients enrolled in the REQUITE trial **[7]** who underwent radical radiotherapy and salvage/adjuvant radiotherapy post-prostatectomy. REQUITE stands for validating pREdictive models and biomarkers of radiotherapy toxicity to reduce side effects and improve QUalITy of lifE in cancer survivors. REQUITE was a prospective multicenter cohort study conducted across 26 hospitals in 8 countries between April 2014 and March 2017, with the primary objective of validating predictive models or biomarkers for assessing the likelihood of post-radiotherapy toxicities in cancer patients.

The rectum and bladder were segmented on planning computed tomography (CT) scans using ADMIRE software (ELEKTA, Sweden). Bladder and rectum DSMs were unwrapped using spherical and cylindrical coordinates, respectively **[6]**, and spatially normalized with a fixed-sized voxel grid of 91 rows and 90 columns. Doses were converted to equivalent doses in 2 Gy fractions (EQD2) with α/β ratios of 1 Gy **[6, 8]**.

We assessed 17 Clinician-rated Common Terminology Criteria for Adverse Events (CTCAE) late toxicities (including proctitis, perforation, bowel obstruction, fistula, bowel stenosis, bowel ulceration, diarrhoea, flatus, rectal bleeding, management of sphincter, haematuria, urinary

tract obstruction, urinary incontinence, urinary frequency, urinary urgency, urinary retention, and bowel fistula) and 21 patient-reported scores through prostate-specific pelvic symptoms (including pain lower tummy, pain open bowels, pain urine, urgency bowels, urgency bowels unable, bowel control, bladder control, urgency urine, urinary frequency, urinary flow, diarrhoea, constipation, urine night (i.e., Nocturea - frequently having to get up in the night to pass urine), blood urine, blood bowels, sticky slimy motions, medication for blood urine, medication for blood bowel, medication for diarrhoea, medication for constipation and medication for bladder (i.e., to control overactive bladder)) obtained at 12 or 24 months.

Two methods were used for the toxicity analysis. In the first method, the maximum of the 12- and 24-month toxicities was normalized by subtracting the baseline toxicity value and then dichotomized at grade 1 to obtain binary toxicity outcomes (0's & 1's). In the second method, the maximum of the 12- and 24-month toxicities was used (without normalization) and dichotomized at grade 1 to obtain binary toxicity outcomes (0's & 1's).

We categorized the DSMs into two groups based on dichotomized toxicity values. Welch's T-test was conducted voxel-wise across the 8190-voxels locations over two stacks of images to obtain a 2D-image of t-values (i.e., a T-map), assuming a normal distribution and based on the null hypothesis that the two populations have equal means, even when the two samples exhibit unequal variances and sample sizes. To address multiple comparisons, permutation testing was performed with 1000 iterations. The maximum T-value over the entire image (called Tmax) was obtained for each of the 1000 iterations. These Tmax values were then thresholded at the 95th percentile (corresponding to an adjusted *P*-value=0.05). This threshold was utilized to threshold T-map for true labels to identify statistically significant voxels to obtain a DRM. Figure 1 illustrates the DRM generation pipeline.

Datasets were excluded if the baseline or 12- and 24-month toxicities were missing, the dose or CT files were absent, the ADMIRE failed to segment rectum or bladder regions, the DSM visually appeared incorrect, the patient received brachytherapy, the difference between the prescribed dose and image-based mean prostate dose was more than ±10%, or the fractionation data was missing.

3. Results

3.1. The results presented here are only for CTCAE-based late toxicities due to space limitations. Bladder and rectum DRMs showed associations with CTCAE-based bladder-related toxicities but not with rectum-related toxicities.

For the bladder, toxicities such as urinary incontinence (Figure 2), urinary tract obstruction (Figure 3), urinary retention (Figure 4) and urinary urgency (Figure 5) showed significant regions highlighted on the DRMs (as shown by the blue color contour in these images). All four DRMs showed significant regions without baseline toxicity normalization, but only the former two showed significant regions with normalization. For the rectum, toxicities such as urinary incontinence (Figure 6), urinary tract obstruction (Figure 7) showed significant regions highlighted on the DRMs. The former only showed significant regions only without baseline toxicity normalization, but the latter showed significant regions for both types of toxicity analyses, with and without normalization.

3.2. Urinary incontinence (Figure 2) and urinary tract obstruction (Figure 3) toxicities for bladder DRMs and urinary tract obstruction (Figure 7) for rectum DRMs, highlighted statistically significant regions that were approximately concentric (but of different sizes) for the two different methods of toxicity analyses, suggesting that the difference in the highlighted regions may be due to the number of events included in the analyses, as other parameters were kept fixed. The toxicity analysis methods without baseline normalization led to an increase in the number of events compared to the other methods with baseline normalization, and therefore larger significant regions were identified without baseline toxicity normalization compared to the other method, as previously reported **[5-6]**.

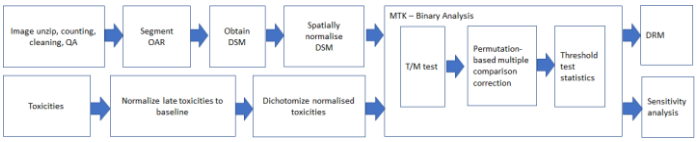


Figure 1. Dose-response maps (DRMs) methodology. Arrows describe the direction of operation. QA=quality assurance; OAR=organs-at-risk; DSM=dose-surface maps; MTK=Manchester tool kit; T/M=Welch t-test.

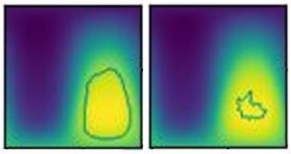


Figure 2. Urinary incontinence: **spherical** unwrapping for **bladder** dose-surface maps, α/β=1, 91x90 voxel grid, t-test, 1000 iterations, and 95th percentile. Without (left: **361** events) and with (right: **226** events) baseline normalization. Colors represent mean radiation dose over all patients included in the analysis for this toxicity type, with yellow indicating the highest dose, followed by green, blue, and black for decreasing dose magnitudes. Significant values are within the drawn contour on these images.

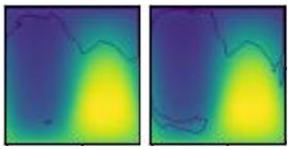


Figure 3. Urinary tract obstruction: **spherical** unwrapping for **bladder** dose-surface maps, α/β=1, 91x90 voxel grid, t-test, 1000 iterations, and 95th percentile. Without (left: **93** events) and with (right: **82** events) baseline normalization.

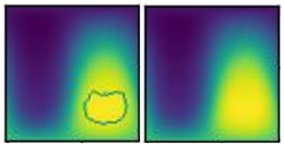


Figure 4. Urinary retention: **spherical** unwrapping for **bladder** dose-surface maps, α/β=1, 91x90 voxel grid, t-test, 1000 iterations, and 95th percentile. Without (left: **104** events) and with (right: **82** events) baseline normalization.

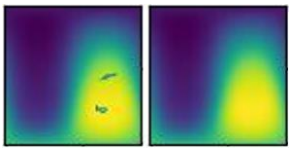


Figure 5. Urinary urgency: **spherical** unwrapping for **bladder** dose-surface maps, α/β=1, 91x90 voxel grid, t-test, 1000 iterations, and 95th percentile. Without (left: **465** events) and with (right: **286** events) baseline normalization.

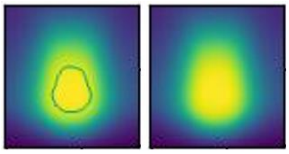


Figure 6. Urinary incontinence: **cylindrical** unwrapping for **rectum** dose-surface maps, α/β=1, 91x90 voxel grid, t-test, 1000 iterations, and 95th percentile. Without (left: **361** events) and with (right: **226** events) baseline normalization.

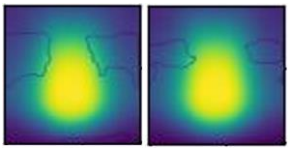


Figure 7. Urinary tract obstruction: **cylindrical** unwrapping for **rectum** dose-surface maps, α/β=1, 91x90 voxel grid, t-test, 1000 iterations, and 95th percentile. Without (left: **93** events) and with (right: **82** events) baseline normalization.

4. Discussion

4.1. Urinary incontinence and urinary tract obstruction toxicities highlighted significant regions on the DRMs of both the bladder and rectum. Urinary retention and urinary urgency toxicities only highlighted significant regions on the bladder's DSM. Therefore, it currently remains unclear if toxicities such as urinary incontinence and urinary tract obstruction occur as a result of the combined effect of radiation damage to both the rectum and bladder, or whether the regions identified on the rectum for urinary incontinence (which only showed significant regions without normalization) and for urinary tract obstruction (which showed significant regions for without and with normalization) are false positives, as the bladder-related toxicities are identified on rectum DRMs.

Test statistics, Tmax, is just a single extreme value property obtained from 8190 T-values from the image that we subject to further testing by applying 1000 permutations. Only the 1000 permutations count as multiple comparisons. Therefore, it was considered appropriate to call the 95th percentile as *P*-value=0.05 in the sense that 5% of the highest value permutations in the histogram of all 1000 permutations. The motivation for performing permutation testing on Tmax was to lower the *P*-value threshold (to minimise false positives) below the commonly chosen 0.05 level, aiming to account for multiple comparisons by considering the proportion of permutation samples producing a higher Tmax value than the one observed with true labels. However, this adjustment results in a stringent threshold, rendering it inadequate for identifying any significant regions on DRM. For instance, the bladder DSM with normalized urinary tract obstruction toxicity (Figure 3) yielded an adjusted *P*-value << 0.05 (i.e., close to zero) after correcting for 1000 multiple comparisons. This led to the failure to identify any significant region on the DRMs. Therefore, the threshold was adjusted to the 95th percentile to identify the significant voxels on DRMs, raising questions about the need for implementing and using permutation testing in the first place. Moreover, this choice of using a *P*-value threshold of 0.05 may have contributed to the identification of regions on the rectum's DRM for bladder-related toxicities (in figures 6-7), potentially resulting in false positives. Consequently, future studies employing VBA should, with caution **[9]**, consider adopting *P*-value thresholds < 0.05 (i.e., higher than 95th percentile) to minimize the occurrence of false positives when identifying significant regions on DRMs.

These results raise a further question about the validity of VBA applied to the combined datasets from radical radiotherapy and salvage/adjuvant radiotherapy post-prostatectomy and provide a compelling case for re-analysing the data in these two cohorts separately. It is because the two treatment approaches are given with different intent. Radical radiotherapy is given with curative intent and involves delivering targeted doses to the prostate gland to eliminate or shrink cancerous cells and, in some cases, the surrounding tissues. Salvage/adjuvant radiotherapy is given after a prostatectomy, with/without signs of recurrence or persistent cancer cell, intended to treat the remaining cancer cells in the prostate bed to prevent the cancer from spreading/returning. Plausibly, prostatectomy may have had a cascading impact on neural bundles around the prostate, potentially resulting in sexual health-related toxicities, which were excluded from the analysis.

4.2. Bladder DRM for urinary retention (104 events) and urinary urgency (465 events) showed statistically significant regions only for the toxicity analysis without baseline normalization. However, these cases may not be considered false positives as these toxicities are bladder-related toxicities and include a larger number of events. The lack of significant regions in the other analysis with baseline normalization could be due to the lack of events for these toxicities (82 and 286 events for urinary retention and urinary urgency, respectively). Possibly, some toxicity types may need more events to show significant regions on DRMs than other toxicity types, based on the strength of relationship between DSM and the toxicity type under investigation.

Moreover, the fact that there is such a good match among Figures 2, 4 and 5 in the significant regions highlighted near the high dose regions provides independent support that the observed statistically significant relationship between the dose and toxicity is not unrealistic.

Even though the reason for the differences in significant regions identified on OARs between the two methods, without and with normalization, appeared to be due to the number of events involved in the analysis, the minimum number of events required for toxicity to start showing significant regions on the DRMs and well as the maximum number of events required beyond which the identified significant regions on DRM show insignificant differences with the increasing number of events is unknown. Answering such questions requires further research.

4.3. The assumptions behind toxicity normalization were (i) that the DSM will only be associated with the post-radiotherapy toxicities due to radiation damage and that the toxicity at baseline (i.e., pre-radiotherapy) will be excluded from the analysis in identifying the high-risk sub-regions within OARs, and (ii) that the toxicity severity grades are linear in nature, such that the difference between two toxicity grades will produce the same effect; for example, grade 3 minus grade 2 is considered to have the same clinical effect as grade 1 minus grade 0, which may not be true. On the other hand, the toxicity analysis without baseline normalization assumes an exponential characteristic of toxicity grades, such that grade 3 minus grade 2 is considered to have a greater clinical effect than grade 2 minus grade 1, which his considered to have greater clinical effect than grade 1 minus grade 0, representing a more realistic clinical scenario. For further studies, since baseline toxicities are expected to be an important factor for toxicity outcomes post-radiotherapy, for the toxicity analysis without normalization (a more realistic clinical scenario), it may be worth using a toxicity prediction model that includes the baseline toxicities *a priori* within the model (e.g., using multivariable VBA).

4.4. Even though parameter values (such as α/β=1, voxel grid size=91x90, statistical test=Welch T-test, test statistics=Tmax, number of iterations=1000 and threshold=95th percentile used to obtain the DRMs were fixed in the present study, whether the choice of these parameter values is valid for each individual toxicity is a separate question that requires further investigation.

5. Conclusion

The choice of toxicity method (without or with normalization) impacts the identification of high-risk sub-regions on DRMs of OARs. Subtracting baseline from post-treatment toxicity grades reduces the effectiveness of dose-response relationship of bladder and rectum in prostate cancer patients undergoing radiotherapy. Therefore, the details of toxicity utilization as well as the number of events included in VBA for each toxicity type should be reported in the manuscripts for external validation. These results may have implications for other OARs in other cancer types (e.g., heart OAR in lung cancer). More work is needed to achieve a more robust methodology to produce reliable and reproducible results. Future work could combine genomics and dose data for toxicity prediction modelling.

6. Acknowledgements

We thank several individuals and entities, including patients, staff across 26 hospitals spanning 8 countries, Corinne Johnson-Hart (for crafting and submitting Concept Form Proposal 42 to secure project approval from REQUITE Consortium) and Peter Remeijer (for the WorldMatch function to perform spherical unwrapping) and others whom we may have inadvertently omitted from this acknowledgment. This research was made possible through funding from the European Union - Seventh Framework Programme for Research, Technological Development, and Demonstration, under Grant ID: 601826. Tanuj Puri's salary is supported by The Christie Hospital NHS Trust, Movember, Prostate Cancer UK.

7. References

[1] G.T.Y. Chen, M. Austin-Seymour, J.R. Castro, et al. "Dose volume histograms in treatment planning evaluation of carcinoma of the pancreas". Proceedings of the Eighth International Conference on the Use of Computers in Radiation Therapy, IEEE Computer Society, Silver Springs (1984), pp. 264-268.

[2] Acosta, O., Drean, G., Ospina, J. D., et al. "Voxel-based population analysis for correlating local dose and rectal toxicity in prostate cancer radiotherapy." Physics in medicine and biology, 58(8), (2013) 2581–2595. https://doi.org/10.1088/0031-9155/58/8/2581

[3] Palma, G., Monti, S., & Cella, L. "Voxel-based analysis in radiation oncology: A methodological cookbook." Physica medica: PM: an international journal devoted to the applications of physics to medicine and biology: official journal of the Italian Association of Biomedical Physics (AIFB), 69, (2020). 192–204. https://doi.org/10.1016/j.ejmp.2019.12.013

[4] McWilliam, A., Palma, G., Abravan, A., et al. "Voxel-based analysis: Roadmap for clinical translation." Radiotherapy and oncology: journal of the European Society for Therapeutic Radiology and Oncology, 188, 109868. (2023). https://doi.org/10.1016/j.radonc.2023.109868

[5] Puri, T., van Herk, M., Vasquez Osorio, E., et al. (2023). "Assessing Interaction between Toxicity Outcomes and Dose Surface Maps at the Bladder and Rectum in Prostate Cancer Patients." Poster presented at the 65th AAPM 2023 Conference, 23 July 2023 – 27 July, Houston, USA. (2023).

[6] Puri, T., van Herk, M., Vasquez Osorio, E., et al. (2024). "Sensitivity analysis of dose-response mapping model parameters at the bladder and rectum in prostate cancer patients." Proffered paper accepted at the ESRO 2024 Conference, 03 May 2024 - 07 May, Glasgow, UK. (2024).

[7] Seibold, P., Webb, A., Aguado-Barrera, et al. "REQUITE: A prospective multicentre cohort study of patients undergoing radiotherapy for breast, lung or prostate cancer." Radiotherapy and oncology: journal of the European Society for Therapeutic Radiology and Oncology, 138, 59–67. (2019). https://doi.org/10.1016/j.radonc.2019.04.034

[8] Brand, D. H., Brüningk, S. C., Wilkins, A., Fernandez, K., Naismith, O., Gao, A., Syndikus, I., Dearnaley, D. P., Tree, A. C., van As, N., Hall, E., Gulliford, S., & CHHiP Trial Management Group (2021). "Estimates of Alpha/Beta (α/β) Ratios for Individual Late Rectal Toxicity Endpoints: An Analysis of the CHHiP Trial." International journal of radiation oncology, biology, physics, 110(2), 596–608. https://doi.org/10.1016/j.ijrobp.2020.12.041

[9] Noguchi, K., Konietschke, F., Marmolejo-Ramos, F. et al. Permutation tests are robust and powerful at 0.5% and 5% significance levels. Behav Res 53, 2712–2724 (2021). https://doi.org/10.3758/s13428-021-01595-5

A technological pipeline for detecting radiation-induced heart damage mechanisms and prevent major cardiac adverse events

Alessandro Cicchetti[1], Alessandra Catalano[1], Silvia Meroni[2], Anna Cavallo[2], Albina Allajbej[3], Chiara Veronese[4], Paola Vallerio[4], and Claudia Sangalli[3]

[1]Data Science Unit, Fondazione IRCCS Istituto Nazionale dei Tumori, Milan, Italy
[2]Division of Medical Physics, Fondazione IRCCS Istituto Nazionale dei Tumori, Milan, Italy
[3]Department of Radiation Oncology, Fondazione IRCCS Istituto Nazionale dei Tumori, Milan, Italy
[4]Cardiopulmonary Unit, Fondazione IRCCS Istituto Nazionale dei Tumori, Milan, Italy

Abstract The heart is believed to be the most critical normal structure responsible for competing events after RT in locally advanced lung cancer patients. A better knowledge of heart substructures' tolerance to RT is essential to define a safe dose escalation protocol and improve patients' overall survival.
Here, we combine a retrospective study to determine dose tolerances for the heart substructures through a prediction model of survival limiting toxicity and a prospective analysis to investigate the pathophysiological process leading to chronic symptoms and major adverse cardiac events.
In silico applications will be adopted to (i) design a decision-making tool for advanced RT patient selection and (ii) test the mechanisms of cardiac dysfunctions.

1 Introduction

Radiation therapy (RT) represents an option for treating lung, lymphoma, oesophagus and breast cancer. For these districts, the partial inclusion of the heart into the medium-high dose region limits the level of curative RT doses and, thus, the treatment success rate [1, 2]. The most robust evidence was reported by the phase 3 trial of RT dose-escalation RTOG 0617 [3], where the median survival for the enrolled lung patients was significantly worse in the higher dose arm. One contributing factor was the dose delivered to the heart.
Along the same line, a mathematical model on 7000 lung patients by Nix [4] showed that up to 20% in 2-year Overall Survival (OS2) gain could be obtained "if normal critical structures in which survival limiting toxicities arise can be identified and selectively spared without compromising the tumour coverage".
The heart is believed to be the most critical normal structure responsible for these competing events, and a better knowledge of its substructures' tolerance to RT is essential. The pathophysiology of radiation-induced heart disease has become more evident in the last decade. However, studies defining the biology that drives the dose-response relationships have been limited, and the results are controversial [5], even when considering the mean dose to the heart (MHD).
A reasonable option is a needful shift to the heart substructures (HSs) to understand better the causality of major adverse cardiac events (MACEs) after RT.
Moreover, robust subclinical warnings linked to the substructure dosimetry must be identified to comprehend cardiac disease pathophysiology better. Indeed, acute radiation-induced damage is usually considered an exclusive inflammatory process. For this organ, such a view is reductive, with chronic consequences of inflammation that can substantially alter the equilibrium in the electrophysiology and the fluid dynamics of the heart. An improvement in describing the detrimental mechanisms of the cardiac tissues can open specific interventions and prevent patient symptoms independently from the model predictions.
In this manuscript, we explained the methodological workflow and the role of technological innovations within the study "Location Matters".

2 Materials and Methods

The study is based on a well-structured research path that begins with a retrospective analysis aimed at identifying statistical associations in a large patient population to establish dose limits for a new set of at-risk structures, namely cardiac substructures. The prospective population will allow for a detailed estimation of the acute impact of radiotherapy on the functional and morphological parameters of the heart and lungs, thus identifying radiation-induced damages. A complex 3D model capable of mimicking cardiac function will enable the investigation of stress areas caused by parameters damaged by radiation therapy to determine potential mechanisms leading to the development of Major Adverse Cardiovascular Events (MACE).
In this process, innovative AI-based software, research machinery capable of determining parameters not yet in use in clinical practice, and everyday devices enabling longitudinal tracking of simple and precise functionalities play a crucial role.

2.1 The retrospective study cohort

A dataset of 1300 European LA-NSCLS patients treated between 2013 and 2020 will be shared and centralized at Fondazione IRCCS Istituto Nazionale dei Tumori. The clinical outcome, the Dicom RT, patient factors, tumour and treatment characteristics will be collected.

Using MVISION, an AI automatic contour software for CT scans, the delineation of the HSs will be obtained for all the patients. The central core of structures is characterized by the aorta, left atrium, right atrium, left ventricular, right ventricular, inferior ventricular cava, superior ventricular cava, and pulmonary artery. Dosimetric variables will be computed for the new structures, overcoming the whole heart dosimetry limitation.
In addition, patients' coronary artery calcifications (CAC) will be automatically contoured on the same imaging modality, and the Agatston calcium score will be computed through AVIEW (Coreline Soft, Seoul, Korea). This baseline parameter has been massively associated with a higher risk of developing heart dysfunctions [6] and will be included in the model with dose descriptors for HSs. Also, CAC location will be registered to understand if calcium deposits are more likely to induce cardiac diseases when localized at a specific point.

The study endpoint will be the OS2 for LA-NSCLC patients. A novel combined approach will be performed to analyse this endpoint. We will start with the mathematical model proposed by Nix et al. to describe positive and negative contributions to the OS2 in LA-NSCLC patients. It considered three components for describing the OS2: the equivalent dose (EQD2) to the tumour to take into account the loco-regional response, the chemotherapy's impact on distant relapse and a last parameter defining the survival limiting toxicity (SLT), i.e. the reduced survival when higher doses to the tumour were prescribed. Unlike Nix et al., where this part is only considered as a mathematical function of the prescribed dose for the tumour, we propose a description of causality for the SLT, as depicted in Fig. 1.
The description will be performed in three progressive methods: (i) substituting the dose to the tumour with the dose to the HSs, (ii) as in (i) + the Agatston calcium score, (iii) as in (ii) + the dose to the normal lung tissue (i.e., lungs – PTV). The reason for choices (ii) and (iii) comes from the previously introduced impact of the CAC on heart disease and the well-described cooperation between the heart and lung, which shows how the loss of cardiopulmonary performance is intensified when the heart and lung are co-irradiated. Through this process, we will complete the findings by Sheperd et al. [7] on the reduction of survival with the increase in the MHD modelling this competing part through the dose to the HSs. Reasonably, we could move from one dosimetric constraint for the whole heart to a set of rules for more substructures involved singularly or in a couple according to the anatomical location of the tumour or to the dose gradient for that treatment technique.
Moreover, the model component on the distant relapse will be enriched, including the impact of immunotherapy on the OS2 for patients with PDL-1 expression.

2.2 A decision-making tool for advance RT patient selections

The derived set of dosimetric constraints for safely sparing the heart substructures and their functionality will guide an in-silico comparison of different treatment modalities. Indeed, when a treatment plan for a patient does not satisfy these limits, the clinician could address the patient toward RT techniques that are more capable of satisfying the dose distribution requirements. Notably, in this study, we could test the clinical gain derived from the use of Tomotherapy (San Raffaele Hospital), hadron-therapy (CNAO Foundation) and MRI-Linac (Policlinico Gemelli of Rome). The differences in the dose distribution will be converted through the model (specifically to the part for SLT) to estimate the risk reduction of radiation-induced heart disease caused by such more conforming treatment modalities. The model developed could allow us to determine if such dose differences are clinically relevant more than statistically significant.

2.3 The prospective study cohort

A cohort of LA-NSCLC patients treated with RT will be enrolled at Fondazione IRCCS Istituto dei Tumori. Comorbidities and cardiovascular risk factors will be recorded for all the patients. We will plan an extensive set of cardiac tests before, at the end of RT, and 8 months after treatment (to assess cardiac remodelling and patient recovery). It will include electrocardiogram, echo cardio, spirometry, blood circulating biomarkers (Troponin and Natriuretic peptide tests) and cardiac calcium score from CT scan (including Follow-up acquisitions). In addition, we will include a set of measures using the cardiovascular ultrasound machine MyLab X8 (Esaote). This machine will allow us to investigate subclinical vascular organ damage indicators, including carotid distention and intima-media thickness (descriptor of the carotid inflammatory state), through a simultaneous and automatic procedure from raw imaging data. This guarantees the extraction of highly reproducible and operator-independent features.
Finally, patients will be equipped with a wearable device (Garmin Venu 3) to collect longitudinal information on heart rate, oxygen saturation, respiration rate, aerobic power and maximal oxygen uptake, and training status, including the recovery time after customised exercise based on the patient status.
The prospective analysis aims to associate the most "radiosensitive" functional parameters or biomarkers (i.e. the more prone to variate after RT) with the dosimetric indications for HSs from the retrospective study. First, we will analyze the variations of the parameters in patients with doses above and below the identified constraints for HSs. Secondly, we will test whether a dose relationship model can continuously describe the changes for the parameters with statistically significant differences. The primary focus will be the comprehension of the dose response for morphologi-

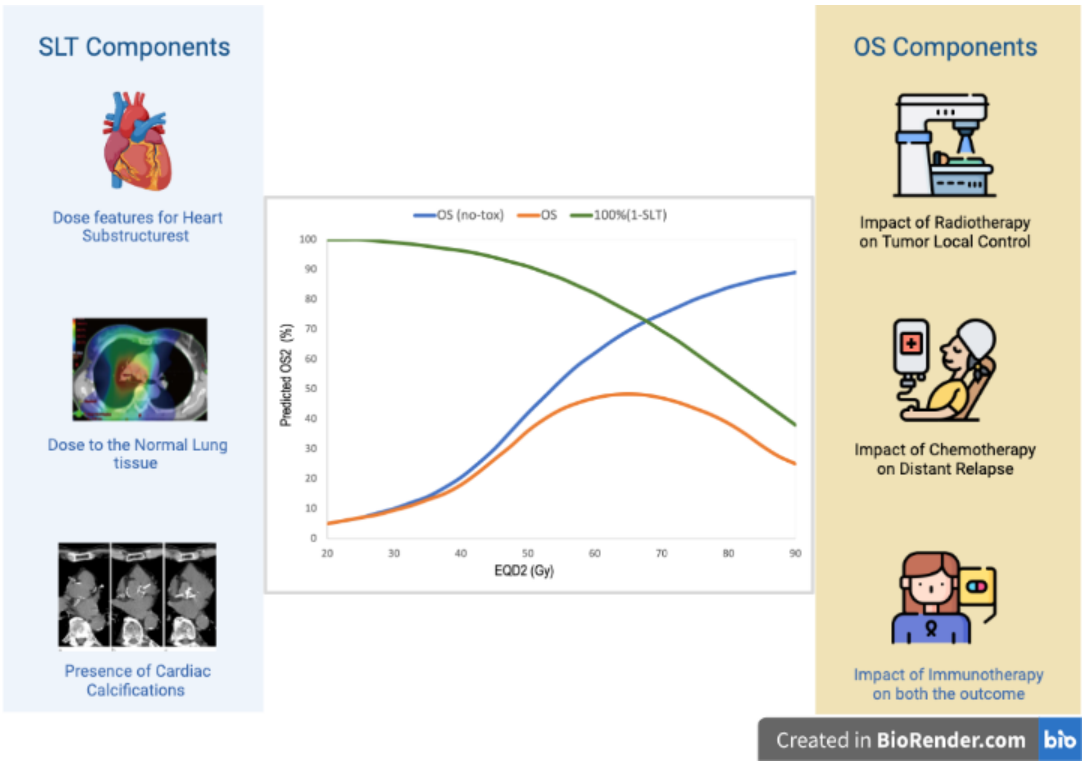


Figure 1: Graphical representation of the two components contributing to the Overall Survival at 2-year post-RT curve (orange line). Blue is the theoretical curve of overall survival caused by the benefits of RT on tumour volume, chemotherapy, and immunotherapy. In green is the complementary of the surviving limiting toxicity components, which includes the dose to the heart substructures, the dose to the normal lung and the presence of cardiac calcification before the start of the RT (impacting on the green line

cal and functional parameters that are altered chronically (8 months after RT). For this study, we estimate to enrol 150 LA-NSCLS patients in 3 years.

2.4 The 3D computational model

In collaboration with Politecnico of Milan, a mathematical approach will use the available patient data of the study to formulate an in-silico mechanistic model for blood flow in large vessels and the heart. The simulation will consider risk factors due to fluid mechanics and the mechanical alteration caused by the radiation on the endothelium and the surrounding tissue. The model could quantitatively describe the differences in the output parameters due to the inclusion of input values from patients below and above the subclinical warnings identified in the prospective study. To summarise, the real-world data coming from the clinical setting will inform an already available computational model for numerical simulations of the cardiac functions to describe in a mechanistic way the impact of such altered parameters on the natural physiology of the heart.

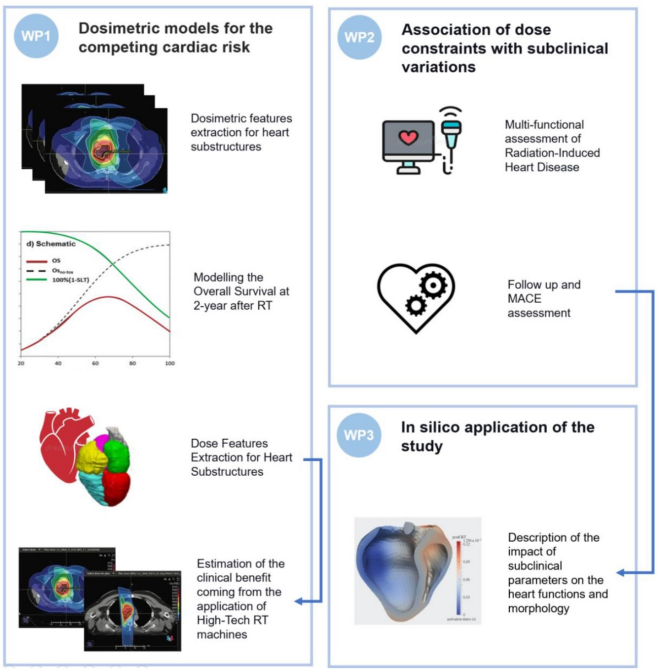


Figure 2: The study workflow: 1) the analysis of heart substructures dose tolerances, 2) the association of dose variables and alteration of functional parameters, 3) the in-silico mechanistic description through a 3D heart model. Arrows indicated the origin of the input parameters for the decision-making tool and the in silico application.

3 Discussion

The identification of the dose tolerance for the most radiosensitive HS, i.e., the one more prone to produce a critical dysfunction in a short period after cancer treatment, could impact the average overall survival time for LA-NSCLC patients,

improving the success rate of CH-RT, which is currently limited by the adverse events. The development of a survival model, including the dosimetric information from HSs, can drive the clinical practice toward a safe dose escalation protocol, exploiting the current technological level of the RT at best. For example, it could allow the performance of a safer and more effective stereotactic body radiation therapy (SBRT). There is considerable interest in combining SBRT with Immunotherapy without any concern about the possible detrimental effects of high doses on small heart substructures. This could lead to an apparent low efficacy of this treatment combination due to the neglect of survival-limiting toxicity. The model could also provide quantitative information for selecting patients who could benefit from using high-tech machines such as the hadron therapy. Indeed, a practical estimate for the risk of cardiac events should be included in the eligibility criteria for the mentioned machines and should not only be based on patient status (presence of cardiovascular comorbidities), taking into account even healthy cardiac patients subjected to doses that the HSs could not tolerate. The definition of subclinical warnings coming from the functional and morphological in-vivo test on patients during the treatment and the follow-up could improve the knowledge of Radiation-Induced Heart Disease (RIHD) and the tissue alteration regarding parenchymal cell composition, structure functionality and electrophysiology. Early identification of deficit values for some parameters that do not manifest into symptoms before inducing MACEs could allow a preventive intervention by cardiologists. The comprehension of such critical parameters is even of help for all the patients who received a safe dose of radiation to the heart but even started to degenerate. Patients with unfavourable genetic characteristics represent some examples of tolerating the RT damage or with already impaired tissue and comorbidities. Finally, the answer to all of these questions is not limited to lung cancer but could give advantages to other thoracic cancer treatments.

4 Conclusion

Establishing dosimetric constraints for heart substructures will provide an instrument for safe dose escalation in thoracic cancer patients, improving the success rate of radiotherapy alone or in combination with other cancer treatments. The estimated gain in terms of Overall Survival for lung cancer patients is 15-20%. A longer life expectancy for these patients could show the real benefit of the combination of RT, Chemotherapy and Immunotherapy. Indeed, Artificial Intelligence (AI) -based automatic contouring and segmentation of the HS can lead to more reliable dosimetric models for RIHD for cancer patients treated with RT in the thoracic region. Moreover, clinicians can exploit these dosimetric models to address patients at risk, i.e., patients with critical tumour location or needing cancer retreatment toward a better-sparing treatment technique or radiation source. The project results can be exploited for other cancer districts where RT is part of the oncological path, and the heart could be included in the medium-high dose region.

5 Acknowledgments

The study was supported by AIRC MFAG, no. 27480.

References

[1] C. Bergom, J. A. Bradley, A. K. Ng, et al. “Past, Present, and Future of Radiation-Induced Cardiotoxicity: Refinements in Targeting, Surveillance, and Risk Stratification”. *JACC: CardioOncology* 3.3 (2021). Radiation and Cardiovascular Disease: Mini-Focus Issue, pp. 343–359. DOI: https://doi.org/10.1016/j.jaccao.2021.06.007.

[2] C. Brink, U. Bernchou, A. Bertelsen, et al. “Causal relation between heart irradiation and survival of lung cancer patients after radiotherapy”. *Radiotherapy and Oncology* 172 (2022), pp. 126–133. DOI: https://doi.org/10.1016/j.radonc.2022.05.002.

[3] J. D. Bradley, R. Paulus, R. Komaki, et al. “Standard-dose versus high-dose conformal radiotherapy with concurrent and consolidation carboplatin plus paclitaxel with or without cetuximab for patients with stage IIIA or IIIB non-small-cell lung cancer (RTOG 0617): a randomised, two-by-two factorial phase 3 study”. *The Lancet Oncology* 16.2 (2015), pp. 187–199. DOI: https://doi.org/10.1016/S1470-2045(14)71207-0.

[4] M. G. Nix, C. G. Rowbottom, S. Vivekanandan, et al. “Chemoradiotherapy of locally-advanced non-small cell lung cancer: Analysis of radiation dose–response, chemotherapy and survival-limiting toxicity effects indicates a low / ratio”. *Radiotherapy and Oncology* 143 (2020), pp. 58–65. DOI: https://doi.org/10.1016/j.radonc.2019.07.026.

[5] “Cardiac Toxicity of Thoracic Radiotherapy: Existing Evidence and Future Directions”. *Journal of Thoracic Oncology* 16.2 (2021), pp. 216–227. DOI: https://doi.org/10.1016/j.jtho.2020.11.002.

[6] P. Greenland, M. J. Blaha, M. J. Budoff, et al. “Coronary Calcium Score and Cardiovascular Risk”. *Journal of the American College of Cardiology* 72.4 (2018), pp. 434–447. DOI: https://doi.org/10.1016/j.jacc.2018.05.027.

[7] A. F. Shepherd, A. F. Yu, M. Iocolano, et al. “Increasing Heart Dose Reduces Overall Survival in Patients Undergoing Postoperative Radiation Therapy for NSCLC”. *JTO Clinical and Research Reports* 2.8 (2021), p. 100209. DOI: https://doi.org/10.1016/j.jtocrr.2021.100209.

Whole-body Organ Dose Assessment of Cancer Patients Undergoing External Beam Radiotherapy Using Body-size-dependent Phantom Library for Secondary Cancer Risk Prediction

Kyeong Hyeon Sung[1], Bo-Wi Cheon[1], Hyung-Joo Choi[1], Soo Min Lee[1], Saerom Seong[1], Yoon Seo[1], Donghyouk Kim[1], Hyun Joon Choi[2], and Chul Hee Min[1,*]

[1]Department of Radiation Convergence Engineering, Yonsei University, Wonju, Korea
[2] Department of Radiation Oncology, Wonju Severance Christian Hospital, Yonsei University Wonju College of Medicine

Abstract: Intensity-modulated radiotherapy (IMRT), while effective for conformal dose irradiation, may inadvertently expose out-of-field areas to radiation due to scattering and leakage doses from the treatment head, potentially leading to secondary cancer. In this study, we evaluated whole-body organ doses to predict the risk of secondary cancer using body-size-dependent tetrahedral mesh-based computational phantoms (TM phantoms) with our developed TET2DICOM software. TET2DICOM software converts TM phantoms into DICOM format, enabling their integration into commercial treatment planning systems (TPS). We utilized MONACO TPS for whole-body organ dose assessment. All selected TM phantoms, with similar BMI to the patients, were successfully transformed into DICOM format. A comparison of average doses at the target volume between patient CT images and TM phantom CT images revealed an error rate of 1.2% to 2.6%. After verifying this method through this comparative analysis, it was finally possible to quantitatively calculate the whole-body organ dose.

1 Introduction

Modern radiation therapy (RT) technology has progressed toward the aim of administering the most optimal dose to the tumor while simultaneously reducing the dose to the organ-at-risk. Intensity-modulated radiotherapy (IMRT) is one of the widely used advanced RT techniques in oncology, it allows the conformal dose delivery to the tumor by dividing a number of beam fractions.

Nevertheless, such hyper-fraction RT technique is reported in an increase of about 3-5 times in the out-of-field dose, attributed to either leakage dose or the scatter radiation than conventional RT technique [1-2]. The TG-158 report from the American Association of Physicists in Medicine (AAPM) suggests that a low dose of out-of-field can give rise to adverse consequences, including skin-related side effects, cardiac toxicity, or secondary cancers in normal tissues [3].

Treatment planning systems (TPS), presently utilized in clinical practice, may exhibit diminished accuracy in dose assessment as the distance from the region of interest expands. Besides, due to the restricted nature of patient Computed tomography (CT) data to the local treatment area, accurate evaluation of the dose to organs outside of the field can be difficult [4].

Meanwhile, the International Commission on Radiological Protection (ICRP) has recently put forth the tetrahedral mesh-based computational human phantoms (TM phantoms) [5] that faithfully replicate the anatomical structures of the human body. Through this, it suggests the possibility of accurately evaluating the dose of whole organs exposed to radiation.

Integrating the TM phantoms into the TPS may be feasible to achieve a comprehensive dose assessment of the whole body. In addition, as it is feasible to predict the dose administered to major organs of individuals receiving therapeutic radiation, patient-specific metadata can be obtained to be employed in subsequent investigations, such as the forecasting study of secondary cancers.

In this study, our research team has developed a conversion software application, called TET2DICOM [6], from TM phantoms into a Digital Imaging and Communications in Medicine (DICOM) format for employment in TPS. Furthermore, the out-of-field dose delivered to major organs according to the treatment target organs was evaluated.

2 Materials and Methods

2.1 Conversion body-size-dependent TM phantoms to DICOM format of the whole-body CT images

In order to convert the TM phantoms into CT images, a total of 6 instances involving cancer patients undergoing radiotherapy were taken into account. As shown in Table 1, taking into consideration both gender and Body Mass Index (BMI), the TM phantom that bore the closest resemblance to the patient in terms of shape was utilized for the conversion process.

Table 1. BMI of candidate patients and TM phantoms.

			Patient		TM Phantom	
Patient case	Gender	Treatment area	Height (cm)	Weight (kg)	Height (cm)	Weight (kg)
PC 1	Male	Brain	163	69	165	70
PC 2	Male	Esophagus	173	67.5	175	70
PC 3	Male	Rectum	172.7	89.3	175	90
PC 4	Female	Cervix	162.8	46.5	165	45
PC 5	Female	Liver	145	42.7	150	40
PC 6	Female	Breast	148	43.7	150	45

This study employed the TET2DICOM software to transform TM phantoms into CT images, which were subsequently utilized in commercial TPS applications. It encompassed two distinct procedures: (1) the conversion of a mesh-type structure into a voxel-type structure, which represents the CT image; and (2) contouring the data about the location, shape, and volume of the organs.

The CT-to-Density curve was measured by acquiring CT images using a cheese phantom containing 12 non-homoge-

neous material rods. The average CT value within the circular region of 8 mm diameter in each distinct material area was taken into account, as shown in Figure 1. It was employed to establish the HU value within the voxel based on the material information derived from the TM phantom.

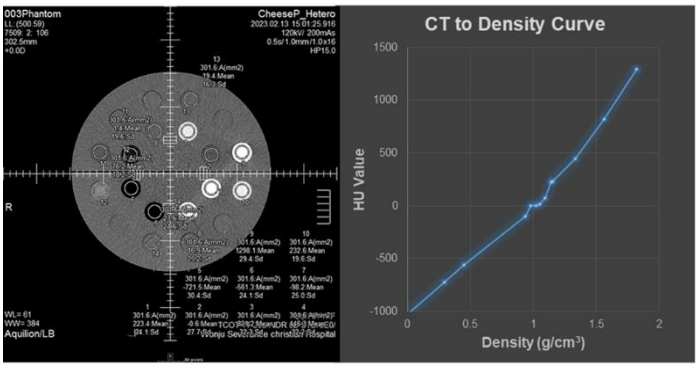


Figure 1. CT-to-Density curve data: the correlation between CT image intensity and material-specific density.

Contour data was defined for major organs in the body with relatively different probabilistic risks during radiation exposure specified in the ICRP 103 report to calculate the organ dose in commercial TPS.

Finally, in order to apply the converted CT image to a commercial TPS, the CT image and contour data were converted to Digital Imaging and Communication in Medicine (DICOM) format. The Grassroots DICOM Library (GDCM), which is a cross-platform library for DICOM medical files was used, and CT image data was converted to binary format according to specific tag rules.

2.2 Whole-body organ dose assessment by commercial TPS

Before performing the calculation of the dose for the entire body, an image registration process was carried out to align the converted DICOM data with the positions of the patient's CT image treatment site and the organs of the entire body, as illustrated in Figure 2. For the purpose of image registration, a commercially available RT plan check system (MIM, MIM Software Inc., Cleveland, USA, version 6.5.6) was utilized. This rigid registration method only includes the rotation and movement of the entire image, without form of distortion such as size or shape transformation of the CT images.

After that, the MONACO 6.1 version TPS software was utilized to calculate the whole-body organ dose using a CT image of the TM phantom. The dose assessment algorithm employed in this software is a simplified Monte Carlo algorithm, specifically designed to evaluate the whole-body organ dose through scattered radiation.

The contour data for each organ was generated as an RT-structure file so that the dose distribution was quantitatively evaluated using a dose volume histogram (DVH). The dose calculated in the Clinical Target Volume (CTV) or Planning Target Volume (PTV) of the radiotherapy plan established

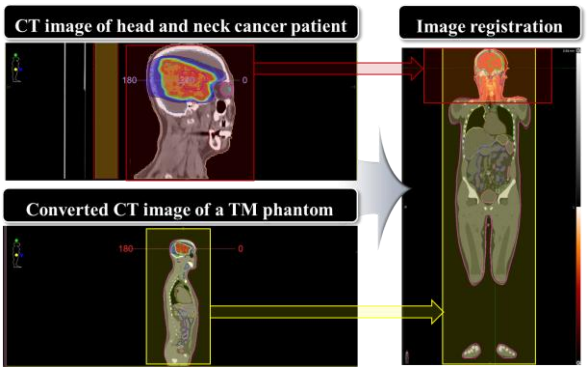


Figure 2. Rigid registration of the CT image of a cancer patient and the converted CT image from the TM phantom.

from each patient's CT image was compared and analyzed with the dose calculated in the CTV or PTV formed in the TM phantom.

3 Results

3.1 Conversion of TM phantoms with similar BMI to 6 patients

All 6 TM phantoms, which had similar BMI to the patients, were successfully transformed into DICOM CT images, as illustrated in Figure 3. For the conversion parameter, as outlined in Table 2, the field of view (FOV) was stated to be 500 mm, which was automatically adjusted to accommodate the size of the phantom. The resolution of the images was set at 512 × 512, and the voxel size was determined to be 1.281 × 1.281 × 3 mm^3, depending on the FOV. The number of slices produced for the phantoms varied from 532 to 590, depending on the height of each phantom.

Table 2. Parameters for conversion of CT image with TM phantoms.

	CT image parameters			
	Field of view (mm)	Pixel size (mm^2)	Pixel resolution	Number of slices
PC 1	500	1.281 ×1.281	512 × 512	564
PC 2	500	1.281 ×1.281	512 × 512	590
PC 3	500	1.281 ×1.281	512 × 512	590
PC 4	500	1.281 ×1.281	512 × 512	572
PC 5	500	1.281 ×1.281	512 × 512	532
PC 6	500	1.281 ×1.281	512 × 512	532

The generation of a DICOM dataset for 6 TM phantoms using TET2DICOM required a time range of 80 to 100 minutes, respectively, with a specific duration dependent on the size of the phantom. Moreover, it is worth noting that all converted CT images were found to be compatible with commercial TPS.

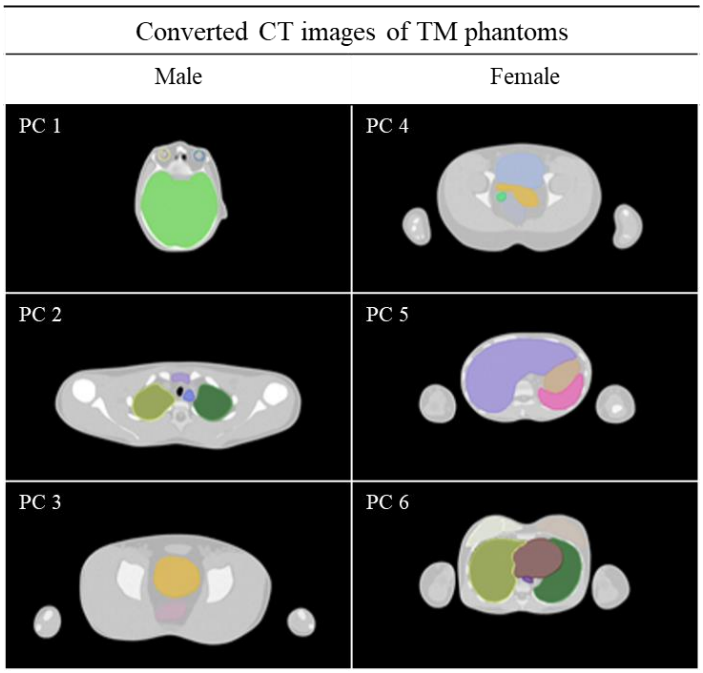


Figure 3. Converted CT images of TM phantoms with similar BMI for each patient case.

3.2 Whole-body organ dose assessment with the converted DICOM CT data

Figure 4 shows the rigid registration results between the CT images of the 6 patients and the TM phantoms which are similar in gender and BMI with 6 patients. Although the specific target volume is not explicitly defined in the converted CT image derived from the TM phantom, and there exists a slight disparity in the dimensions and placement of organs, the distribution of dose was calculated by aligning the patient-specific parameters of the skeletal attributes.

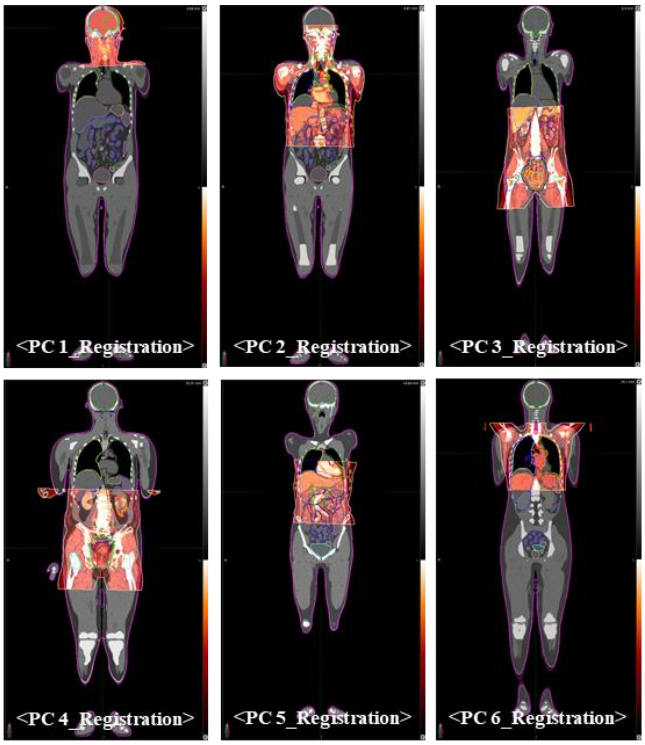


Figure 4. Registration of the patient's CT image and the TM phantom's CT image

By comparing the dose calculated on the CTV or PTV of the existing radiation treatment plan established from the patient's CT image with the dose calculated on the CTV or PTV calculated by converting and matching the TM phantom, the dose difference was confirmed as shown in Table 3. The maximum dose within the target volume in all patient cases except for the breast cancer patient case showed an error rate of 0.6% to 5.1%, and the average dose showed a difference of 1.2% to 2.6%.

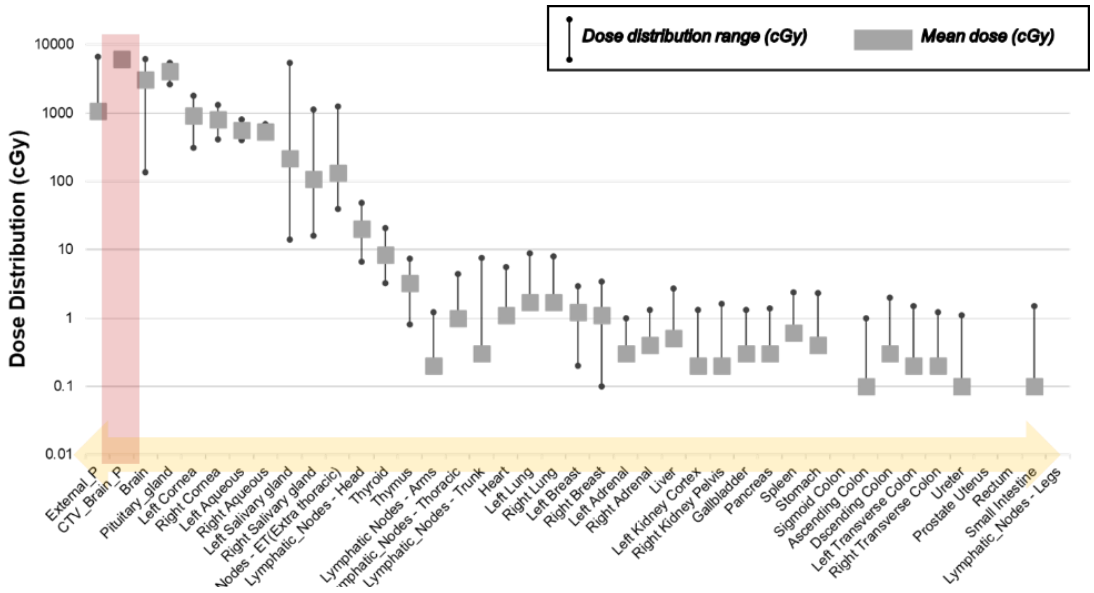


Figure 5. Dose distribution for each organ of the whole body for brain tumor patients.

Table 3. Comparison of calculated dose between the CTV or PTV of each patient's original CT image and the converted CT image in the TPS.

	Maximum dose within target volume (cGy)			Average dose within target volume (cGy)		
	CT image	TM Phantom	Error (%)	CT image	TM Phantom	Error (%)
PC 1	6381.4	6709.1	5.1	5973.7	6129.3	2.6
PC 2	4455.6	4673.6	4.9	4177.1	4273.4	2.3
PC 3	4838.9	4866.6	0.6	4524.8	4466.1	1.3
PC 4	4333.7	4396.3	1.4	3177.4	3217.0	1.2
PC 5	4837.8	5010.4	3.6	4550.6	4663.6	2.5
PC 6	4590.7	4689.6	2.2	4311.1	3728.5	13.5

After verifying the evaluation method by calculating the dose in the target volume, the delivered dose for organs of the whole-body was calculated and quantitatively evaluated based on the contour data, and the results as shown in Figure 5.

4 Discussion

In this study, we aimed to calculate the dose of whole-body organs by converting the body-size-dependent phantom library into DICOM format and applying it to the commercial treatment planning system. TET2DICOM software was developed and used to convert the tetrahedral-mesh computational phantom to DICOM format, and for calculating whole-body organ doses, MONACO treatment planning system was used.

TM phantoms that had a similar BMI to the chosen 6 patients were converted into DICOM format, as depicted in Figure 3. Subsequently, a dose calculation and comparison were conducted to validate the method used to assess the doses on whole-body organs. This involved comparing and analyzing the average dose and maximum dose on the clinical target volume (CTV) or planning target volume (PTV) within the converted DICOM CT data and the actual patient's CT image. In all cases, except for patient 6 whose treatment area was breast cancer, the average dose in the CTV or PTV for each patient demonstrated a tendency to be consistent within a range of 2.6%. However, for patient 6, the average dose error was evaluated to be 13.5%. This discrepancy was deemed relatively substantial due to the concentration of radiation irradiation on the body surface and the notable differences in breast volume and shape.

The graph displayed in Figure 5 exhibits the quantitatively calculated dose to the entirety of the patient's organs, resulting from radiotherapy administered for the treatment of a brain tumor. This graph effectively categorizes the organs of the entire body, including the target volume, in a hierarchical manner from the top to the bottom of the body. It has been verified that the dose in the target volume amounts to 6129.3cGy, while lower doses have been observed to be distributed below 100mGy in organs situated further away from this specific area. It has been ascertained that the quantification of the whole-body organ dose is not only applicable to the treatment of brain tumors but also encompasses all five patients involved in this study. Furthermore, it has been determined that the evaluation of the whole-body organ dose, which is the aim of this study, can indeed be accomplished.

5 Conclusion

The confirmation of the possibility to quantitatively assess the dose received by the patient's organs has been achieved through this study. Currently, there are limitations in calculating accurate organ dose due to distortion of the volume and organ shape between actual patient CT and converted CT data. However, in upcoming years, it is envisaged that the construction of an extensive database containing a multitude of patient cases, utilizing the approach employed in this study, will allow for the provision of factors for referencing secondary carcinogenic risk. Furthermore, the establishment of a framework for the presentation of globally standardized guidelines aimed at optimizing novel radiation treatment plans can be expected.

References

[1] Ruben, Jeremy D., et al. "A comparison of out-of-field dose and its constituent components for intensity-modulated radiation therapy versus conformal radiation therapy: implications for carcinogenesis." *International Journal of Radiation Oncology* Biology* Physics* 81.5 (2011): 1458-1464. DOI: 10.1016/j.ijrobp.2010.08.008

[2] Taylor, Michael L., and Tomas Kron. "Consideration of the radiation dose delivered away from the treatment field to patients in radiotherapy." *Journal of medical physics/Association of Medical Physicists of India* 36.2 (2011): 59. DOI: 10.4103/0971-6203.79686

[3]Kry, Stephen F., et al. "AAPM TG 158: measurement and calculation of doses outside the treated volume from external-beam radiation therapy." Medical physics 44.10 (2017): e391-e429. DOI : 10.1002/mp.12462

[4] Howell, Rebecca M., et al. "Accuracy of out-of-field dose calculations by a commercial treatment planning system." *Physics in Medicine & Biology* 55.23 (2010): 6999. DOI: 10.1088/0031-9155/55/23/S03

[5] Kim, Chan Hyeong, et al. "ICRP publication 145: adult mesh-type reference computational phantoms." *Annals of the ICRP* 49.3 (2020): 13-201. DOI: 10.1177/0146645319893605

[6] Cheon, Bo-Wi, et al. "Development of a novel program for conversion from tetrahedral-mesh-based phantoms to DICOM dataset for radiation treatment planning: TET2DICOM." *Journal of Applied Clinical Medical Physics* 23.1 (2022): e13448. DOI: 10.1002/acm2.13448

Towards Gaussian processes modelling to study the late effects of radiotherapy in children and young adults with brain tumours.

Angela Davey[1], Arthur Leroy[2], Eliana Vasquez Osorio[1], Kate Vaughan[1], Peter Clayton[3,4], Marcel van Herk[1], Mauricio A Alvarez[2], Martin McCabe[1,4] and Marianne Aznar[1]

[1]Division of Cancer Sciences, Faculty of Biology, Medicine and Health, The University of Manchester, Manchester, United Kingdom.
[2]Department of Computer Sciences, Faculty of Science and Engineering, The University of Manchester, Manchester, United Kingdom.
[3]Division of Developmental Biology and Medicine, The University of Manchester, Manchester, United Kingdom.
[4]The Christie NHS Foundation Trust, Manchester, United Kingdom.

Abstract: Survivors of childhood cancer need life-long monitoring for side-effects from radiotherapy. However, longitudinal data from routine monitoring is often infrequently and irregularly sampled, and subject to inaccuracies. Due to this, measurements are often studied in isolation, or simple relationships (e.g., linear) are used to impute missing timepoints. In this study, we investigated the potential role of Gaussian Processes (GP) modelling to make population-based and individual predictions, using insulin-like growth factor 1 (IGF-1) measurements as a test case. With training data of 23 patients with a median (range) of 4 (1-16) timepoints we identified a trend within the range of literature reported values. In addition, with 8 test cases, individual predictions were made with an average root mean squared error of 31.9 (10.1 – 62.3) ng/ml and 27.4 (0.02 – 66.1) ng/ml for two approaches. GP modelling may overcome limitations of routine longitudinal data and facilitate analysis of late effects of radiotherapy.

1 Introduction

Each year an estimated 400,000 children are diagnosed with cancer worldwide [1]. Radiotherapy plays a main role in curative-intent treatment with more than 75% of children now surviving more than 5 years [2]. Despite modern radiotherapy techniques, brain tumour survivors have the highest incidence of severe chronic health conditions compared to other childhood cancers, in part due to the impact of radiation on the developing brain [3]. There is a need to study the relationship between radiation to brain substructures and late outcomes so we can plan 'gentler' radiation treatments and ultimately improve patient outcomes.

A common outcome post-radiotherapy in survivors of childhood brain tumours is growth hormone deficiency (GHD), which, if it goes untreated, can lead to short-stature, fatigue and increased risk of diabetes and heart disease among others [4,5]. Routine blood tests are used as follow-up for brain tumour survivors, which include a specific test for Insulin-like growth factor 1 (IGF-1), a primary index for diagnosis of GHD.

To enable the study of the association between dose to brain substructures and GHD, we have so far dichotomised the outcome as the likelihood a child is placed on treatment for GHD (yes vs no). A dichotomous outcome could lead to loss of valuable information on patterns of change for those within each group. In future, we plan to make use of longitudinal measurements of IGF-1 for a more granular study of outcome. Follow-up routine blood rests are typically recorded for children many years post-treatment. However, these tests are infrequent, sampled at irregular timepoints, and subject to inaccuracies. To move towards dose-response modelling with longitudinal data, we need to understand population trends with age and make predictions of an individual's trajectory to provide outcomes at interpolated standardised time-points such that patients can be compared. Gaussian processes (GP) modelling has the potential to enable this.

GP modelling is a machine learning approach that can be used to capture the function that maps any input variable (i.e., time) to any output variable (i.e., IGF-1) [6]. Fundamentally, it is non-parametric regression and more flexible in comparison with typical parametric approaches (e.g., linear and polynomial regression) that are somewhat limited in the spectrum of trends they are able to capture. Due to this flexibility, GPs are particularly powerful in situations where there is limited data and/or relationships are complex. GPs are probabilistic in nature and hence leverage prior information during training to provide uncertainty quantification for the output estimates of the model. Multi-task GPs encompass the situation where a GP may be trained on data from multiple individuals. In this case, one can take advantage of a latent mean process, common to all individuals, to share information across them and enhance modelling and prediction accuracy. This is useful for the situation where some individuals do not have many timepoints sampled or have large gaps in recording, as they can 'borrow' information from other individuals. This would allow us to impute data based on population trends. One package for performing this type of prediction is MAGMA (Multi tAsk GPs with common MeAn) developed by Leroy et al. [7]. MAGMA has proved advantageous in multiple simulated and realistic scenarios (e.g., predicting athletes swimming performance and modelling body mass index during childhood). However, the ability to use such techniques in routine follow-up data for childhood cancer survivors has not yet been explored.

Here, we train multi-task GPs with a common mean to measure how IGF-1 changes with age from routine measurements of children and young adults who received

radiotherapy for a brain tumour. We will test **1)** the ability to identify underlying population trends, and **2)** the ability to interpolate individual timepoint predictions in unseen data. By performing these tasks, we highlight the potential of using GPs to impute data at a given time, thereby enabling study of the late effects of radiotherapy.

2 Materials and Methods

Data

We collected data from a single institutional archive on blood tests for IGF-1 recorded after radiotherapy for 31 children and young adults with brain tumours. The dataset was divided quasi-randomly to ensure a 75% / 25% split for training and testing (23 training, 8 test). The quasi-random component was to make sure patients with one timepoint were only included in the training set (as there would not be enough points for testing).

Training

We trained a MAGMA model, which is defined as the sum of a shared mean GP and an individual specific GP. We test two variations model hyperparameters (HPs) in training: **1)** consider *common* mean and covariance HPs to be equal for all individuals, assuming individuals correspond to different trajectories of the same process, and **2)** individual-specific HPs to generate a more flexible model. Model training is performed via likelihood maximisation within an expectation-maximisation algorithm. To account for the influence of the HPs random initialisation, we repeat the training process 25 times per variation and select the GP models with the maximum log-likelihood. We compare the trained population GP to literature reported values.

Testing

We ensured that each individual the in the test set had at least two timepoints. Per test patient we randomly sample half of the timepoints for prediction, and half for evaluation. Using the prediction points for each patient, we can predict the entire curve of IGF-1 values with associated uncertainty, in particular at the location of the testing points for evaluation purposes. To evaluate the individual GP model, we calculate the root mean squared error (RMSE) and coverage of the 95% credible interval (CIC-95) for each of the evaluation timepoints. We report the average statistics across the evaluation points for each case. The evaluation is repeated for the common and individual-specific HPs models.

3 Results

Data

Longitudinal blood test data was available for 149 data points (114 for training, 35 for testing) across the 31 individuals. All patients had received photon radiotherapy for a brain tumour of mixed diagnosis. Patients were of median age 21.0 (3.79 – 25.5) years at radiotherapy. IGF-1 measurements were recorded at a median (range) of 4.82 (0 - 14.4) years after radiotherapy. All patients were on growth hormone replacement therapy and had begun this intervention treatment 7.19 (0.49 – 14.6) years after radiotherapy which should return IGF-1 to normative ranges. In terms of data availability, the median (range) number of recordings of IGF-1 for each patient was 4 (1-16). Six patients had only one recording and hence were included in the training dataset, and the rest of the data was randomly selected for training and testing.

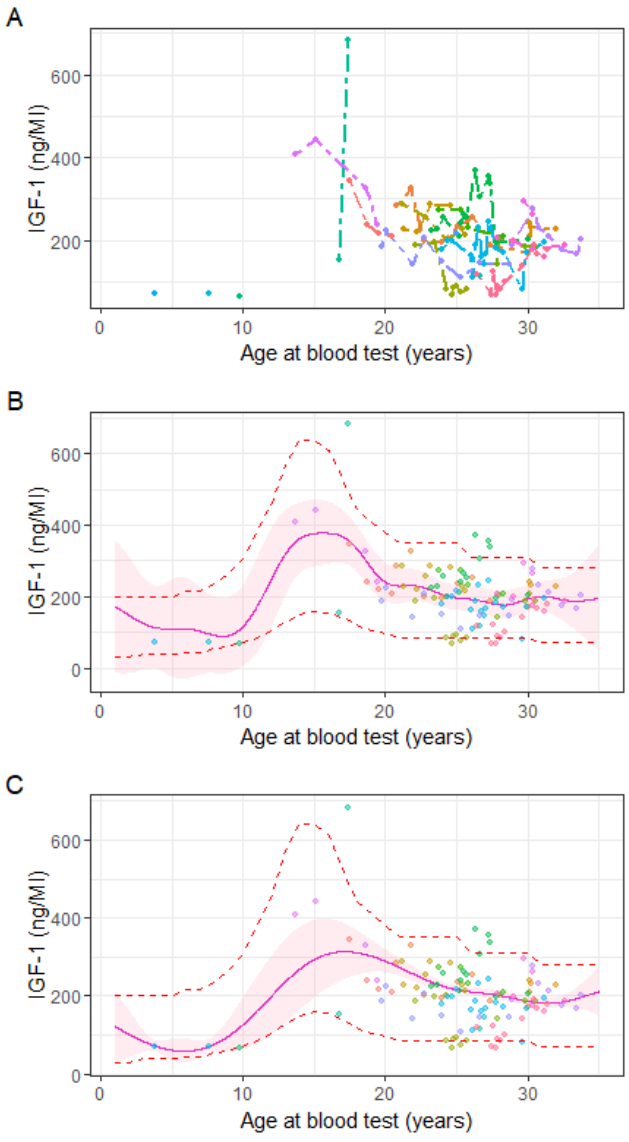


Figure 1. A) Individual trajectories for patients within the training dataset. Each colour represents a patient, with lines connecting time-points from the same individual. B) Fitting of a GP model with common hyperparameters, the light pink represents the 95% credible interval. C) Fitting GP model without individual-specific hyperparameters. The red dashed lines display normative trends for IGF-1.

Training

In Figure 1A an individual's 1GF-1 trajectory is mapped out over the age at which the test was taken. Figure 1B shows the GP model assuming common mean

hyperparameters. The maximum log-likelihood for this model fit was -804.56 with a minimum of -821.74 across the 25 training runs. In the GP model with individual-specific HPs (Figure 1C) the maximum log-likelihood was -744.88 within a minimum of -779.31. In comparison to normative reference values (red dashed lines on Figure 1 [8]), when there is a common HP most timepoints are within the ranges until very few data-points for training are available (i.e., <12 years old) and the GP follows the shape of the normative trajectory. For individual-specific HP, the credible interval is within the range for all timepoints, but the shape is smoother with a smaller peak. The models trained in 64 and 76 seconds respectively.

Testing

The results of the 8 test cases for individual-specific GPs are reported in Table 1.

Case	Prediction	Evaluation	RMSE (common HP)	RMSE (individual-specific HP)
1	2	3	24.96	31.94
2	2	2	38.24	92.55
3	2	3	25.56	41.24
4	1	1	10.12	18.74
5	4	5	62.29	66.06
6	1	1	46.07	0.02
7	3	3	10.89	22.78
8	1	1	56.65	5.53
			31.90	***27.36***

Table 1. A table of the test results for the GP models with common and individual-specific HPs. The case is an individual patient, prediction gives the number of timepoints included for prediction, and evaluation is the number of timepoints included for evaluation. The RMSE represents the average across the evaluation timepoints (ideal value is 0), and overall means are included in the bottom row.

On average, the lowest RMSE is observed in the individual-specific HP model, however, not all evaluation points were covered by the 95% credible interval suggesting a slight underestimation of the uncertainty. There was 100% CIC-95 in all cases apart from case 5 who had 50% coverage. For the common HP setting all evaluation points had 100% CIC-95. Considering individuals, in 6 out of 8 cases, the common HP approach has an improved model fit for prediction. But the overall average RMSE is skewed by the predictions where only one timepoint is available. For cases with one timepoint it appears there is an advantage to the more flexible model.

Model fit examples are displayed in Figure 2. In case one (A and B) the datapoints show fluctuations around the common mean, so, in both cases the error is small. In case 5 (C and D) an uncharacteristic peak is observed in the years post-treatment, and here the population mean process does not fit well in either case. In case 6 and 8 only one timepoint is available for fitting the individual prediction (and one for testing), so the individual-specific HP performs better (E-F and G-H respectively).

4 Discussion

We have studied the use of MAGMA to perform Gaussian Processes modelling on the dependency between IGF-1 and age in a dataset of children treated with radiotherapy for brain tumours and under growth hormone treatment. We have shown the GPs can correctly model the underlying expected trends from normative samples.

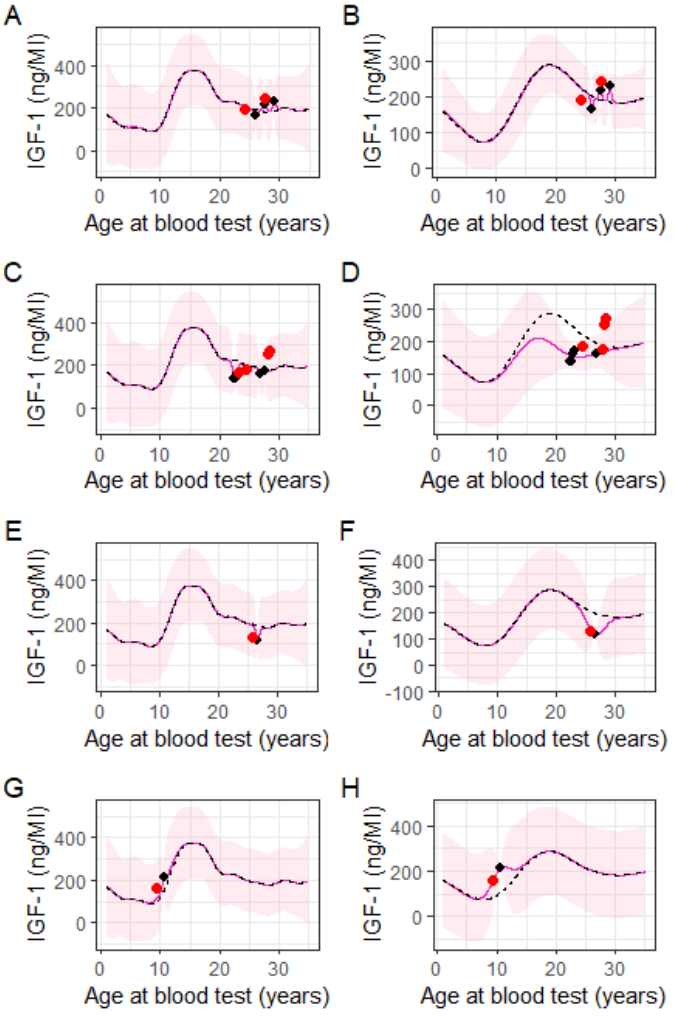


Figure 2. Examples of individual GP models (pink line) for 4 out of 8 example cases using the common mean process (dotted line) trained using common HPs (left) and individual-specific HPs (right). The prediction points for that case are in black, and evaluation points (unseen in training and prediction) are in red.

Individual predictions were performed with an ~30ng/ml root mean squared error, and in all but one case (94%) evaluation points were covered by the 95% credible

interval. In future work, we aim to incorporate the longitudinal predictions modelled by GPs into our existing image-based data-mining pipeline [9].

Gaussian processes have previously been explored in radiotherapy for motion-modelling for image-guided treatment [11] and motion compensation in robotics treatment [12]. However, as far as the authors are aware they have not yet been utilized for dose-response outcome modelling.

The large credible intervals displayed by the model are unsurprising given the noise in the dataset. It is known that IGF-1 scores are also influenced by the time after radiotherapy, the age at radiotherapy, puberty status, weight, diet, and day-to-day variations in recording [13,14]. In addition, we are typically interested in measuring what is 'abnormal' so the ability to model normative ranges may appear a redundant test. However, all these patients were administered growth hormone therapy at some point post-radiotherapy. The potential confounding variables and external factors influencing this data raises concern for its use for showcasing the advantages of GPs in dose-response modelling. Incorporating extra covariates including time-varying ones, using MAGMA will be explored to overcome some of the issues presented here for future work with this cohort.

5 Conclusion

We demonstrated how multi-task GPs with a common mean process (MAGMA) can be used in routine and 'inadequately sampled' outcome data. The model was able to identify population-level trends that match known values and make individual predictions on 'missing' data-points where the true values were within the 95% credible interval for 94% of cases. This tool will aid the analysis of late effects of radiotherapy in childhood cancer as we could interpolate missing timepoints in routine follow-up data, and eventually study patterns of development across different treatment groups or dose levels. This will also complement our ongoing image-based data mining work for endocrine outcomes.

References

[1] Steliarova-Foucher et al. *International incidence of childhood cancer*, 2001-10: a population-based registry study. Lancet Oncol. 2017;18(6):719-731. 10.1016/S1470-2045(17)30186-9

[2] Cancer Research UK, *Young people's cancer statistics*, 2021. Available at: https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/young-peoples-cancers

[3] Bhakta et al. *The cumulative burden of surviving childhood cancer: an initial report from the St Jude Lifetime Cohort Study (SJLIFE).* The Lancet 390.10112 (2017): 2569-2582. https://doi.org/10.1016/S0140-6736(17)31610-0

[4] Baunsgaard et al. *Growth hormone deficiency in adult survivors of childhood brain tumors treated with radiation.* Endocr Connect. 2023 Jan 19;12(2):e220365. doi: 10.1530/EC-22-0365.

[5] Ranke et al. *Short and Long-Term Effects of Growth Hormone in Children and Adolescents With GH Deficiency.* Front Endocrinol (Lausanne). 2021 Sep 1;12:720419. doi: 10.3389/fendo.2021.720419.

[6] Deisenroth et al. *A practical guide to Gaussian Processes*. 2020. Available at: https://infallible-thompson-49de36.netlify.app/

[7] Leroy, A. MAGMA: inference and prediction using multi-task Gaussian processes with common mean. *Mach Learn* **111**, 1821–1849 (2022). https://doi.org/10.1007/s10994-022-06172-1

[8] University of Michigan Pathology Handbook, 2017. Available at: https://www.pathology.med.umich.edu/handbook/#/details/855

[9] Wilson et al. *Image-based data mining applies to data collected from children.* Phys Med. 2022 Jul;99:31-43. doi: 10.1016/j.ejmp.2022.05.003.

[10] Huttinger et al. *Gaussian Processes for real-time 3D motion and uncertainty estimation during MR-guided radiotherapy.* Medical Image Analysis, 2023;88:102843. doi: 10.1016/j.media.2023.102843

[11] Dürichen et al. *Multitask Gaussian processes for multivariate physiological time-series analysis*. IEEE Transactions on Biomedical Engineering 62.1 (2014): 314-322. 10.1109/TBME.2014.2351376

[12] Lee et al. *Discordance of insulin-like growth factor-1 results and interpretation on four different platforms.* Clin Chim Acta. 2023 Jan 15;539:130-133. doi: 10.1016/j.cca.2022.11.034.

[13] Clemmons. *Clinical utility of measurements of insulin-like growth factor 1.* Nat Clin Pract Endocrinol Metab. 2006 Aug;2(8):436-46. doi: 10.1038/ncpendmet0244.

Explainable Artificial Intelligence for the Prediction of Non-small Cell Lung Carcinoma Recurrence

Lorna Tu[1,2], Hervé H.F. Choi[2,1], Haley Clark[3,1] and Samantha A.M. Lloyd[2,1]

[1]Department of Physics and Astronomy, University of British Columbia, Vancouver, Canada
[2]Department of Medical Physics, BC Cancer – Vancouver, Vancouver, Canada
[3]Department of Medical Physics, BC Cancer – Surrey, Surrey, Canada

Abstract Despite careful radiotherapy treatment planning, recurrence may occur in patients with non-small cell lung carcinoma (NSCLC). If clinicians can identify patients at high risk of recurrence, treatment plans could be tailored with risk-adapted approaches. Although artificial intelligence (AI) can predict patient outcomes, it often does not explain its underlying decision-making process and therefore is not widely adopted in clinical practice. In this retrospective study, we use explainable AI methods to predict post-stereotactic ablative radiotherapy NSCLC recurrence in 74 patients treated across a Canadian province and interpret the results. The eXtreme Gradient Boosting model outperformed the random forest and Explainable Boosting Machine models, achieving an accuracy of 0.73, F1-score of 0.5, and area under the curve of 0.65. Radiomic features were considered to be the most important features, whereas clinical features were not as important overall. This work shows the utility of explainable AI in the context of recurrence prediction.

1 Introduction

Stereotactic ablative radiotherapy (SABR) is recommended for patients with stage I and II non-small cell lung cancer (NSCLC) who are medically inoperable, or who are potentially operable but considered high-risk, elderly, or prefer not to undergo surgery [1]. While SABR for early-stage NSCLC reports over 90% local control, there remains a substantial risk of distant metastasis, with rates ranging between 10% and 20% [2]. A retrospective study reports that the median times to local and distant recurrence after SABR are 14.9 and 9.6 months respectively [3]. Currently, the integration of individualized recurrence risk estimation into the radiotherapy planning process is not a standard practice. If the recurrence risk of each patient could be precisely predicted, treatment strategies could be tailored with risk-adapted approaches, such as escalation or de-escalation, to potentially enhance treatment effectiveness and durability.

Recent strides in artificial intelligence (AI) have revolutionized many medical fields, including radiation oncology. AI models have been developed for tasks including image reconstruction, registration, segmentation, diagnosis, and outcomes prediction [4]. Despite their popularity and accuracy, their low interpretability prevents their widespread adoption in medicine. A clinician trusting a prediction without understanding its limitations or the underlying rationale may cause adverse consequences for the patient.

In response to the black-box nature of traditional AI, explainable artificial intelligence (XAI) has emerged to help the user better understand model predictions [5]. Among them, Shapley Additive exPlanations (SHAP) is a method based on game theory that determines the importance of each feature for a given prediction [6]. SHAP values can be used to identify the most important features on a global or individual basis. Another approach proposes the use of glass box models designed to be inherently more interpretable. Although XAI has generated increasingly interest in medicine, to our knowledge, its application to predicting NSCLC recurrence after SABR remains unexplored.

While traditional AI models have been developed for recurrence prediction, they have not gained enough trust of clinicians to be used broadly in practice. This work aims to explore the use of XAI to predict NSCLC recurrence in SABR patients treated across the Canadian province of British Columbia. XAI would provide clinicians with outcomes predictions that are interpretable and related to patient-based features, which would help clinicians to assess the prediction's reliability and to justify any treatment plan modifications. It has the potential to support clinical decision making and ultimately improve treatment outcomes in British Columbia.

2 Materials and Methods

This retrospective study was approved by the British Columbia Cancer Research Ethics Board (H21-02186). The patient dataset was collected retrospectively and consists of 74 NSCLC patients treated with SABR across British Columbia, Canada between 2010 and 2017. Patients receiving simultaneous treatment to multiple disjointed tumour sites, receiving atypical chemotherapy treatment, or lacking sufficient follow-up time (one year) were excluded. The dataset included medical records and radiotherapy planning computed tomography (CT) scans containing contours delineated by expert clinicians. Chart review was performed to determine the two-year recurrence status of each patient. Locoregional recurrence and distant metastasis were both considered disease recurrence in this study, and not distinguished. Patient data was anonymized.

Clinical features from medical records were one-hot encoded. Radiomic features were extracted from gross tumour volumes on CT scans using IBEX radiomic software [7]. Subsequent analysis was performed using Python 3.7. Transforms were performed on heavily skewed continuous features to achieve a more Gaussian-like distribution. Feature selection was performed using Pearson coefficient correlation with a cut-off threshold of 0.8, resulting in a total of 15 radiomic and 28 one-hot encoded clinical features.

The data set was split into training (80%) and testing (20%) sets with balanced classes. Random Forest (RF), eXtreme Gradient Boosting (XGBoost), and Explainable

Boosting Machine (EBM) classification models were trained using stratified ten-fold cross validation. RF and XGBoost are commonly used tree-based models that can be explained using SHAP values. EBM is a glass-box model designed to have comparable accuracy to RF and XGBoost, but with increased interpretability. Hyperparameters for RF and XGBoost were optimized using Bayesian optimization. Following EBM author recommendations [8], the default EBM hyperparameters were used. Model performance on the test set was evaluated using accuracy, F1-score, and area under the curve (AUC). Model explainability was investigated using SHAP values and EBM visualizations.

Study Population Statistic	Values
Number of patients	74
Mean age of patients	75.6
Number of patients by sex	
Female	39 *(53%)*
Male	35 *(47%)*
Number of patients by recurrence status	
Did recur	23 *(31%)*
Did not recur	35 *(69%)*
Number of patients by health authority	
Fraser	34 *(46%)*
Interior	7 *(9%)*
Northern	4 *(5%)*
Vancouver Coastal	20 *(27%)*
Vancouver Island	8 *(11%)*
Number of patients by AJCC stage	
IA	44 *(59%)*
IB	13 *(18%)*
IIB	5 *(7%)*
IIA	10 *(14%)*
IIIB	2 *(3%)*
Number of patients by histology	
Adenocarcinoma	35 *(47%)*
Carcinoma	9 *(12%)*
Lepidic adenocarcinoma	1 *(1%)*
Mucinous adenocarcinoma	1 *(1%)*
Neoplasm	12 *(16%)*
Non-small cell carcinoma	7 *(9%)*
Papillary adenocarcinoma	1 *(1%)*
Squamous cell carcinoma	8 *(11%)*
Number of patients by site	
Lung, lower lobe	26 *(35%)*
Lung, middle lobe	3 *(4%)*
Lung, upper lobe	43 *(58%)*
Main bronchus	1 *(1%)*
Number of patients by laterality	
Left origin of primary	30 *(41%)*
Right origin of primary	44 *(59%)*
Number of patients by diagnosis method	
Clinical diagnosis only	3 *(4%)*
Histology	50 *(68%)*
Radiology/imaging only	21 *(28%)*

Table 1: Demographics of the study population.

Table 2 (continued): Demographics of the study population.

Model Hyperparameter	Values
Random forest	
Maximum depth	8
Maximum features	log2*
Number of estimators	64
XGBoost	
Lambda	0.1
Learning rate	1
Maximum depth	4
Number of estimators	256

*Table 2: Optimum hyperparameters for the models. *Calculated as log2 of number of features*

3 Results

The dataset consists of 74 non-metastatic stage I–III NSCLC patients treated in five provincial health authorities across British Columbia, Canada. The majority of patients were in AJCC (7th edition) stage I and had adenocarcinoma. The mean age of patients was 75.6 years, with a 31% recurrence rate. Demographics of the study population are provided in Table 1.

Bayesian optimization was performed with 50 iterations. Optimized model hyperparameters are shown in Table 2.

Performance metrics of the models are summarized in Table 3. The best performing model was XGBoost with an accuracy of 0.73, F1-score of 0.5, and AUC of 0.65. Figure 1 shows AUC curves of all three models. The most important features based on XGBoost SHAP values are shown in Figure 2 and mostly consist of radiomic features.

Global feature importance of the models represented by SHAP values or EBM visualizations was compared. Radiomic features (Autocorrelation, ClusterShade, and LongRunLowGrayLevelEmpha) were identified within the top five important features of all three classifiers. Two clinical features ranked among the top ten important features overall by specific classifiers. Adenocarcinoma histology was the sixth most important feature for the XGBoost model, while age at diagnosis ranked tenth for the RF model.

Model	Accuracy	F1-score	AUC
RF	0.73	0.33	0.6
XGBoost	0.73	0.5	0.65
EBM	0.67	0.0	0.5

Table 3: Performance metrics for the models on the test set.

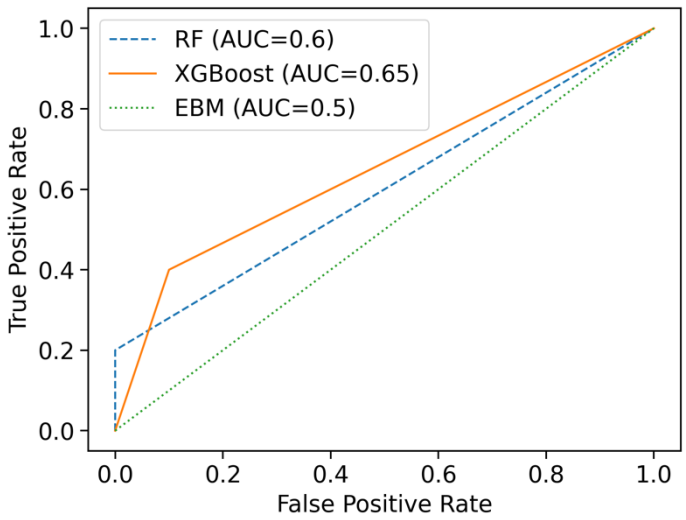


Figure 1: Area under the curve (AUC) of RF, XGBoost, and EBM models.

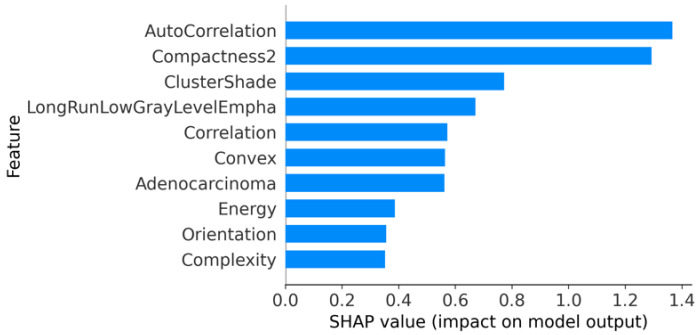


Figure 2: Features with the highest SHAP values identified by the XGBoost model.

4 Discussion

Although EBM models are often as accurate as black-box models, it was outperformed by XGBoost and RF models for our cohort. By incorporating the use of SHAP values, the RF and XGBoost models achieved similar interpretability as the glass box EBM model. Our best performing model achieved similar performance to a RF model in a previous radiomics-based study [9] despite being trained on a smaller dataset.

All models considered radiomic features to be most useful for recurrence prediction, while clinical features were not important. This may be because some clinical features were found to be correlated to radiomic features and therefore dropped during feature selection. A lower Pearson correlation coefficient cut-off threshold during the feature selection process may help decrease the influence of radiomic features. Note that the evaluation of SHAP values may be also performed for individual patients rather than at the level of the study population. A personalized SHAP assessment may place more weight on clinical features when making outcomes predictions. In addition, these features may not necessarily rank among the overall most important features due to patient variability.

Although the model explains the underlying 'motivation' behind predictions, it cannot recommend specific treatment plan modifications because it is based on radiomic and clinical features alone. An AI recommender system would involve specific features that can be modified by clinicians, such as details of the treatment plan. In our case, the model would support clinical decision making by indicating which patients may be at higher risk of recurrence and therefore justify an increased risk of toxicity in order to pursue a more aggressive treatment.

Patient data was acquired using standardized procedures in a provincial framework. The dataset is expected to be high quality, but is still likely to contain heterogeneity due to the inclusion of patients treated in different geographical regions. This study is limited by its use of a small subset of the available data. Future studies will leverage a larger provincial dataset and allow more efficient model training and robust analysis. A larger dataset will also provide a more complete assessment of the model's generalizability and better represent the five health authorities featured in this study.

Another limitation of the present work is the lack of recurrence examples present in the datasets, which is a common issue in medical studies. The percentage of recurrence cases in the training and testing datasets were balanced, but mild class imbalance was observed. To decrease this effect, resampling techniques may be warranted.

5 Conclusion

XAI methods were used to predict NSCLC recurrence in SABR patients across British Columbia and provide understandable explanations for its predictions. It was found that the XGBoost model architecture had the highest predictive performance. SHAP values showed that radiomic features were most informative for model predictions. Ultimately, XAI would provide outcomes recurrence predictions that can be trusted to support clinical decision making and potentially improve patient outcomes.

References

[1] National Comprehensive Cancer Network. "Non-small cell lung cancer (Version 1.2024)". Available: https://www.nccn.org/professionals/physician_gls/pdf/nscl.pdf

[2] J. Y. Chang. "Stereotactic ablative radiotherapy for stage I NSCLC: Successes and existing challenges". J Thorac Dis 3.3 (2011), pp. 144–146. DOI: 10.3978/j.issn.2072-1439.2011.06.02.

[3] S. Senthi, F. J. Lagerwaard, C. J. A. Haasbeek, et al. "Patterns of disease recurrence after stereotactic ablative radiotherapy for early stage

non-small-cell lung cancer: A retrospective analysis". Lancet Oncol 13.8 (2012), pp. 802–809. DOI: 10.1016/S1470-2045(12)70242-5.

[4] E. Huynh, A. Hosny, C. Guthier, et al. "Artificial intelligence in radiation oncology". Nat Rev Clin Oncol 17.12 (2020), pp. 771–781. DOI: https://doi.org/10.1038/s41571-020-0417-8.

[5] S. Cui, A. Traverso, D. Niraula, et al. "Interpretable artificial intelligence in radiology and radiation oncology". Br J Radiol 96.1150 (2023), p. 20230142. DOI: 10.1259/bjr.20230142.

[6] S. Lundberg and S.I. Lee. "A unified approach to interpreting model predictions". arXiv (2017). DOI: 10.48550/arXiv.1705.07874.

[7] L. Zhang, D. V. Fried, X. J. Fave, et al. "IBEX: An open infrastructure software platform to facilitate collaborative work in radiomics". Med Phys 42.3 (2015), pp. 1341–1353. DOI: 10.1118/1.4908210.

[8] H. Nori, S. Jenkins, P. Koch, et al. "InterpretML: A unified framework for machine learning interpretability". arXiv (2019). DOI: 10.48550/arXiv.1909.09223.

[9] C. Parmar, P. Grossmann, J. Bussink, et al. "Machine learning methods for quantitative radiomic biomarkers". Sci Rep 5.1 (2015), p. 13087. DOI: 10.1038/srep13087.

Is comorbidity associated with interruption of radiotherapy in patients with HNSCC cancer?

Golnoosh Motamedi – Ghahfarokhi[1], *Azadeh Abravan*[1], *Eliana M Vasquez Osorio*[1], James Price[1], Marcel van Herk[1,2], Hitesh Mistry[1], Gareth Price[1,2]

[1]Division of Cancer Sciences, The University of Manchester, *U*K
[2]The Christie NHS Foundation Trust, UK

Abstract In cancer management, timely completion of radiotherapy (RT) is crucial, as interruptions impact overall survival. We investigated whether comorbidity burden is associated with RT completion in a cohort of 1,877 patients with head and neck squamous cell carcinoma (HNSCC) who underwent curative-intent RT with/without chemotherapy. Multinomial logistic regression explored associations between treatment completion categories, comorbidities, and overall comorbidity scores. Patients with moderate and severe overall comorbidity scores had a poorer survival than those with no comorbidity. Based on the ΔAIC analysis, accounting for each comorbidity category has the most pronounced effect on predicting treatment completion (ΔAI>4). The risk of not completing treatment increased for patients with moderate comorbidity score, and for patients with cardiovascular comorbidities. The risk of late treatment completion increased for patients with both respiratory and cardiovascular comorbidities. We found that comprehensive consideration of each comorbidity category significantly enhanced the accuracy of predicting treatment completion.

1 Introduction

Radiotherapy (RT) plays an important role in the management of patients with cancer. One of the key elements in radiation oncology is timely completion, which is associated with treatment outcome, including overall survival (OS) [1]. In addition, higher comorbidity scores are associated with poorer treatment outcomes and incomplete treatment [2] However, the degree to which comorbidity burden affects RT interruption is not fully understood. Therefore, in this study, we investigated the association between RT completion and OS and furthermore assessed whether overall score comorbidities and baseline comorbidities are associated with RT completion in a large cohort of patients with head and neck squamous cell carcinoma (HNSCC) cancer treated with RT.

2 Materials and Methods

Data were collected from 1,877 HNSCC patients treated at a single institute with curative-intent RT (60/63/66 Gy in 30 fractions, with/without concurrent or sequential chemotherapy) between January 2013 and February 2023 and for whom Adult Co-Morbidity Evaluation (ACE27) score were available. RT was given to patients on weekdays over six weeks. Overall treatment duration (days) was defined as the number of weekdays from treatment initiation to the last fraction received. ACE-27 comorbidity burden was collected at diagnosis and were considered as binary variables (with/without comorbidity) for twelve organ systems (3). Incidence of all comorbidities was calculated, but we only report on those exceeding a 5% occurrence to ensure the robustness of our outcome. Besides ACE-27 comorbidity burden, we also had overall comorbidity scores which were scored as none, mild, moderate, and severe based on the validate comorbidity instrument ACE-27 [3]. Treatment completion was categorized into three groups, patients who did not complete the treatment, those who completed the treatment in the scheduled 30 days, and those who completed the treatment after more than 30 days (late completion). The association of treatment completion and OS was assessed via a multivariable Cox regression model. The association of treatment completion category with different comorbidities and overall comorbidity score was assessed via individual multinomial logistic regression models. The goodness of fit of all multinomial models was evaluated using AICs. All models were adjusted for age, gender, performance status, log gross tumour volume, chemotherapy and multiple deprivation index (IMD). IMD deciles were calculated by ranking geographic areas of the UK from most to least deprived and dividing them into 10 equal groups (decile 1, most deprived 10%; decile 10, least deprived 10%).

3 Results

1,655 (88%) patients completed treatment in time, 180 (10%) patients had late completion and 42 (2%) patients did not complete their treatment. Cardiovascular (40%), respiratory (14%), and endocrine (11%) conditions were the most common comorbidities. Overall comorbidity scores showed almost half of patients had no diagnosed comorbidity (46%), and 35%, 14% and 5% of patients had mild, moderate, and severe comorbidity burdens, respectively. An adjusted multivariable cox regression model revealed that the survival of patients did not complete treatment was significantly worse than those who did complete treatment (HR (95% CI) = 3.32 (1.84, 5.99), P < 0001), Figure 1. Also, patients with moderate overall comorbidity scores (HR (95% CI) = 1.89 (1.27, 2.80), P = 002) and severe overall comorbidity scores (HR (95% CI) = 1.87 (1.11, 3.14), P = 0.019) had worse survival than those with no comorbidity. We fitted various

adjusted multinomial logistic regression models to assess the most influential predictor of comorbidity associated with the treatment completion category. The AIC for the model incorporating overall comorbidity score was 516.42, while including all comorbidities it was 513.53. The model with categorized comorbidities had an AIC of 515.82, and the lowest AIC among individual categories was observed for cardiovascular disease at 507.22. Considering the ΔAIC, there is strong evidence that factoring in all comorbidity categories would provide the most significant prediction of treatment completion (ΔAIC = 9.2). The adjusted multinomial logistic regression revealed the risk of not completing treatment increased for patients with moderate (OR (95%CI) = 9.93 (1.91, 51.62), P = 0.006) overall comorbidity. No significant association was found between the risk of late completion and overall comorbidity scores. The risk of not completing treatment increased for patients with cardiovascular conditions at baseline (OR (95%CI) = 6.70 (1.98, 22.63), P = 0.002), Figure 2. The risk of late completion was not significantly associated with other comorbidities. An adjusted multinomial logistic regression with interaction of cardiovascular and respiratory comorbidities indicated the risk of late treatment completion increased for patients with both respiratory and cardiovascular comorbidities (OR (95%CI) = 6.83 (1.26, 37.00), P = 0.03), Figure 2. The risk of not completing treatment increased for patients with cardiovascular conditions at baseline (OR (95%CI) = 6.70 (1.98, 22.63), P = 0.002), Figure 2. The risk of late completion was not significantly associated with other comorbidities in our cohort. An adjusted multinomial logistic regression with interaction of cardiovascular and respiratory comorbidities indicated the risk of late treatment completion increased for patients with both respiratory and cardiovascular comorbidities (OR (95%CI) = 6.83 (1.26, 37.00), P = 0.03), Figure 2.

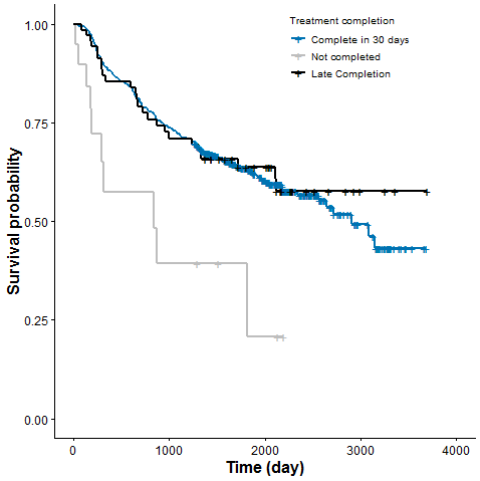


Figure 1 – Kaplan-Meier of adjusted Cox regression model according to RT treatment completion. Not completed the RT relative to completed RT is associated with worse survival. Also, age, performance status, smoking, and log(GTV) were associated with overall survival.

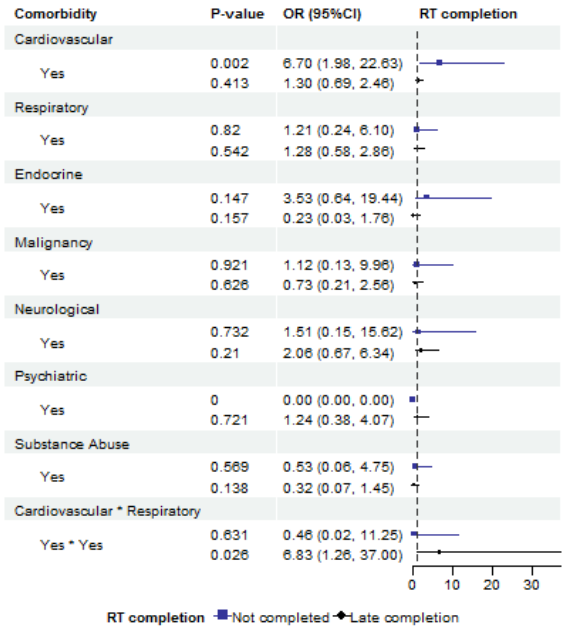


Figure 2 – Forest plot of adjusted multinomial logistic regression model including an interaction between cardiovascular and respiratory. Risk of not completing treatment relative to complete RT treatment in 30 days is increased in patients with cardiovascular comorbidities. Risk of late completing RT is increased with presence of both respiratory and cardiovascular comorbidities.

4 Discussion and conclusion

In this cohort of HNSCC patients treated with curative intent (chemo) radiotherapy, treatment completion was associated with improved OS, but late completion had no effect. Risk of not completing RT is associated with increasing comorbidity score, regardless of late completion. Our findings suggest that considering each comorbidity category generates a substantial improvement on the prediction of treatment completion. This underscores the importance of comprehensive comorbidity assessment for tailoring effective strategies for optimizing treatment outcome. In addition, individual comorbidities are associated with treatment completion. Risk of not completing was associated with presence of cardiovascular comorbidities and risk of late treatment completion was associated with the presence of both respiratory and cardiovascular comorbidities. This cohort come from observational data and includes small number patients who did not complete RT treatment, consequently the prevalence of comorbidities were low among this group.

Therefore, the obtained outcome came from no sufficient data. Further research is required to validate these associations and to propose methods to stimulate treatment completion.

References

[1] Tubiana M. Repopulation in human tumors. A biological background for fractionation in radiotherapy. Acta Oncol. 1988;27(2):83-8. DOI:10.3109/02841868809090328

[2] Sarfati D, Koczwara B, Jackson C. The impact of comorbidity on cancer and its treatment. CA Cancer J Clin. 2016;66(4):337-50. DOI: 10.3322/caac.21342

[3] RCP. (n.d.). Adult co-morbidity evaluation (ACE-27) | RCP London. 2–3. https://www.rcplondon.ac.uk/file/3058/download

Prediction of Acute Pulmonary Toxicity Events with 3D Convolutional Neural Networks from Radiotherapy Dose Maps

Pedro J Soto Vega[1], Vincent Bourbonne[1,2], Wistan Marchadour[1], Gustavo Andrade-Miranda[1], Francois Lucia[1,2], Martin Rehn[2], Ulrike Schick[1,2], Dimitris Visvikis[1], Franck Vermet[3,1], and Mathieu Hatt[1]

[1]Laboratory of medical information processing (LaTIM), INSERM UMR 1101, Univ Brest, Brest, France
[2]Radiation Oncology Department, University Hospital of Brest, Brest, France
[3]Laboratory of mathematics of Atlantic Britanny (LMBA), CNRS UMR 6205, Univ Brest, Brest, France

Abstract Predicting toxicity events in radiation therapy (RT) is highly beneficial for managing patients effectively. Identifying patients who are at a high risk of experiencing toxicity early on during their treatments can help in taking measures to reduce the risk of adverse events produced by this undesirable effect. Recent works in a related application, namely, acute pulmonary toxicity (APT) in lung cancer patients treated by RT, have demonstrated high accuracy in predicting such an event using dose maps features processed by a multilayer perception network. Thus, motivated by the success of convolutional neural networks (CNN) in learning semantically rich representation directly from images, this work investigates the suitability of CNN architectures in predicting APT directly from dose maps. Our results demonstrate the ability of some CNN models to predict APT from planning dose maps with an accuracy of up to 81% in terms of receiver operative characteristic's area under the curve. However, most of the architectures and configurations under evaluation led to non-satisfactory accuracy, as only shallower architectures using resized dose maps as inputs were able to train models with good accuracy in the testing set.

1 Introduction

Prediction of toxicity events related to radiotherapy (RT) could be very useful for better managing patients and personalizing treatment and follow-up. Indeed, identifying patients at risk of suffering from these toxicity events early in the course of patients monitoring (ideally, before initiating RT), could allow, in the best case scenario, modifying treatment (optimization of treatment planning, de-escalation of dose to sub-volumes of the organs-at-risk (OAR) most responsible for the toxicity, etc.) to reduce the risk, or at least, identifying patients that could benefit from intensified monitoring after RT to better prevent and treat toxicity symptoms. Recently, we have shown that analysis of dose maps can provide predictive markers of toxicity events in cervical [1] and lung cancer [2]. We used radiomics engineered features (i.e., intensity or textural metrics) extracted from the delineated OAR in the dose maps to train multiparametric models that demonstrated higher predictive value than the usual dose-volume histogram (DVH) approach. In another work, we developed an alternative approach where all dose maps were co-registered to a common spatial reference, and a specific region in the lung most correlated with the acute pulmonary toxicity (APT) event was identified through statistical analysis on a voxel-by-voxel level in the dose maps. Dose map features (e.g., mean dose) from this specific area, combined with clinical variables through a multilayer perceptron (MLP), allowed good accuracy (AUC 81%) in predicting patients with APT[3, 4]. The main advantage of this approach was the ability to both predict which patients are most likely to suffer from APT and to identify a spatial area in the lung that was most responsible. In the present work, we investigated the feasibility of achieving similar predictive power by relying on deep learning (DL) convolutional neural network (CNN) architectures trained using the dose maps as inputs, with or without the help of clinical variables. Our long term goals are i) to achieve similar (if not better) predictive power compared to the previous approach, ii) to provide some explainability of the prediction made by the network and thereby identify the area in the lung most responsible for the APT, on a patient-by-patient basis. Objective ii) will be investigated using interpretability tools such as saliency maps of the trained CNN model. Preliminary results will be reported in the present paper, focusing on reporting methods and results of objective i), while reporting on first results using saliency maps for objective ii).

2 Materials and Methods

2.1 Workflow

The APT classification task was conducted following the procedure presented in Figure 1, where the prediction model takes two inputs: the RT dose maps and clinical features (e.g., age, gender, tumor location...). The dose maps, in the figure represented by X_1 are forwarded through a CNN backbone, e.g., ResNet or DenseNet, from which a feature vector of size 512 is obtained and later concatenated with the clinical features X_2 to serve as input to a set of fully connected layers which delivers the final prediction outcome.

2.2 Dataset and pre-processing

The dose maps as images were previously generated and co-registered to a common spatial reference, namely a thoracic phantom, using a segmentation-based elastic registration via MIM Maestro (MIM v7.0.0, Cleveland, OH, USA) [3]. The segmentation used for registration was the volume combining the lungs and the heart. The dataset contains dose maps of dimension $512 \times 512 \times 237$ from 207 patients, out

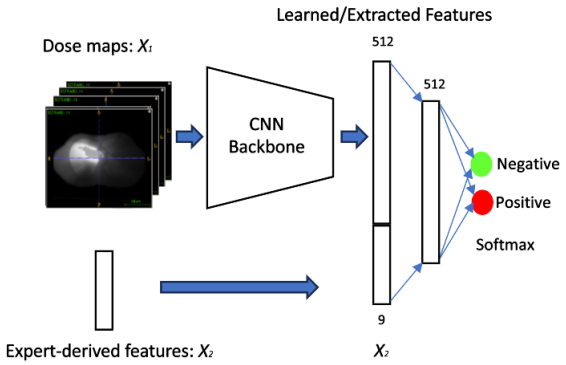


Figure 1: Proposed framework overview.

of which 45 experienced APT whereas 164 did not. Additionally, the dataset includes 37 clinical features for each patient, as described in [3]. Expert-derived knowledge considered as additional input consisted of features previously identified as predictive, i.e., Volume of the heart receiving at least 40Gy ($V40_{Heart}$), Volume of the homolateral lung receiving at least 10Gy ($V10_{LungH}$), Mean dose received by the two lungs ($DMean_{2Lungs}$), American Joint Committee of Cancer (*AJCC Stage*), Chronic Obstructive Pulmonary Disease (*COPD*), Mean Expiratory Volume/Second (*MEVS*), and smoking status [3]. The training set included the same 165 patients used in [3], while the remaining 45 patients were included in the test set. For validation purposes during training, we isolated 20% of the training samples.

2.3 CNN architectures

When using dose maps as input, 3D ResNet [5] and DenseNet-based [6] architectures were used, specifically, ResNet-10, 18, 34, 50, and DenseNet-121. We note that diverse other image classification architectures have been proposed, some with outstanding performance on different applications. To the best of the authors' knowledge, no specific architecture has been designed so far for APT classification. Therefore, we opted to evaluate seminal, general-purpose models, extensively used and well-documented, which we believe would facilitate the reproducibility of our experiments and results.

Succinctly, Residual Networks (ResNet), introduced by [5], aimed at improving convergence issues while training very deep network architectures. ResNet addressed the problem of vanishing gradients, which hinders the optimization process, and the degradation problem, i.e., adding more layers to a deep model leads to higher training error. Such problems, including residual learning blocks among the network layers, were dealt with. The degradation problem suggested that when the network's full capacity was underused for solving a particular task, the optimization process would have difficulty approximating nonlinear layers into identity mappings, which could automatically adjust network depth. Then, instead of hoping that every few stacked layers directly fit a desired underlying mapping, e.g., identity, ResNet explicitly lets those layers fit a residual mapping, which is easier to optimize. DenseNet, on the other hand, introduced by [6], similar to ResNet [5], addresses the vanishing gradient problem by adding densely connected blocks, which consists of linking each layer to any other previous layer. This allows the network to learn more efficiently by reusing features and reducing the number of trainable parameters. The principle behind this proposal states that concatenating the feature maps learned in previous layers allows each layer to access all features of all preceding levels.

2.4 Experiments implementation

In the experiments conducted in this work, we considered different input dose map sizes, see Table 1, the use (✓) or not (×) of data augmentation, and to take into account (✓) or not (×) the dataset imbalance. Each dose map and clinical features, except for AJCC Stage, was normalized with min-max normalization between 0 and 1. AJCC Stage feature was converted from categorical to binary variable. When data augmentation was considered, Gaussian noise, smooth, sharpen, and histogram shift were applied randomly. We used the MONAI library[1] to perform these transformations, as well as to set the network architectures. We used the Adam [7] optimizer and learning rate decay during training following [8]. We set the initial learning rate μ_0 and momentum β_1 equal to 0.0001 and 0.9, respectively. The batch size was set to 2, and the early stopping procedure was used to avoid over-fitting after 20 epochs without signals of accuracy improvements in the validation set.

The training/testing procedures were executed three times, each with a different (random) initialization of the trainable parameters. When class imbalance was addressed in experiments, we adopted a *weighted cross-entropy* cost function to compensate for class imbalance to train the networks, by assigning larger weights to the underrepresented ones, i.e., 0.2 for the negative and 0.8 for the positive class. Those weights were set experimentally.

2.5 Evaluation

Following [3]; the performance of models was evaluated using Receiver Operative Characteristic's (ROC) Area Under Curve (AUC). Considering each network architecture was trained and tested three times, the final performance is expressed in terms of the average AUC. It is worth noting that the standard deviation over the three runs was very low and thus not included in the table for space purposes.

3 Results

Table 1 presents the AUC (average of 3 runs) of each model in predicting APT events according to different criteria: using only dose maps, clinical variables, or both.

[1] https://monai.io/

Input	Clinical only				Dose Map only												Dose Map & Clinical											
Input Size	9				64 × 64				128 × 128				256 × 256				64 × 64				128 × 128				256 × 256			
Balance	×		✓		×		✓		×		✓		×		✓		×		✓		×		✓		×		✓	
Data Aug.	×	✓	×	✓	×	✓	×	✓	×	✓	×	✓	×	✓	×	✓	×	✓	×	✓	×	✓	×	✓	×	✓	×	✓
MLP	43.7	56.0	46.5	60.2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
DenseNet121	-	-	-	-	73.7	69.2	67.4	68.5	76.9	73.2	73.9	74.7	72.0	75.3	64.7	73.7	65.2	64.8	70.9	73.0	72.5	73.3	71.7	71.8	74.2	74.8	73.5	74.4
ResNet10	-	-	-	-	**80.2**	61.2	**80.4**	71.3	67.0	78.4	79.9	76.7	25.3	41.0	73.0	74.9	**80.9**	63.3	73.6	61.2	76.0	66.8	78.5	66.9	41.5	71.4	73.0	47.5
ResNet18	-	-	-	-	77.7	60.6	79.9	63.2	74.2	75.8	76.8	72.3	37.3	37.0	61.5	75.4	76.6	65.4	71.9	58.2	77.7	67.9	76.9	76.8	31.1	61.8	72.3	68.7
ResNet34	-	-	-	-	47.1	57.1	76.6	54.2	39.1	71.7	80.0	59.1	36.7	48.3	50.5	66.1	64.9	50	67.9	65.6	79.5	72.5	77.9	53.7	38.1	43.7	50.7	53.4
ResNet50	-	-	-	-	49.8	44.5	40	59.9	25.2	59.0	39.6	22.7	47.0	23.8	52.6	50.9	58.1	42.3	59.7	53.8	22.5	59.6	70.9	59.1	37.8	43.3	41.0	44.0

Table 1: Average AUC (%) for 3 executions of models.

First of all, results obtained by using only expert-derived knowledge as input (i.e., features previously identified in [3]) to a MLP, the highest performance of 60.2 % of AUC was obtained with balancing and data augmentation. On the other hand, the highest performances were obtained when using dose maps as input in dimensions of 64 × 64 and shallower CNN architectures, i.e., ResNet10, for which the best-achieved performance was 80.2% (80.4%, without data augmentation). Addressing the imbalance using weights in the cost function delivered the best AUC (80.4%) among these two results. Still considering dose maps-based results, ResNet50 obtained the lowest performance among all the experiment's configurations, i.e., around 22.5% for the lower result and 70.9% for the highest, which corresponded to the dimension 128 × 128 input dimension and without data augmentation. In the remaining experimental configurations, ResNet50 achieved results between 40% and 60% AUC.

Considering the results where dose maps and clinical variables were jointly used as inputs, the ResNet10 architecture outperformed the remaining evaluated architectures with 80.9% AUC. Such a result was reached without data augmentation and weighted cross-entropy to alleviate the imbalance. Although still low, It is worth highlighting the significant increase obtained by ResNet50's performance when clinical variables and dose maps were added.

The highest results, e.g., 80.2%, 80.4%, and 80.9%, were obtained with dimension 64 × 64, no data augmentation, and with the shallowest architecture, namely ResNet10. On the other hand, regardless of the use of data augmentation and *weighted cross-entropy*, consistently lowest performances were achieved with the larger dose maps size (256 × 256) and the deepest CNN architecture, ResNet50, with AUCs from 23.8% to 52.9% with only dose maps, and from 37.8% to 44% when both types of data were forwarded through the proposed prediction model. The results obtained by the DenseNet121 architecture were between those achieved by ResNet10 and ResNet50, i.e., in the range 64.7% to 76.9%.

4 Discussion

Our results suggest it is feasible to achieve satisfactory prediction of APT events in lung cancer patients following RT by training a CNN with co-registered dose maps as input. Interestingly, the overall accuracy reached by our best models (around 80% AUC) is similar to that reached through the previous approach [3], which relied on a combination of clinical and dosimetry features extracted from the lung area identified as most correlated with APT events. We should, however, emphasize that this level of accuracy could be reached without any clinical variables as inputs, as the results were not significantly improved using both dose maps and clinical variables, compared to the use of dose maps alone, whereas the use of clinical variables alone led to poor predictive power (60% AUC or less). Another important observation is that the choice of the CNN architecture, the use of balancing and data augmentation, as well as the choice of the input dose maps resize had a strong impact on the final performance. The best results were obtained with shallower architectures (ResNet 10), without data augmentation or balancing, and using smaller dose maps (64 x 64). We also tested using the original size of dose maps as input (512 x 512), but the results were even poorer, so we did not include them.

Considering the results obtained by larger CNN architectures, e.g., ResNet50, it is well known that the number of training labeled samples required for efficient learning increases with the depth of the network. Since ResNet50 contains significantly more trainable parameters than the other variants of ResNet, it is not surprising that it performs worse given the amount of available data. By contrast, the few configurations that led to the best performance in the testing (around 80% AUC) corresponded to the shallowest architectures. Additionally, with respect to this last point, the performance (around 70% AUC) of DenseNet deserves to be highlighted, as even though DenseNet encompasses more layers than ResNet50, its performance was consistently superior, which may be due to the dense blocks to alleviate the vanishing gradient problem.

Although we focused here on our first objective, we include in Figure 2 a preliminary result of using saliency maps generated with Grad-CAM [9] to illustrate how these could be further exploited in our future work to provide some interpretability of the CNN prediction output, as well as to confirm the location of the area most responsible for the APT event occurrence. The figure shows an overlaid representation of the phantom used for the co-registration of dose maps and the average of saliency maps computed by Grad-CAM. Note that the average saliency maps were computed only over the samples correctly classified by the CNN model, separately

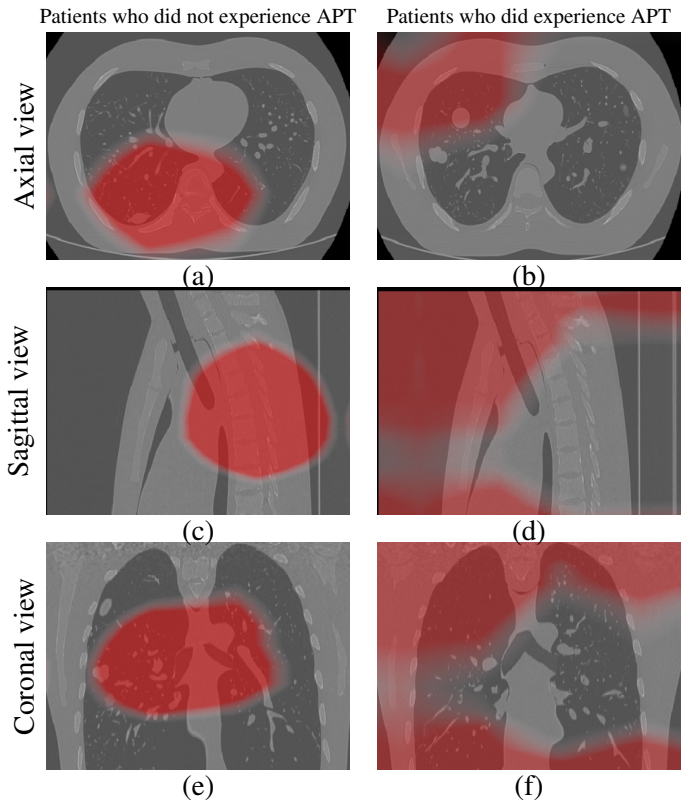


Figure 2: Saliency maps representations computed over ResNet10 architecture using Grad-CAM algorithm. The average of all saliency maps is overlaid on the phantom used for co-registration to visualize the localization of the parts of the dose maps relied upon by the CNN in axial (a/b), sagittal (c/d), and coronal (e/f) views for classifying patients without/with APT respectively.

for cases with and without APT. According to these saliency maps, prediction of the absence of APT events is achieved by the CNN by relying mostly on the dose information contained in lower lung areas, whereas prediction of the occurrence of APT events is achieved by relying on features located in higher lung areas but also strangely highlights top and bottom (coronal/sagittal view) of dose maps where the dose is near-zero, which is a bit puzzling. More work is necessary to understand better the explainability value of these saliency maps (including the comparison with other methods such as integrated gradient with or without smoothgrad) [10].

5 Conclusion

This study evaluated five 3D CNNs for predicting toxicity events based on 3D RT planning dose maps and associated clinical features. We also reported on the potential of Grad-CAM saliency maps to provide some interpretability of the toxicity predictions. Our results suggest that CNN can perform satisfactorily in this classification task. However, only a small part of the configurations of CNN architectures we tested led to good predictive performance. There is clearly room for improvement regarding the generalization capacity of these models. This could be achieved by better preprocessing the training data, using additional data augmentation techniques, or by simplifying the models regarding the number of parameters to avoid overfitting.

6 Acknowledgments

This study was funded by the CHIST-ERA grant [CHIST-ERA-19-XAI-007] with project acronym INFORM, by General Secretariat for Research and Innovation (GSRI) of Greece, National Science Centre (NCN) of Poland [2020/02/Y/ST6/00071] and Agence Nationale de la Recherche (ANR) of France.

References

[1] F. Lucia, V. Bourbonne, D. Visvikis, et al. “Radiomics analysis of 3D dose distributions to predict toxicity of radiotherapy for cervical cancer”. *Journal of Personalized Medicine* 11.5 (2021), p. 398.

[2] V. Bourbonne, R. Da-Ano, V. Jaouen, et al. “Radiomics analysis of 3D dose distributions to predict toxicity of radiotherapy for lung cancer”. *Radiotherapy and Oncology* 155 (2021), pp. 144–150.

[3] V. Bourbonne, F. Lucia, V. Jaouen, et al. “Development and prospective validation of a spatial dose pattern based model predicting acute pulmonary toxicity in patients treated with volumetric arc-therapy for locally advanced lung cancer”. *Radiotherapy and Oncology* 164 (2021), pp. 43–49.

[4] V. Bourbonne, F. Lucia, V. Jaouen, et al. “Combination of Radiomics Features and Functional Radiosensitivity Enhances Prediction of Acute Pulmonary Toxicity in a Prospective Validation Cohort of Patients with a Locally Advanced Lung Cancer Treated with VMAT-Radiotherapy”. *Journal of Personalized Medicine* 12.11 (2022), p. 1926.

[5] K. He, X. Zhang, S. Ren, et al. “Deep residual learning for image recognition”. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 2016-Decem (2016), pp. 770–778. DOI: 10.1109/CVPR.2016.90.

[6] G. Huang, Z. Liu, L. Van Der Maaten, et al. “Densely connected convolutional networks”. *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 4700–4708.

[7] D. P. Kingma and J. Ba. “Adam: A method for stochastic optimization”. *arXiv preprint arXiv:1412.6980* (2014).

[8] Y. Ganin, E. Ustinova, H. Ajakan, et al. “Domain-adversarial training of neural networks”. *Journal of machine learning research* 17.59 (2016), pp. 1–35.

[9] R. R. Selvaraju, M. Cogswell, A. Das, et al. “Grad-cam: Visual explanations from deep networks via gradient-based localization”. *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 618–626.

[10] L. Brocki, W. Marchadour, J. Maison, et al. “Evaluation of importance estimators in deep learning classifiers for Computed Tomography”. *International Workshop on Explainable, Transparent Autonomous Agents and Multi-Agent Systems*. Springer. 2022, pp. 3–18.

Deep learning based Three-dimensional Radiation Dose Prediction with Dose level Guidance - a Spine SBRT example

Tushar Shinde[1], Gregory Arthur Szalkowski[1], Cynthia Chuang[1], Lei Wang[1], Lianli Liu[1], Scott Soltys[1], Erqi Pollom[1], Elham Rahimy[1], Quoc-Anh Ho[1], Neelan Marianayagam[2], David Park[2], Ulas Yener[2], Antonio Meola[2], Hao Jiang[3], Mingli Chen[4], Zi Yang[1,4], Weiguo Lu[4], and Xuejun Gu[1,4]

[1]Department of Radiation Oncology, Stanford University, School of Medicine, Palo Alto, CA
[2]Department of Neurosurgery, Stanford University, School of Medicine, Palo Alto, CA
[3]NeuralRad LLC, Madison, WI
[4]Department of Radiation Oncology, University of Texas, Southwestern Medical Center, Dallas, TX

Abstract
Deep learning models have emerged as promising tools in predicting radiation doses, but their application to spine stereotactic body radiotherapy (SBRT) encounters challenges like critical structure proximity and dose heterogeneity, resulting in sub-optimal clinical accuracy. We introduce an innovative workflow for three-dimensional radiation dose prediction. Our goal is to enhance the accuracy of 3D dose prediction tasks using dose-level guidance. We propose a two-stage model approach: a dose level prediction model employing nnU-Net and a voxel-wise dose prediction model utilizing a cascaded 3D U-Net. The first model predicts clinically relevant dose levels. Whereas the second model incorporates the weighted predicted dose levels as an additional input to generate precise voxel-wise radiation dose predictions. We trained and evaluated these models using a dataset of 102 spine SBRT plans, comparing them to a baseline cascaded 3D U-Net. Our comprehensive evaluation involved metrics such as Dice score, dosimetric indices (D2, D98, and Dmean) for planning target volumes (PTVs) and 0.035 cc for spinal cord volumes, as well as clinical indices (conformity index, Paddick conformity index, and gradient index). The results showed a superior performance as compared to the baseline approach, demonstrating the potential of dose level guidance to boost the performance of radiation dose prediction tasks.

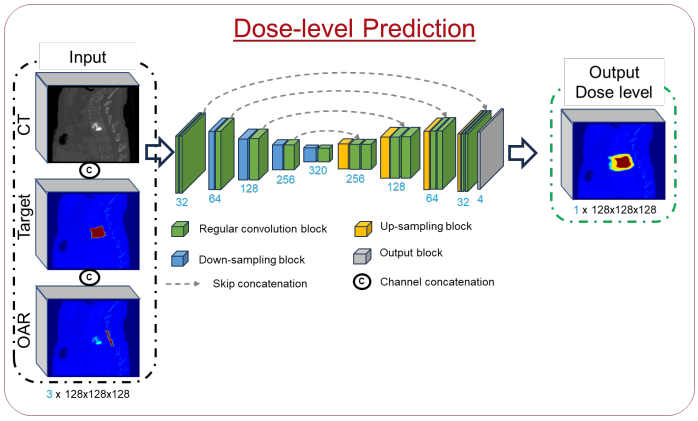
1 Introduction

In recent years, the emergence of stereotactic body radiotherapy (SBRT) has shown promise as an effective therapeutic modality for spine cancer. However, due to the proximity of critical radio-sensitive organs at risk (OARs), such as the spinal cord, the treatment planning process becomes increasingly complex. Also, traditional treatment planning is a time-consuming and labor-intensive process. To address these issues, in literature, various studies have explored the integration of machine learning (ML) and deep learning (DL) techniques to enhance dose distribution prediction [1–3]. The U-Net architecture [4, 5], initially designed for image segmentation, had gained significant popularity in predicting dose distribution without the need for conventional inverse optimization processes. Despite recent progress on U-Net-based 3D dose predictions [6], focusing mainly on prostate and head/neck cancers, spine cancer presents distinct challenges such as critical structure proximity and dose heterogeneity, among others, making it difficult to design an effective model [7].

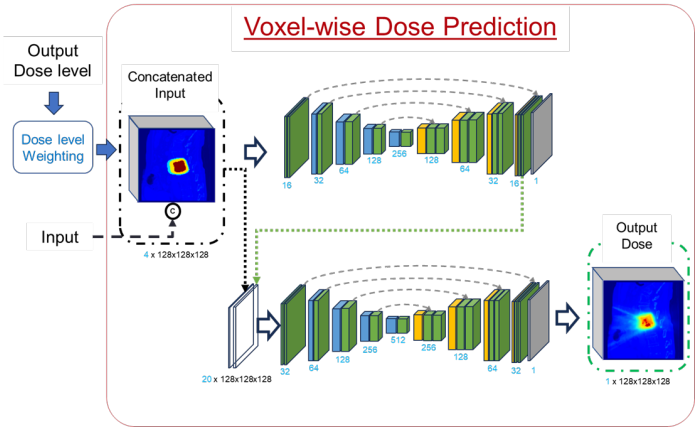
In this study, we introduce an innovative workflow consisting of DL-based radiation dose prediction with dose level guidance in spine SBRT. Our approach shown in Figure 1, consists of two key stages: first, predicting 3D dose levels, and second, leveraging a weighted version of the predicted dose levels to predict 3D dose distributions for spine cancer patients. Here, we introduce an innovation by incorporating the weighted version of the predicted dose levels as an additional input to train a conventional dose prediction network.

2 Materials and Methods

Data: An in-house institutional dataset used in this study consists of 102 spine SBRT plans, including cases of cervical-spine, thoracic-spine, lumbar-spine, and sacral-spine malignancies. The PTV volumes ranged between 8.312 cm³ and 131.280 cm³, with an average of 38.492 ± 30.303 cm³. The target prescription dose for the plans ranged from 16 Gy to 30 Gy, with an average of 20.609 ± 2.581 Gy. The dataset was randomly divided into training (85) and testing (17) sets. Network architecture: Cascading models have shown great ability in dense prediction tasks [6] and in this study, the off-the-shelf segmentation and traditional 3D dose prediction models cascaded for the first time for 3D dose prediction tasks. In this pursuit, the first model is used to predict the dose-level using a segmentation network, and the second model is used to further predict 3D dose distribution. The proposed workflow is a combination of the following two stage models: a) Dose level prediction model, b) Voxel-wise dose prediction model. The schematic of our proposed workflow for the three-dimensional radiation dose prediction is shown in Figure 1. The input of the dose level prediction model has three channels (CT image, PTV mask, and OAR mask) and the output has four channels corresponding to four dose levels such as dose level 3 (> 100 % of Rx), dose level 2 (50-100% of Rx), dose level 1 (40-50% of Rx), and dose level 0 (<40% of Rx). The predicted dose level is further used as an additional input to the voxel-wise dose prediction model. In this pursuit, we introduce a dose level weighting scheme to guide their utilization before employing



(a) a dose level prediction using nnU-Net [8] with three inputs: CT image, target (PTV) mask, and OAR mask, and a 4-channel dose level output



(b) a voxel-wise dose prediction using cascaded 3D U-Net [6] with four inputs. The fourth channel is a weighted predicted dose levels

Figure 1: Schematic diagram of the proposed workflow and network architecture.

it as an input to the voxel-wise dose prediction task. To this end, we propose a weight vector, denoted as [w0, w1, w2, w3], which can be empirically chosen to adjust the dose levels 0, 1, 2, and 3 respectively. As a result, the voxel-wise dose prediction model has four (3+1) input channels and an output dose map. To make the image resolution consistent among all data, images were resampled to a voxel size of 2mm×2mm×2mm and cropped to the image size of 128×128×128. The resampling is carried out in such a way that the PTV is located at the center after resampling.

3 Results

The evaluation of the first stage, dose level prediction model for spine SBRT plans is done through dice scores. Dice scores of 0.901 ± 0.034, 0.756 ± 0.082, 0.385 ± 0.109 are achieved across dose levels 3, 2, and 1, respectively. This effectively demonstrates the model's strength in predicting dose level 3, its notable performance with dose level 2, and the challenges encountered when predicting dose level 1. High dice scores for dose levels 2 and 3 in the first stage model indicate a better initial estimate which could help drive the second stage model for the three-dimensional dose prediction task. To analyze the dose prediction results for the second stage model we compare the dose map for the baseline C3D U-Net [6] dose prediction model and the proposed model. Figure 2 shows some examples of voxel-wise dose prediction from the proposed model and baseline model. It can be visually seen that our proposed method is best in detail in the three-dimensional dose prediction task.

Furthermore, we present the comparison for critical dose-volume indices clinical parameters, including Dmean, D98, and D2 for PTVs and 0.035 cc for the spinal cord volume. Figure 3 shows a visual illustration of these critical dose-volume indices clinical parameters (measured in Gy). The clear visual distinction reveals that our proposed approach tends to closely align with the clinical reality, reflecting a more accurate prediction of these parameters. Also, we present the dosimetric clinical parameters such as Conformity Index (CI), Paddick Conformity Index (PCI), and Gradient Index (GI) for evaluating the quality of the dose predictions. Figure 4 shows a visual illustration of CI, PCI, and GI, as assessed across the actual dose, nnU-Net [8] based dose level prediction, Baseline C3D U-Net [6] based dose prediction, and Proposed dose prediction. These results indicate a better resemblance to the actual dose distribution. Whereas the predicted doses obtained through the baseline C3D U-Net method exhibit poor alignment with the actual dose.

4 Discussion and Conclusion

We have developed a new deep learning-based workflow for 3D radiation dose prediction and tested on an in-house spine SBRT dataset. It is observed that the predicted 3D dose distribution of the proposed approach is more accurate than those of the other widely adopted dose prediction approaches. Here, we analyzed the new paradigm of deep learning-based 3D dose prediction with dose level guidance. In this direction, we presented a two-stage model approach combining a dose-level prediction task and a voxel-wise dose prediction task. The weighted version of the output of the first model is fed as input to the second model to enhance the prediction accuracy. Certain efforts were made to explore the various weighting schemes in improving the performance of the dose prediction approach on spine SBRT plans. Secondly, the evaluation of the spinal cord volume (0.035 cc) revealed limitations in alignment with actual doses.

In the future, we could further investigate the adaptability of our approach to different treatment sites, and leveraging larger datasets can enhance its generalizability in various clinical settings. The weight vector scheme introduced in this study, based on the mean of dose level ranges, opens new avenues for fine-tuning radiation therapy planning.

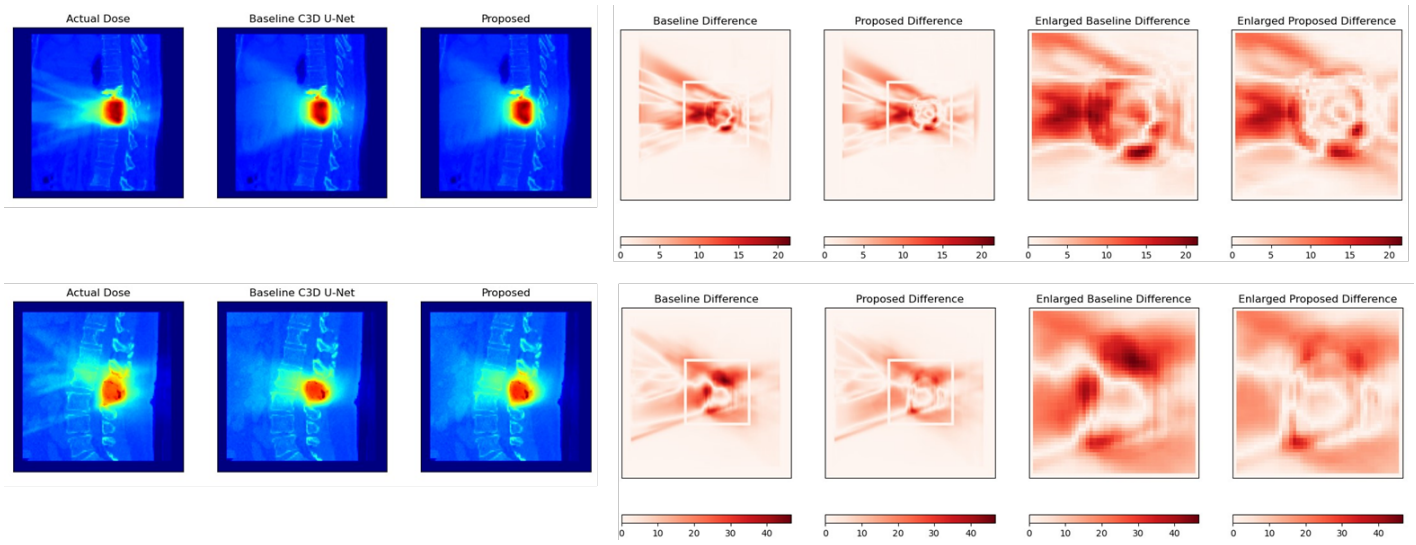


Figure 2: Comparison of actual dose vs. predicted dose for two test cases. Each row features columns, presented from left to right, depicting the following: the actual dose, the baseline C3D U-Net [6] dose, the proposed dose, and corresponding dose difference maps (in % of Rx). The dose difference maps include the baseline difference between the actual and baseline dose, the proposed difference between the actual and proposed dose, and corresponding enlarged difference maps.

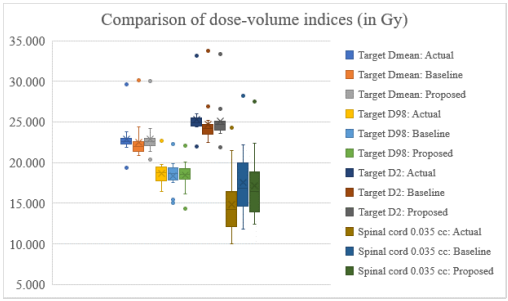
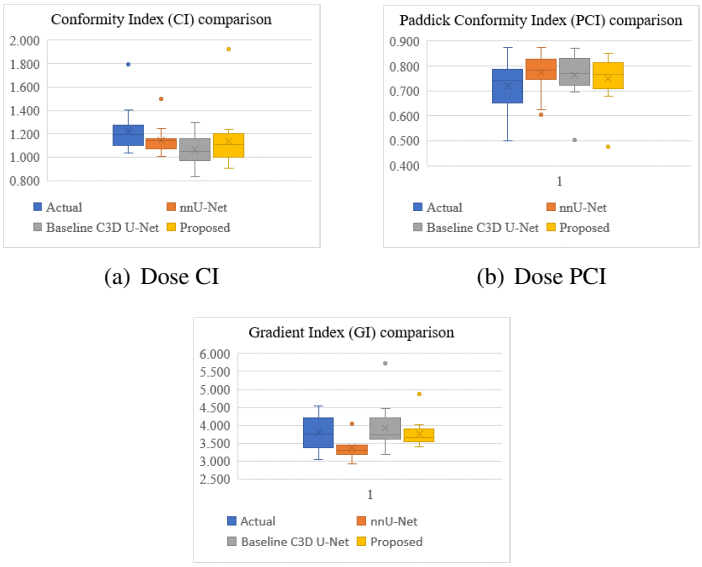


Figure 3: Illustration of dose-volume indices clinical parameters (in Gy) for PTV and Spinal cord volumes. This figure compares dose-volume indices clinical parameters between the actual dose, the baseline C3D U-Net [6], and our proposed approach.



(a) Dose CI (b) Dose PCI

(c) Dose GI

Figure 4: Comparison of Dose Conformity Index (CI), Paddick Conformity Index (PCI), and Gradient Index (GI) across four scenarios: the actual dose, nnU-Net-based [8] dose-level prediction, Baseline C3D U-Net-based [6] dose prediction, and Proposed dose prediction.

References

[1] J. Fan, J. Wang, Z. Chen, et al. "Automatic treatment planning based on three-dimensional dose distribution predicted from deep learning technique". *Medical physics* 46.1 (2019), pp. 370–381.

[2] J. Ma, D. Nguyen, T. Bai, et al. "A feasibility study on deep learning-based individualized 3D dose distribution prediction". *Medical physics* 48.8 (2021), pp. 4438–4447.

[3] M. Mashayekhi, I. R. Tapia, A. Balagopal, et al. "Site-agnostic 3D dose distribution prediction with deep learning neural networks". *Medical physics* 49.3 (2022), pp. 1391–1406.

[4] O. Ronneberger, P. Fischer, and T. Brox. "U-net: Convolutional networks for biomedical image segmentation". *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer. 2015, pp. 234–241.

[5] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, et al. "3D U-Net: learning dense volumetric segmentation from sparse annotation". *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19*. Springer. 2016, pp. 424–432.

[6] S. Liu, J. Zhang, T. Li, et al. "A cascade 3D U-Net for dose prediction in radiotherapy". *Medical physics* 48.9 (2021), pp. 5574–5582.

[7] J. Liu, Q. J. Wu, J. P. Kirkpatrick, et al. "From active shape model to active optical flow model: a shape-based approach to predicting voxel-level dose distributions in spine SBRT". *Physics in Medicine & Biology* 60.5 (2015), N83.

[8] F. Isensee, P. F. Jaeger, S. A. Kohl, et al. "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation". *Nature methods* 18.2 (2021), pp. 203–211.

Can input reconstruction be used to directly estimate uncertainty of a dose prediction U-Net model?

Margerie Huet Dastarac[1], Dan Nguyen[2], Steve Jiang[2], John Lee[1], and Ana Barragan Montero[1]

[1]Molecular Imaging, Radiation and Oncology (MIRO) Laboratory, Institut de Recherche Expérimentale et Clinique (IREC), UCLouvain, Belgium
[2] Medical Artificial Intelligence and Automation (MAIA) Laboratory, Department of Radiation Oncology, UT Southwestern Medical Center, USA

Abstract Artificial Intelligence is gaining momentum in fields like radiation therapy. However, there is a lack of reliable and easy to implement quality assurance tools. To address this problem, we present a direct uncertainty estimation method for dose prediction models using input reconstruction error. Our method achieves a higher correlation between uncertainty and prediction loss than the state of the art and allows an easier identification of out-of-distribution data.

1 Introduction

The reliable and efficient estimation of uncertainty in artificial intelligence (AI) models poses an ongoing challenge in many fields such as radiation therapy. AI models are intended to automate manual steps involved in the treatment planning workflow [1]. We focus in this study on models that predict an optimal dose trade-off for each new patient for a specific treatment modality. They can be used to guide physicians in the optimization [2], for automatic treatment plan generation [3] or for decision support [4]. The clinical implementation of AI models is relatively slow compared with other fields. It is primarily due to the lack of reliable and easy to implement quality assurance tools [5]. Indeed, in critical environments like radiotherapy, AI must be used with caution since large uncertainty in the predictions can jeopardize patients safety. Uncertainty estimation techniques must also consider the constraints of the clinic such as limited time and resources. The two prominent methods are deep ensembling (DE) [6] and Monte Carlo dropout (MCDO) [7]. DE uses the variance of multiple networks predictions while MCDO combines several forward passes with randomly dropped out weights. Both require high inference time and MCDO has been argued to be unreliable on real world datasets [8], particularly in out-of-distribution (OOD) detection.

Direct uncertainty estimation methods try to remove the requirement of multiple passes by proposing one shot uncertainty estimation. For instance, direct epistemic uncertainty prediction (DEUP) [9] disentangles epistemic and aleatoric uncertainties by subtracting an estimation of the latter to the prediction of the error. The aleatoric uncertainty is approximated with the variance of labels for the same input by oracles. Despite its elegance, applications like radiotherapy dose prediction would require the generation of multiple treatment plans by different experts for the same set of patients. It is currently very difficult and unrealistic to obtain such data due to the time constraints of clinical environments.

In this work, we present a direct uncertainty estimation method using input reconstruction error. We propose to reconstruct the CT scan given in input from the latent representation of a dose prediction model. A CT reconstruction branch stems from the bottleneck of the U-Net based dose prediction model (see Figure 1). The mean squared error (MSE) of the CT reconstruction can be used as a surrogate of the uncertainty. Our method achieves a higher correlation between uncertainty and prediction loss than the state-of-the-art and requires less computation time. We also investigate its performance as an OOD detection by testing with patients treated for other head and neck sites. Our method allows an easier identification of OOD data, shown by a higher Z-score and no overlap between the distributions.

2 Materials and Methods

2.1 Architecture:

Our model is based on the HDUNet [10], a convolutional neural network, with a hybrid architecture between the popular U-Net [11] and DenseNet [12]. Our model plugs a second decoder branch to the bottleneck after the encoder, which aims to reconstruct the CT scan given in input, like an auto-encoder (see Figure 1). The training loss is the sum of two MSE, one for dose prediction and the other for CT reconstruction.

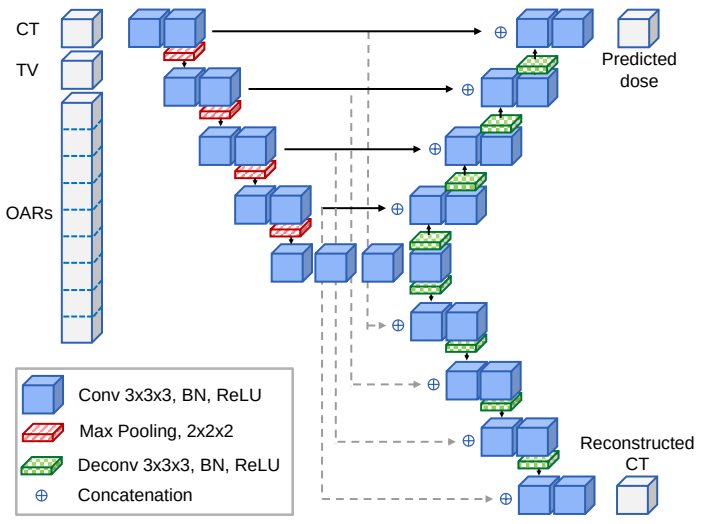


Figure 1: Architecture of the network.

2.2 Data:

Dataset 1 - ID

The training set of the DL dose prediction network is an anonymized database of 60 patients with oropharynx cancer. The treatment planning system used to generate the plans is RayStation 11B. For consistency regarding the planning strategy and the beam configuration, all selected patients have two dose prescription levels: 70Gy (CTV_HIGH) and 54.25Gy (CTV_LOW) on tumor and elective nodal volumes, respectively. PT plans are planned for pencil beam scanning delivery, using four beams. Each patient has a PT dose distribution, a CT image and a set of contours for the target volumes and the OARs.

Dataset 2 - OOD

The second database used to investigate its performance with OOD is a set of 10 patients obtained from the same institution. The cancer sites are also head and neck but not oropharynx: larynx, hypopharynx, nasopharynx, supraglottis and pyriform sinus. Each patient has a CT image and the same set of contours as the ID dataset.

2.3 Training:

Networks were trained to minimize the MSE loss functions from the two branches.In the ID analysis as described below, we used a 11-fold nested cross-validation approach. This method entails to set aside 3 patients as an outer test set and train multiple times the networks with moving training, validation, and test sets. The validation and test sets consist of 5 patients each, and we use the remaining 47 patients for the training. The process is repeated 11 times, allowing us to obtain predictions and uncertainty values on 55 patients (number of patients per test set times number of folds).

2.4 Uncertainty estimation:

The proposed surrogate of the uncertainty is the MSE between the clinical and reconstructed CT images. We consider the voxels in the body contour to keep relevant information in our metric and avoid a positive bias caused by zero values of the background around the patient. Moreover, couch setup can vary and should not be considered by our metric.

2.5 Evaluation:

We evaluated the performance of the input reconstruction method in three parts: its impact on the model performance, the accuracy of the uncertainty metric when fed ID data and its ability to perform OOD detection.

- **Impact of the additional branch on the model**: We compute the difference in dose-volume histogram (DVH) metrics between predictions of standard HDUNet and our model. A paired-samples Wilcoxon test was performed. The threshold for statistical significance is set at p-value < 0.05.

- **ID analysis**: The dose prediction model is trained and tested on the ID dataset following the training process described in the Training subsection. At inference time, we computed the Pearson correlation coefficient between the uncertainty estimation and the dose prediction error and compared it with MCDO and DE.

- **OOD analysis**: The dose prediction model is trained on the ID dataset and the inference is performed on the ID and OOD datasets. The two distributions are displayed as histograms for CT reconstruction, MCDO and DE. The Z-score indicates the distance between the distributions. We also computed the number of patients in the overlap between the ID and OOD histograms for the different methods.

  For MCDO, a Bernoulli law was used to drop the weights at inference time with probabilities of 0.1 to 0.5. The standard deviation is computed voxelwise on 20 inferences and the mean over the body is considered.

  For DE, we trained 20 models with the weights following the He initialization [13]. The standard deviation is computed voxelwise on the inference of the 20 models and we consider the mean over the body.

3 Results

The performance of the dose prediction model is shown in Figure 2. The median absolute error is close to zero in the D95 and D99 of the target volumes (Figure 2, bottom) and less than one percent of the highest prescribed dose for Dmean and D2 of OARs (Figure 2, top). The addition of the CT reconstruction branch does not impact in a statistically significant way the dose prediction of the network, as seen in Figure 2. The paired-samples Wilcoxon test over DVH metrics also showed that there is no statistical significance brought by adding the second decoder branch.

	Pearson Correlation	p-value
Deep Ensemble	0.447	6.2e-4
MC Dropout (0.1)	0.599	1.3e-6
MC Dropout (0.2)	0.606	9.1e-7
MC Dropout (0.3)	0.609	8.0e-7
MC Dropout (0.4)	0.613	6.4e-7
MC Dropout (0.5)	0.612	6.6e-7
CT reconstruction	0.620	4.5e-7

Table 1: Pearson correlation coefficients of mean squared error between the predicted and clinical dose in the body and mean squared error between predicted and clinical CT scan. The data was computed for in-distribution patients.

In Figure 3, we can see the distribution of the different uncertainty methods for the ID and the OOD databases. While the five MCDO methods fail to distinguish OOD from ID, DE almost separates between the two distributions with only one patient on the right-hand side of the histogram. Our input reconstruction technique on the other hand allows us to easily set a threshold between ID and OOD patient data. Table 2 shows distance metrics between ID and OOD for each uncertainty estimation method.

	Z-score	Intersection
Deep Ensemble	2.347	0
MC Dropout (0.1)	0.368	9
MC Dropout (0.2)	0.877	4
MC Dropout (0.3)	1.098	2
MC Dropout (0.4)	1.177	2
MC Dropout (0.5)	1.124	3
CT reconstruction	34.050	0

Table 2: Z-score and intersection (in number of patients) between in and out-of-distribution data for each method.

The uncertainty estimation of our method is obtained simultaneously to the main task which takes 10 seconds. MCDO and DE require 20 inference passes per patch.

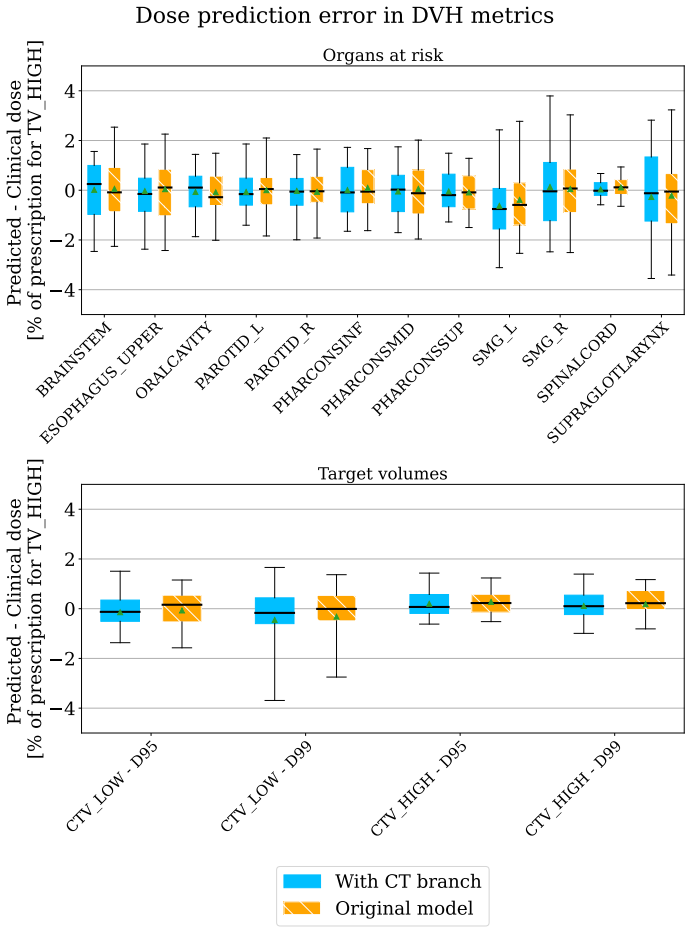


Figure 2: Performance in terms of clinical DVH metrics of the dose prediction model.

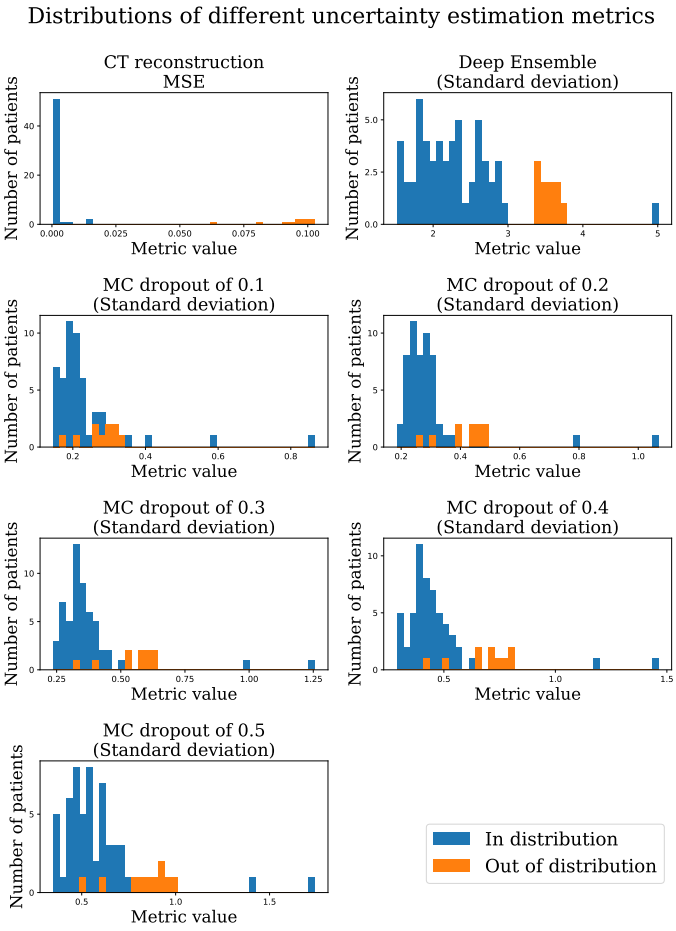


Figure 3: Distributions of the uncertainty estimation metrics

4 Discussion

In this study, we investigate a novel architecture to directly estimate the global uncertainty and detect OOD for a regression U-Net-like architecture by adding an input reconstruction branch. We applied it to an increasingly popular application in the medical field: the prediction of radiation therapy dose distribution. The surrogate of the uncertainty is the MSE of the CT reconstructed by the second decoding branch, grafted to the regular encoder-decoder architecture of a U-Net, as illustrated in Figure 1. The addition of the CT reconstruction branch and duplicate skip connections does not impact the prediction of the dose in a statistically significant way, as seen in Figure 2.

Table 1 shows that the CT reconstruction method presents the highest Pearson correlation coefficient compared with DE and MCDO with different dropout rates. Therefore, our method is a better indicator of patients susceptible to get a prediction with larger error.

In Figure 3, we notice that CT reconstruction is the only method allowing a clear separation with no overlap between the ID and OOD data. This observation is confirmed with the metrics in Table 2, where CT reconstruction has a Z-score of 34 whereas MCDO are less than 1.2 and about 2.2 for DE. Distribution shapes also differ. CT reconstruction distribu-

tion is positively skewed, DE is difficult to define and the five MCDO distributions are also positively skewed but with a wider spread. Two patients present much higher MCDO variances. They would be falsely considered as outliers if we used MCDO for OOD flagging. Despite having been used effectively for OOD detection in 2D image classification tasks in [14], MCDO does not appear to perform well for our regression task with 3D clinical data. DE also has one patient that would have been classified as an outlier. Interestingly, the patient with highest DE or MCDO variances is the same for all metrics. After a visual check by a physician, no striking difference with other oropharynx patients was detected. The patient with the lowest CT reconstruction MSE value from the OOD dataset is a patient with a centered tumor site of supraglottis whereas the furthest apart has a tumor in the lowest part of the larynx. This observation suggests a connection between the high uncertainty and the distance between ID and OOD tumor location.

The widespread adoption of MCDO and DE can be attributed, in part, to their ease of adaptation to various models. However, both techniques require on average 20-50 inference passes, which means that either you need more computational resources to parallelize inferences or more time. Our solution allows us to compute the uncertainty for free, since it is computed simultaneously to the main task. In a real-time scenario such as adaptive radiotherapy workflow, reducing the time spent by a patient on the couch is crucial.

Contrary to MCDO and DE, our method outputs a single value as uncertainty measure whereas the former techniques can produce 3D uncertainty maps. Reliable uncertainty maps are definitely an interesting tool to help physicians apprehend the regions of uncertainty and would be a great addition to the interface of the clinical workflow. There are still applications where a single value can be used and where our model could be beneficial. For example, to flag OOD patients, these patients could either be redirected to another network trained on closer patients or require the attention of a clinician to proceed with the treatment. Another application could be in the curation of networks with active learning. Active learning is the task of optimizing the selection of the training set among a pool of unlabeled data [15, 16]. Current active learning methods use MCDO or DE, rank unlabeled data according to their uncertainty and select highly uncertain samples for oracles to label [15]. We think that active learning would also benefit from our fast method.

One limitation of our study is that the addition of the CT reconstruction branch has only been investigated for a regression task. This implies that both outputs (predicted dose and CT reconstruction) are trained with the same MSE loss and therefore the correlation between the error and the uncertainty is more straightforward. It would be interesting to investigate the behavior of such an addition for a task where the main loss and the reconstruction loss are different. For example, also in the field of radiation therapy, a segmentation task uses Dice or cross-entropy as the main loss function.

5 Conclusion

In this study, we investigate a novel architecture to directly estimate the global uncertainty and detect out-of-distribution detection for a proton therapy dose prediction model by adding an input reconstruction branch. This method showed a higher Pearson correlation coefficient between the dose prediction error and uncertainty estimation metric than state-of-the-art techniques like Monte Carlo dropout (MCDO) and Deep ensemble (DE). Moreover, our method allows an easier identification of out-of-distribution data. Contrary to MCDO and DE, the uncertainty is computed simultaneously to the main task, which is ideal in time-sensitive applications. This technique could be investigated in the future for other U-Net like architecture like segmentation or detection.

References

[1] A. de Biase et al. “Standardization of Artificial Intelligence Development in Radiotherapy”. *Radiat. Oncol.* 32.4 (2022).

[2] M. Mashayekhi et al. “Artificial intelligence guided physician directive improves head and neck planning quality and practice Uniformity: A prospective study”. *CTRO* 40 (2023).

[3] D. Maes et al. “Automated treatment planning for proton pencil beam scanning using deep learning dose prediction and dose-mimicking optimization”. *JACMP* (2023).

[4] M. Huet-Dastarac et al. “Patient selection for proton therapy using Normal Tissue Complication Probability with deep learning dose prediction for oropharyngeal cancer”. *Med. Phys.* (2023).

[5] A. Barragán et al. “Towards a safe and efficient clinical implementation of machine learning in radiation oncology by exploring model interpretability, explainability and data-model dependency”. *PMB* 67.11 (2022).

[6] B. Lakshminarayanan et al. “Simple and scalable predictive uncertainty estimation using deep ensembles”. *NeurIPS* 30 (2017).

[7] Y. Gal et al. “Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning”. Vol. 48. PMLR, 2016.

[8] Y. Liu et al. “The Peril of Popular Deep Learning Uncertainty Estimation Methods” (2021).

[9] S. Lahlou et al. “DEUP: Direct Epistemic Uncertainty Prediction” (2021).

[10] D. Nguyen et al. “3D radiotherapy dose prediction on head and neck cancer patients with a hierarchically densely connected U-net deep learning architecture”. *PMB* 64.6 (2019).

[11] O. Ronneberger et al. “U-Net: Convolutional Networks for Biomedical Image Segmentation”. *MICCAI*. 2015.

[12] G. Huang et al. “Densely Connected Convolutional Networks”. *CVPR*. 2017.

[13] K. He et al. “Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification” (2015).

[14] A. T. Nguyen et al. “Out of Distribution Data Detection Using Dropout Bayesian Neural Networks”. *AAAI* 36.7 (2022).

[15] S. Budd et al. “A survey on active learning and human-in-the-loop deep learning for medical image analysis”. *MEDIA* 71 (2021).

[16] M. Abdar et al. “A review of uncertainty quantification in deep learning: Techniques, applications and challenges”. *IF* 76 (2021).

Investigation of Unsupervised Race-based Predictive Modeling Strategies to Improve Prediction Power of Tumor Recurrence in Head and Neck Cancer Patients in Diverse Racial Groups

Hassan Bagher-Ebadian[1,2,3,4], Marissa V. Gilbert[1], Farzan Siddiqui[1], Benjamin Movsas[1], Kundan Thind[1,5,6], Indrin J. Chetty[4,7]

[1]Department of Radiation Oncology, Henry Ford Cancer Institute, Detroit, USA, [2]Department of Radiology, Michigan State University, East Lansing, USA, [3]Department of Osteopathic Medicine, Michigan State University, East Lansing, USA, [4]Department of Physics, Oakland University, Rochester, USA, [5]Department of Medicine, Michigan State University, East Lansing, USA,[6]Department of Oncology, School of Medicine, Wayne State University, Detroit, USA, [7]Department of Radiation Oncology, Cedars-Sinai Medical Center, Los Angles, USA

Abstract: Purpose: This study investigates the predictive value of different race-based models constructed from radiomics and clinical factors, to predict tumor recurrence at three years for African American and Caucasian head and neck cancer patients. Two-hundred-two patients (78 African American, and 124 Caucasian), treated with definitive radiation/chemoradiation (70Gy, 35-Fraction) were studied. A five-fold nested-cross-validation (NCV) technique was used to evaluate forty-five unsupervised Kohonen-Self-Organizing Maps (KSOMs, topology: 5×5, five-folds for nine different models corresponding to different combined cohorts and information modalities) constructed from twenty-six one-hot encoded clinical factors and 447 radiomics features extracted from patients' gross-tumor-volume on planning-CT images. The NCV-based unsupervised predictive accuracy for African American and Caucasian-based models were: 0.712±0.048 and 0.633±0.066 for clinical factors, 0.736±0.069 and 0.678±0.074, for radiomics, and 0.756±0.088 and 0.693±0.093, for the ensemble model (combined information: radiomics + clinical-factors). The accuracy for the combined cohort ensemble models (combined clinical and Radiomics information) were: 0.798±0.093 and 0.734±0.084 for African American and Caucasian patients, respectively. This study demonstrates the value of different unsupervised modeling strategies for combining different information modalities, their sparsity levels and influence on the model sensitivity according to their class imbalance, and patient cohort sizes to reveal and improve the outcomes prediction power in diverse racial groups.

1 Introduction

Significant racial disparities currently exist in head and neck (H/N) cancer outcomes, with 5-year overall survival being ~30% lower in African American (AA) versus Caucasian (CC) patients [1-4]. After adjusting for age, sex, marital status, tumor site, and year of diagnosis, African American race has been found to be associated with worse OS and Disease-Specific Survival (DSS) Hazard Ratio=1.67, p<0.0001) [5]. Epidemiological studies have suggested several reasons for the disparity including lack of access of AA patients to high volume surgical centers, limited access to care, cost, and transportation limitations, etc. [1-4]. It is important to examine how imaging biomarkers and clinical factors can contribute to these outcome disparities. Image-based biomarkers in combination with relevant clinical factors can be used to develop predictive models, which can risk-stratify individual patients prior to cancer treatment to improve outcomes and ultimately reduce the disparities between these racial groups. This study investigates different unsupervised model development strategies and their predictive powers for inference of tumor recurrence after three years in H/N cancer patients in diverse racial groups.

2 Materials and Methods

2.1 Patient Population, Treatment, Outcome information, and Imaging: Two-hundred-two subjects were identified from a head and neck cancer database (REDCap 10.6; Nashville, TN) under an institutional review board (IRB: 8751-27) approved as a single institution retrospective study. Subjects diagnosed with head and neck cancer primary with an HPV status of unknown or negative were included for analysis. The recurrence Free Interval (RFI) information of the patients at three years was used as the outcome of interest for the model development. Seventy-eight African American patients (68 poor responders and 10 good responders) and 124 Caucasian patients (92 poor responders and 32 good responders) were studied. All patients were treated with definitive radiation/chemoradiation at our institution with fractionated radiation therapy (total dose was 70 Gy in 35 fractions) between February 2005 and December 2021. The patients were treated either with intensity modulated radiation therapy (IMRT) or volumetric modulated arctherapy (VMAT). Patient planning CT (pCT) images were acquired by a Brilliance Big Bore, 16-slice CT scanner (Philips Health Care, Best, the Netherlands). Reconstructed axial pCT images were 512×512 pixels with an in-plane resolution of 1.23×1.23 mm^2 and image thickness of 3mm.

2.2 Clinical Factors Information and Radiomics Analysis: We retrieved the following 6 clinical factors from our institutional head and neck database for the two cohorts (African American and Caucasian): 1- Age at Diagnosis, 2- Gender, 3- Total Charlson Comorbidity Index (CCI), 4- Alcohol Use, 5- Smoking History, and 6- Clinical T Stage. The clinical factors were selected based on their correspondence with the primary tumor since in the radiomic analysis, we only analyzed the primary tumor and did not investigate the lymph nodes. As shown in Table-1, for each patient, the 6 different clinical factors were encoded into 26 channels and mapped from categorical to quantitative/numeric information using One-Hot Encoding technique [6]. Quantitative image analysis of radiomics features was performed using RODiomiX [7, 8], a software system consists of a series of computation cores that has been developed by our group, and validated on the recommendations provided by the image biomarker standardization initiative (IBSI) [9, 10]. Data quantization, and image re-segmentation [9, 10] were performed prior to radiomic feature extraction as follows: Effects of variable

voxel sizes (SNR inconsistency) from pCT images were minimized by resampling the image voxels into a uniform and isotropic voxel size (1 mm^3/voxel) using proximal or Nearest-Neighbor interpolation technique [9, 10]. GTV intensities were then quantized and resampled into 64 intensity levels using a Fixed Bin Number technique [9, 10], as recommended by the IBSI. Volume re-segmentation was also performed to exclude any out of range voxels (due to resampling and FBN quantization) [9, 10].

Table-1: This table shows the clinical factors and their respective One-Hot encoding labels.

Ch.	Encoded Clinical Factor	Ch.	Encoded Clinical Factor
1	Race - (0: AA, 1: Caucasian)	14	CCI-0
2	Alcohol Use- None	15	CCI-1
3	Alcohol Use- Social Drinker	16	CCI-2
4	Alcohol Use- Occasional (1-10/wk)	17	CCI-3
5	Alcohol Use - Frequent (>10/wk)	18	CCI-4
6	Alcohol Use -Current drinker	19	CCI-5
7	Alcohol Use - Alcoholism- in the past, not currently drinking	20	CCI-6
8	Alcohol Use – Unknown	21	CCI-7
9	Smoking History - Never Smoker	22	CCI-8
10	Smoking History - Former use (Quit >3 months ago)	23	HPV Status – Negative
11	Smoking History -Current use	24	HPV Status - Not tested/Not done
12	Smoking History - Unknown/No Data	25	HPV Status – NA
13	Gender (0: Female, 1: Male)	26	HPV Status - Unknown/No data

Radiomic feature extraction was performed using the ROdiomiX software system [7, 8]. A series of 2D, 2.5D, and 3D IBSI-validated radiomic features (447 features, Chebyshev distance=1.0, symmetrical, 26-connected Neighborhoods, with 13 and 8 unique directions,) from 11 feature categories (LI, IBS, IH, IVH, MORPH, GLCM, GLRLM, GLSZM, GLDZM, NGLDM, and NGTDM) [9, 10] with 10 different IBSI recommended aggregation techniques (DHQ4, BTW3, SUJT, JJUI, ZW7Z, ITBB, IAZD, 8QNN, 62GR, and KOBO) were extracted from pCT images of the GTVs [9, 10].

2.3 Data Augmentation and Unsupervised Model Development: The radiomics and encoded clinical features were augmented (~ 10 times) using Synthetic Minority Over-sampling TEchnique (SOMTE [11]) to address the data imbalance between the number of patients within the two cohorts (AA and CC) and their respective poor and good responder sub cohorts. An unsupervised Kohonen self-organizing map (K-SOM)[12] or self-organizing feature map analysis was used to reveal the intrinsic association of radiomics and clinical factors and radiomics information with the RFI at three years. A five-fold nested-cross-validation (NCV, [13]) technique was used to evaluate forty-five unsupervised Kohonen-Self-Organizing Maps (KSOMs, topology: 5×5, five-folds for nine different models corresponds to different combined cohorts and information modalities) constructed from twenty-six one-hot encoded clinical factors and 447 radiomics features. The KSOM is an unsupervised machine learning technique used to produce a low-dimensional representation of higher dimensional data set with complex structures while preserving the topological structure of the data. During the K-SOM analysis, a competitive learning algorithm along with Best Matching Unit (BMU) strategy were employed to identify the winner nodes/neurons for a 5×5 topology. The cover steps for initial mapping of the input space for ordering the phase steps of the K-SOM was set to 100. The K-SOM's architecture was hexagonal with an initial neighborhood size of three, and the maximum epoch was set to 200 epochs for batch training mode. For each fold, a BMU-based hit map and its corresponding probability and fuzzed maps [14] were generated for the test and training cohorts using different information modalities and combined cohorts (Figure 1). Ultimately, the NCV-based classification accuracy for poor and good responders for different KSOM-based test models were estimated (at different probability thresholds) and averaged over five different folds.

3 Results

Subfigures 1A-J (left column) demonstrate the AA- and CC-based KSOM fused probability maps for outcome prediction of AA and CC patients using Clinical Factors (1A-B), Radiomics information (1E-F), and combination of Clinical Factors and Radiomics information (1I-J). Subfigures 1C-L (right column) demonstrate the (AA+CC)-based KSOM fused probability maps for outcome prediction of AA and CC patients using Clinical Factors (1C-D), Radiomics information (1G-H), and combination of Clinical Factors and Radiomics information (1K-L). The green, and red zones on the K-SOM topology maps (B–F and H–L) correspond to the good and poor responders predicted by the KSOM models. Subfigures A-F demonstrate 2D and 3D contribution level of diverse racial groups (green for AA and red for CC) into the information content of the KSOM's feature space for Clinical Factors (2A-B), Radiomics(2C-D), and Clinical Factors + Radiomics (2E-F). For a better visualization, the 2D and 3D surface maps are up sampled by 4 and 8, respectively. The NCV-based unsupervised predictive performances for African American and Caucasian-based models were: 0.712±0.048 and 0.633±0.066 for clinical factors, 0.736±0.069 and 0.678±0.074, for radiomics, and 0.756±0.088 and 0.693±0.093, for ensemble model (radiomics + clinical-factors). The accuracy for the combined cohort (AA+CC) ensemble models (combined clinical and Radiomics information) were: 0.798±0.093 and 0.734±0.084 for African American and Caucasian patients, respectively.

4 Discussion

The KSOM technique is an unsupervised learning method which enables detection and preservation of the topological relationship of the training features based on the information similarity. This implies that the underlying

properties of the feature space are uncovered to discover associations that are not easily identified from the raw information. We used a competitive learning algorithm along with Best Matching Unit strategy to identify the winner nodes for a 5×5 hexagonal topology.

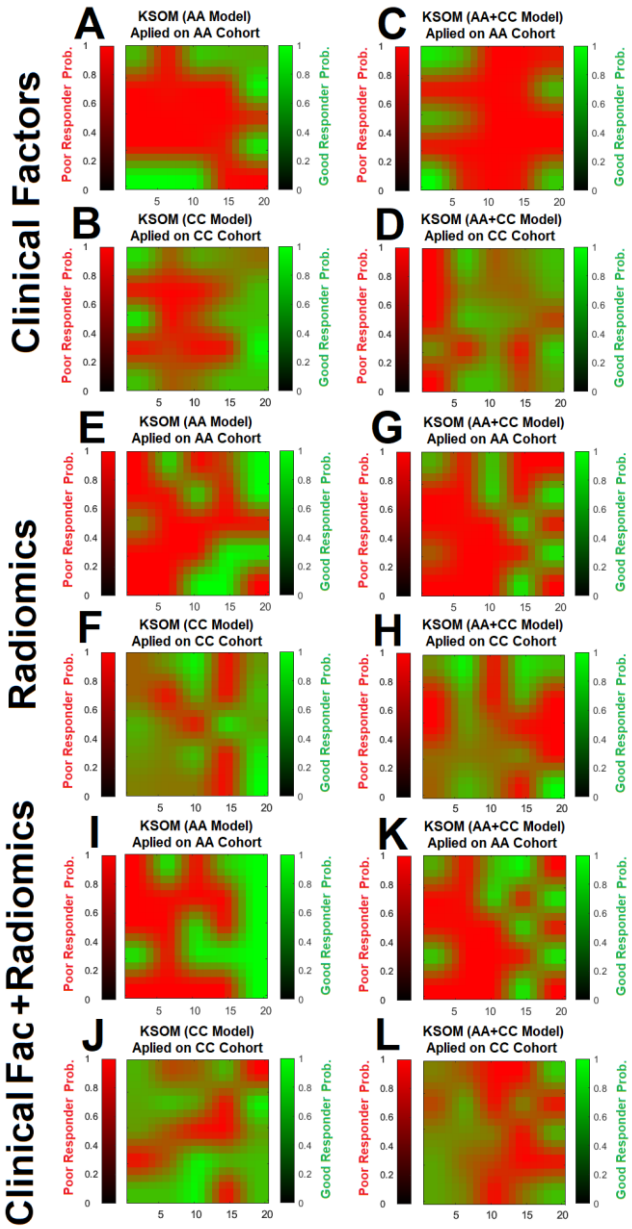


Figure-1: Subfigures 1A-J (left column) demonstrate the AA- and CC-based KSOM prediction probability maps for AA and CC patients for Clinical Factors (1A-B), Radiomics information (1E-F), and their combination (1I-J). Subfigures 1C-L (right column) demonstrate the (AA+CC)-based KSOM prediction probability maps for AA and CC patients for Clinical Factors (1C-D), Radiomics information (1G-H), and their combination (1K-L). The green, and red zones on the K-SOM maps correspond to the good and poor responders predicted by the KSOM models.

However, recruitment of different topology sizes can affect the quality of the KSOM algorithm in detecting the association and distinction of the information clusters on the KSOM's feature space. We also used a very conservative method of cross-validation technique, NCV, to prevent any potential information leakage between different training and test cohorts across different folds. The datasets used in this study were highly imbalanced (~ 7 times and ~2.9 times for AA and CC, respectively) with significant impact on the performance of the developed models. Although, the SMOTE technique was used to circumvent the class (poor vs. good responders) and cohort (AA vs. CC) imbalance issues, it can generate samples between actual samples that represent noise. thus, the augmented dataset can contain more noise than the original one, which can negatively impact on the performance of the model.

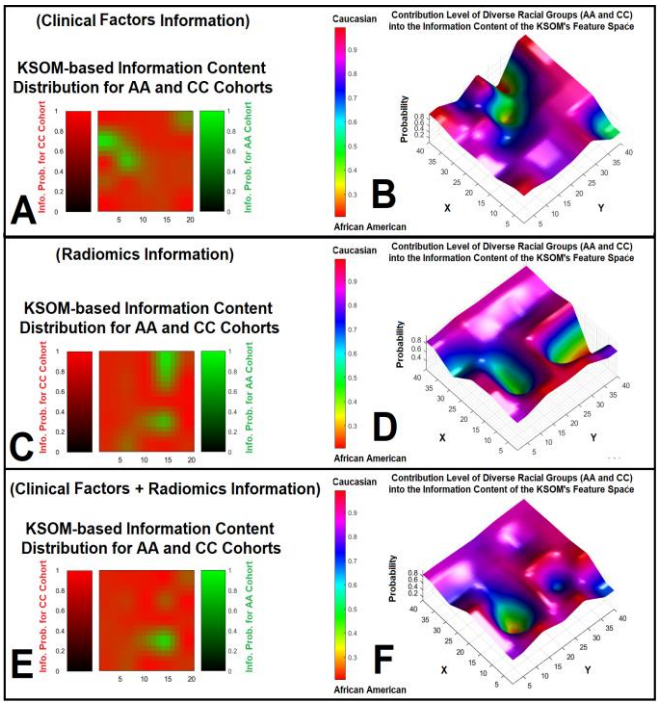


Figure-2: Subfigures A-F demonstrate 2D and 3D contribution level of diverse racial groups (green for AA and red for CC) into the information content of the KSOM's feature space for Clinical Factors (2A-B), Radiomics(2C-D), and Clinical Factors + Radiomics (2E-F). For a better visualization, the 2D and 3D surface maps are up sampled by 4 and 8, respectively.

As shown in subfigures 2A-F, compared to the CC cohorts, the information content for different information modalities (Clinical Factors, Radiomics, and Clinical + Radiomics) provided by the AA cohorts for all the KSOM models are insignificant. Thus, there is a domination effect from the CC patients' information content on the feature space with more information variabilities (see distribution of the information on 2A-F). The NCV analysis results imply that compared to the CC cohort, the predictive power of the Clinical information is higher (~8%) for the AA cohort and AA would benefit more from a separate model constructed from clinical factors. Also, when the clinical information of the two cohorts (AA+CC) is used to construct the KSOM, the predictive powers increase for both racial groups (~4% and 8% for AA and CC, respectively) while the uncertainty of the predictions for both races increase.

As shown in figures 1 and 2, the Radiomics information somehow outperforms the clinical information for both races, while it is more beneficial to the CC patients (2% vs. 4% for AA and CC, respectively). When a combined model

(AA+CC) is constructed from the Radiomics information, the prediction powers for both races increase (~2-3%). Thus, the AA patients benefit more from a combined model (AA+CC) constructed from Radiomics information. When the two information modalities are combined (Radiomics + Clinical Factors), the predictive power for both races are slightly improved (~2-3%). As expected, results imply that when the two information modalities (Radiomics + Clinical information) and cohorts (AA+CC) are used, the predictive power increases for both races (~3-4% for AA and CC, respectively).

In this study, distribution of different information modalities for the two racial groups were mapped on the feature space using an unsupervised method to reveal their cohort-based and information modality-based similarities and dissimilarities. Such analysis can reveal the importance and contribution level of different information modalities in construction of predictive models. Different modeling strategies for combining patient cohorts and their respective information modalities and their predictive powers were investigated. It was shown that the way these information modalities and patient cohorts are used in model development, can highly affect the performance and uncertainty of the prediction. The results of this study can be significantly influenced by the sample size, the cohort and class imbalance, and the data augmentation. In this study, no feature engineering analysis (feature selection, multicollinearity elimination, etc.) was performed prior to the KSOM analysis. Incorporation of these analyses as well as using different supervised models could also improve the predictive power of the models. In this study, the three-year RFI information was used as the outcome information. Investigation of the RFI information at different timepoints to discover the trend of its intrinsic frequencies can provide more insight into the available information content, our strengths, and limitations in the model development.

Although radiomics can serve as a useful tool for distinguishing differences in outcome between AA and CC patients, it is important to note that this study does not offer causal associations. Causality must include consideration of health care disparities such as access to healthcare, costs of care, and other factors that might limit ability for AA patients to receive equivalent quality health care for the same cancers as CC patients.

5 Conclusion

This pilot study demonstrates the value of careful modeling considerations for combining imbalanced information modalities and patient cohorts to improve the outcomes prediction in diverse racial groups.

Acknowledgement: This work was supported in part by a grant from Varian Medical Systems (Siemens Healthineers, Palo Alto, CA), and an Integration Grant (#F80020) from Henry Ford Health and Michigan State University.

References

[1] V. Chiruvella *et al.*, "Analysis of Race and Gender Disparities in Mortality Trends from Patients Diagnosed with Nasopharyngeal, Oropharyngeal and Hypopharyngeal Cancer from 2000 to 2017," (in eng), *Int J Gen Med*, vol. 14, pp. 6315-6323, 2021, doi: 10.2147/ijgm.S301837.

[2] A. Kompelli *et al.*, "An assessment of racial differences in epidemiological, clinical and psychosocial factors among head and neck cancer patients at the time of surgery," *World Journal of Otorhinolaryngology - Head and Neck Surgery*, vol. 6, no. 1, pp. 41-48, 2020/03/01/ 2020, doi: https://doi.org/10.1016/j.wjorl.2019.01.002.

[3] U. C. Megwalu *et al.*, "Racial disparities in oropharyngeal cancer survival," (in eng), *Oral Oncol*, vol. 65, pp. 33-37, Feb 2017, doi: 10.1016/j.oraloncology.2016.12.015.

[4] N. Osazuwa-Peters, *et al.*, "Race and sex disparities in long-term survival of oral and oropharyngeal cancer in the United States," *Journal of Cancer Research and Clinical Oncology*, vol. 142, no. 2, pp. 521-528, 2016/02/01 2016, doi: 10.1007/s00432-015-2061-8.

[5] U. C. Megwalu *et al.*, "Racial Disparities in Oropharyngeal Cancer Stage at Diagnosis," (in eng), *Anticancer Res*, vol. 37, no. 2, pp. 835-839, Feb 2017, doi: 10.21873/anticanres.11386.

[6] D. Harris *et al.*, *Digital design and computer architecture*, 2nd ed. (no. 1). San Francisco, Calif.: Morgan 2012, p. 129.

[7] H. Bagher-Ebadian *et al.*, "Technical Note: ROdiomiX: A validated software for radiomics analysis of medical images in radiation oncology," *Med Phys*, vol. 48, no. 1, pp. 354-365, Nov 9 2020, doi: 10.1002/mp.14590.

[8] H. Bagher-Ebadian *et al.*, "ROdiomX: A New Validated Software for the Radiomics Analysis of Medical Images in Radiation Oncology," *Annual Jiont AAPM (American Association of Physicists in Medicine), Vanccouver, Canada, July 12-16 2020*, vol. 1, PO-GeP-M-371, July-2020.

[9] M. V. Hatt *et al.*, "Image Biomarker Standardisation Initiative (IBSI): An international community radiomics standardization initiative," *J Nucl Med* vol. 59 no. supplement 1 287 no. 1, p. 287, 2019 2019.

[10] Z. A. L. S. V. e. M. L. o. S, "Image biomarker standardisation initiative (IBSI)," *Creative Commons Attribution 4.0 International License (CC-BY)- arXiv preprint arXiv:1612.07003*, **2019**. [Online].

[11] V. N. Chawla *et al* , "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, no. 1, pp. 321-357, 2002.

[12] S. P. Lim *et al.*, "Cube Kohonen self-organizing map (CKSOM) model with new equations in organizing unstructured data," *IEEE Trans Neural Netw Learn Syst*, vol. 24, no. 9, pp. 1414-24, Sep 2013, doi: 10.1109/TNNLS.2013.2259259.

[13] GC T. Cawley, et al.,"On over-fitting in model selection and subsequent selection bias in performance evaluation," *Journal of Machine Learning Research*, vol.11, no.1, pp: 2079-2107, 2010.

[14] H. Bagher-Ebadian *et al.*, "Radiomics characterization of tissues in an animal brain tumor model imaged using dynamic contrast enhanced (DCE) MRI," *Sci Rep*, vol. 13, no. 1, p. 10693, Jul 2 2023, doi: 10.1038/s41598-023-37723-8.

Robust Models for Esopohageal Toxicity Prediction from Radiation Dose Maps

Matthew Gil[1,2], William H Nailon[2,3], Stephen Harrow[2], Paul Murray[1], and Stephen Marshall[1]

[1]Department of Electronic and Electrical Engineering, University of Strathclyde, Glasgow, United Kingdom
[2]Edinburgh Cancer Centre, NHS Lothian, Edinburgh, Scotland
[3]School of Engineering, University of Edinburgh, Edinburgh, Scotland

Abstract
Esophageal toxicity is a common group of adverse events caused by radiotherapy treatment in the chest region. Previous work has shown that dose based prediction models are possible but there has been a lack of testing of the robustness of these models. We use patient dose information and outcomes (N=397) from the RTOG-0617 clinical trial to create esophageal toxicity prediction models using an ANN and LSBoost achieving an AUC of 0.707 and 0.691 for the prediction of grade ≥ 3 and grade ≥ 2 esophageal toxicity respectively. A MAE of 0.884 was achieved when predicting exact toxicity grades. The robustness of the models was evaluated by applying random noise to the test sets and observing the degradation of performance. Methods to improve the robustness including L2 regularisation, SMOTE, random noise-based data augmentation and ensemble model approaches were shown to be effective for different models.

1 Introduction

Esophageal toxicity often occurs after the treatment of lung cancer, or other cancers, by external beam radiotherapy (RT) with esophagitis being the most common esophageal toxicity. Many studies have been produced aiming to predict and limit the occurrence and severity of esophageal toxicity with older studies generally aiming to establishing radiation dose volume effects [1] and more recent studies generally aiming to build machine-learning based models to predict esophageal toxicity [2]. These recent studies generally report the AUC ROC for grade ≥ 2 or ≥ 3 toxicity which usually falls in the range of 0.6 to 0.75 when only using dose features [3].
This study applies two popular models, artificial neural networks (ANNs) and boosted decision trees, for the prediction of esophageal toxicity from dose and clinical information using data available from the RTOG-0617 clinical trial [4]. The robustness of these models was evaluated by applying random noise to the test data with incrementally increasing magnitude. Data augmentation by SMOTE and the application of random noise during training were used to increase model robustness as well as increasing the L2 regularisation of the ANN and combining the models as an ensemble to produce final recommendations for building robust toxicity prediction models.

2 Materials and Methods

Unless mentioned otherwise, all computational methods were applied using MATLAB R2022b v9.13.0 (The Mathworks Inc).

2.1 Dataset

This study uses data from the RTOG-0617 clinical trial [4], a multi-centre trial investigating the effects of standard (60 Gy) versus high (74 Gy) doses of radiation to treat lung cancer patients as well as the effects of using cetuximab during treatment. This is, to the authors knowledge, currently the largest public RT dataset with the necessary information for our analysis. Patients in the RTOG-0617 trial received either IMRT or 3DCRT from one of 185 different institutions in the USA and Canada. Of the 544 patients initially recruited to the RTOG-0617 trial, we excluded 147 patients from our study due to non-completion during the RTOG-0617 trial (N=49), an overall survival of less than 6 months post RT (N=57) or due to issues with the radiotherapy data (N=41) leaving N=397 patients for our study. The data from the RTOG-0617 study is available and was accessed through The Cancer Imaging Archive (TCIA) [5], [6].

Toxicity Grade	0	1	2	3	4	5
Number of Patients	84	99	148	62	2	2

Table 1: The number of patients that developed a maximum toxicity for each toxicity grade (as defined by CTCAE v3 [7]).

The following esophageal toxicities, as defined by CTCAE v3 [7] (the version used in the RTOG-0617 study), were considered; esophagitis, acquired tracheo-esophageal fistula, dysphagia, dyspepsia, esophageal ulcer, esophageal stenosis and esophageal perforation. The number of patients that developed an esophageal toxicity for each CTCAE grade is displayed in table 1.

2.2 Dose and Clinical Features

Dose-volume histogram features were calculated from the RT planning dose maps isolated to the esophagus volume that was available from the RT planning contours. The dose features calculated were the V_x (the percent of the total volume receiving over x Grays), the Vol_x (the total volume receiving over x Grays), the mean esophageal dose and the max esophageal dose. The V_x and Vol_x were calculated for the range 5-80 Gy in steps of 5 Gy. Available clinical features were also included, these features were; age, gender, Zubrod performance, carcinoma type, RT technique (IMRT or 3DCRT), smoking status and the primary tumour volume.

All features were converted to z-scores (mean = 0, std dev =1) before model training.

2.3 Model Training and Hyperparameter Tuning

The models investigated were an ANN and the LSBoost algorithm [8] for boosted decision trees, both using the features from section 2.2 as the model input. The ANN consists of an input layer, fully connected layer, activation layer, fully connected layer and an output layer connected in order. The models were trained as regression models where the output is a continuous variable that correlates to the expected toxicity grade. After training, both model outputs were corrected for variance bias by setting the mean and variance of the model predictions on the training set equal to the mean and variance of the training set ground truth by applying the following equation:

$$\hat{y}_C = (\hat{y} - mean(\hat{Y}_{train})) * \sqrt{\frac{var(Y_{train})}{var(\hat{Y}_{train})}} + mean(Y_{train})$$

Where Y_{train} and $\hat{Y}_{train}$ are the set of all ground truth and model predictions for the training set respectively, $\hat{y}$ is a new prediction and $\hat{y}_C$ is the corrected new prediction. The models were evaluated as an exact grade prediction model and as a binary classifier by defining a cutoff value, 1.5 for grade ≥ 2 and 2.5 for grade ≥ 3 classification. The models were evaluated using the AUC for both grade ≥ 2 and grade ≥ 3 binary classification and the mean average error (MAE) for the exact grade output. Additionally, the ANN and LSBoost models were combined as an ensemble model by taking a weighted average of the two model outputs where the weighting was a tunable parameter.

Test Data Split A 4-fold data split was taken where one fold was held back for testing and the other three folds were used for training and hyperparameter tuning. The training and hyperparameter tuning was repeated four times, leaving out a separate fold for testing each time so that testing could be repeated for the full dataset.

Training and Hyperparameter Tuning Data Split The data in the three folds left out for training and hyperparameter tuning were again split into four folds. One of these folds was used for hyperparameter tuning and the others were used for training. The loss function used during training for both models was the mean square error (MSE). Cross-validation was performed so the hyperparameter tuning was repeated on each fold and the average hyperparameters were used to train the final model on all of the folds (not including the test data). Hyperparameter tuning was applied by Bayesian optimisation to minimise the MSE on the validation set. The hyperparameters tuned and their average values are given in section 3.1.

2.4 Robustness Tests

To test the robustness of the models, random noise was incrementally added to the features of the test set and the degradation of model performance in terms of the AUC for grade ≥ 3 toxicity and the MAE was recorded. For all features in the test set, random noise, X, was added to feature F converting it to feature F_N by $F_N = F + X$. Here $X \sim U(-\alpha, \alpha)$ where $U(-\alpha, \alpha)$ is a uniform distribution with maximum and minimum values α and $-\alpha$. As the features had been converted to z-scores, a value of $\alpha = 1$ means that the noise has a range of 1 standard deviation in the positive and negative directions. The noise was added incrementally with α increasing in steps of 0.1 from 0.0 to 5.0. For all robustness tests, no hyperparameter tuning was performed and the average hyperparameter values given in section 3.1 were used.

2.4.1 Improving Robustness

Techniques were employed to improve the robustness of the ANN and LSBoost models. Again the AUC for grade ≥ 3 classification and the MAE for the exact grade prediction were calculated.

L2 Regularisation: L2 regularisation is known to increase model robustness by forcing a model to rely on features more evenly, reducing overfitting. The robustness test was repeated for the ANN with different values of λ, the L2 regularisation term.

SMOTE: Synthetic minority oversampling technique (SMOTE) [9] is commonly applied in radiotherapy prediction models [2] to improve model performance by generating synthetic training data. SMOTE was applied to the training data to oversample the underrepresented grades. Grade 2 was the most common grade for esophageal toxicity so SMOTE was applied to oversample grade 0, 1 and 3 cases to be as represented as grade 2 cases. SMOTE was not applied to the grade 4 and 5 cases as there were only two of each case in the full dataset so it is not possible to oversample these cases. The performance of the ANN and LSBoost model was recorded as noise was applied to the model after training with SMOTE.

Adding Noise During Training: The training dataset was expanded by adding random noise to the training data features to generate new training examples. This was applied using the equation $F_N = F + X$, described in section 2.4, with an α value of 0.2. The training dataset was expanded to 5 times its original size using this technique. This was done for both the ANN and LSBoost models both on its own and after the application of SMOTE.

Ensemble Model: Using a weighted ensemble, as detailed in section 2.3, was additionally investigated as a method for increasing robustness to noise.

3 Results

3.1 Performance Metrics and Hyperparameter Tuning

The results of the model training with hyperparameter tuning are displayed in Figure 2. The hyperparameter values most

Model	AUC (Grade $\geq$ 3)	AUC (Grade $\geq$ 2)	MAE
ANN	**0.709**	0.690	0.894
LSBoost	0.682	0.669	0.925
Ensemble	0.707	**0.691**	**0.884**

Table 2: Results for the hyperparameter tuning of the ANN and LSBoost models.

commonly selected by the Bayesian hyperparameter tuning for both the ANN and LSBoost models are given below. **ANN Hyperparameters:** L2 regularisation $\lambda = 0.1$, fully connected layer size = 20, activation function = sigmoid, loss gradient tolerance = $1e^{-4}$. **LSBoost Hyperparameters:** Maximum number of splits = 2, minimum leaf size = 2, minimum parent size = 5, number of learning cycles = 400, learning rate = 0.01. **Ensemble Weighting:** The weighted average ensemble weighted the ANN 4 times higher than the LSBoost model on average. Further details regarding the hyperparameters can be found in the MATLAB R2022b files for the functions *'fitensemble()'* and *'fitrnet()'* for the LSBoost and ANN models respectively.

3.2 Robustness Tests

The results of the robustness tests are displayed in figures 1 to 5. Figures 2 and 1 show that increasing L2 regularisation increases robustness in terms of both MAE and AUC for the ANN model. Figure 3 shows that using SMOTE or added noise reduces the performance of the ANN in terms of the MAE. Figure 4 shows an improvement to the robustness of the LSBoost model in terms of the MAE. The best-performing ANN and LSBoost models were selected as the final models. No observable change was observed in the robustness of the ANN or LSBoost model in terms of the AUC when applying SMOTE or noise during training so the figures are not included here. The final ANN model uses a high L2 regularisation term of $\lambda = 0.95$ and only the original data during training. The final LSBoost model uses both SMOTE and added noise during training. The AUC and MAE responses of these final models as well as their weighted ensemble are displayed in figures 5 and 6 respectively.

4 Discussion

We have shown that a model for the prediction of esophageal toxicity from dose maps can be produced to achieve an AUC of 0.707 and 0.691 for the classification of grade ≥ 3 and grade ≥ 2 respectively with a MAE of 0.884 when trying to predict exact toxicity grade. The ANN beats the LSBoost model on all metrics when no test set noise is added. The ANN also outperforms the LSBoost model for all robustness tests provided there is a high level of L2 regularisation present. However, using a weighted ensemble of the two models provides benefit from the LSBoost model by increasing the model robustness to noise in terms of the MAE while not impacting the AUC. To improve the robustness of the models it was observed that a high level of L2 regularisation was beneficial for the ANN and, for the LSBoost model, using SMOTE and adding noise during training was beneficial. L2 regularisation was not available for the LSBoost model but may also be beneficial if implemented.

Many of the improvements to the robustness of the models would not be apparent if the tests to add noise were not applied. For example, in Figure 1, where $\alpha = 0$ and no noise is added, all of the models perform equally. A sharp divergence is then observed for lower values of λ with added noise. This highlights the benefit of these robustness tests for reducing overfitting and building robust models.

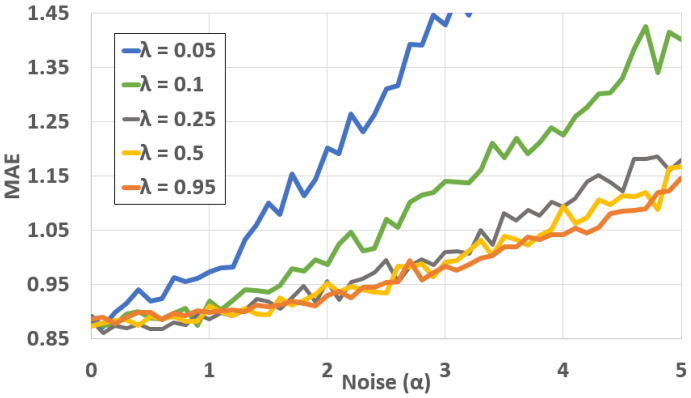


Figure 1: MAE plots for the ANN L2 regularisation tests.

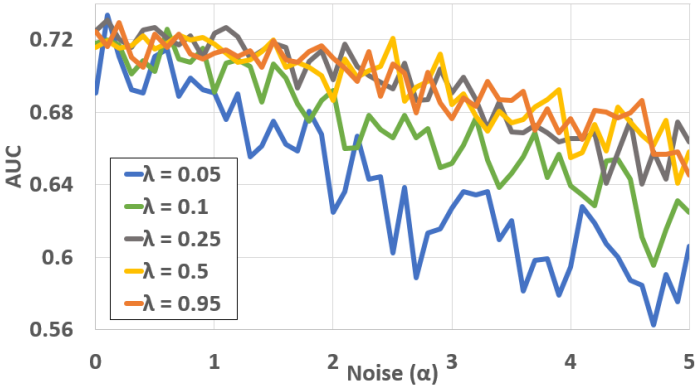


Figure 2: AUC plots for the ANN L2 regularisation tests.

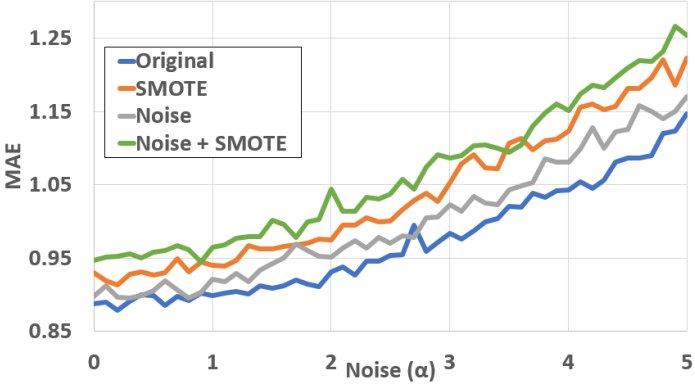


Figure 3: MAE plots for the ANN with SMOTE and noise added.

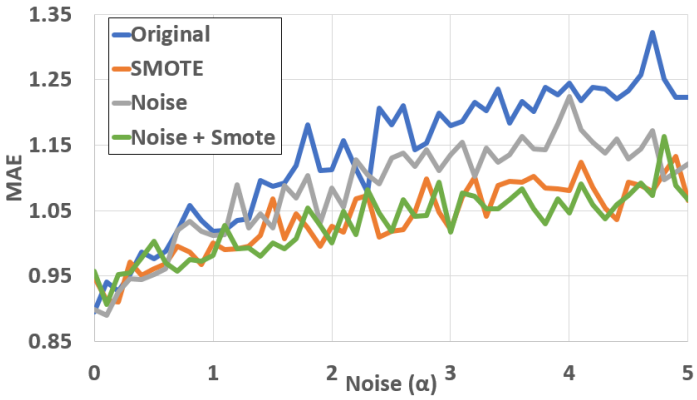


Figure 4: MAE for the LSBoost tests with SMOTE and noise.

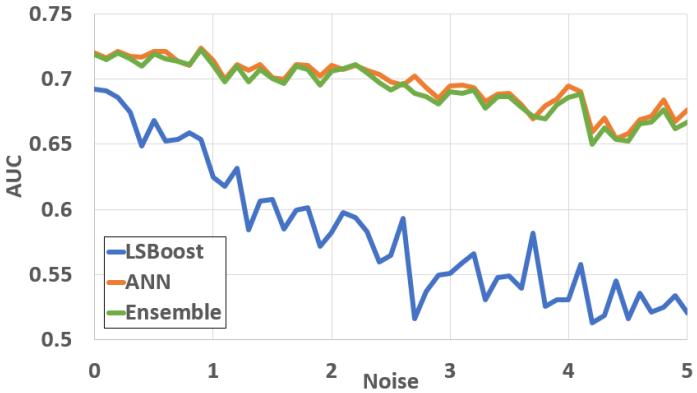


Figure 5: Final LSBoost, ANN and ensemble AUC plots.

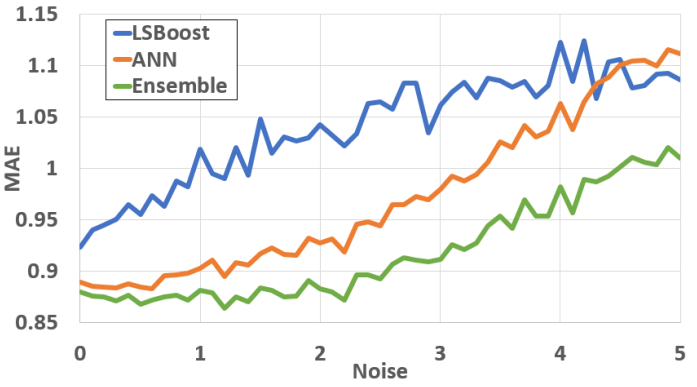


Figure 6: Final LSBoost, ANN and ensemble MAE plots.

This work may be expanded in the future to investigate the application of systematic errors to the test data to study robustness by altering the esophagus volume to simulate contouring errors or by randomly translating the esophagus volume to simulate positioning errors. Additional data augmentation techniques, such as the use of GANs to generate data, and their effect on robustness may also be an interesting avenue of further research.

5 Conclusion

ANN and LSBoost models were trained to predict esophageal toxicity from dose maps using the RTOG-0617 dataset. A weighted ensemble model formed of these two models achieved an AUC of 0.707 and 0.691 for classification of grade ≥ 3 and grade ≥ 2 toxicity respectively and a MAE of 0.884 when predicting exact toxicity grades. Model robustness was evaluated by adding random noise to the test data. This showed that for the ANN model, it is beneficial to include a high level of L2 regularisation and for the LSBoost model it is beneficial to use SMOTE and the application of random noise to augment the training dataset. Finally, it was shown that a weighted ensemble of the LSBoost and ANN approaches can further improve the model robustness.

Acknowledgements

This manuscript was prepared using data from Datasets (RTOG-0617; NCT00533949-D1, D2, D3) from the NCTN/NCORP Data Archive of the National Cancer Institute's (NCI's) National Clinical Trials Network (NCTN). Data was originally collected from clinical trial NCT number NCT00533949, titled "A Randomized Phase III Comparison of Standard-Dose (60 Gy) Versus High-Dose (74 Gy) Conformal Radiotherapy With Concurrent and Consolidation Carboplatin/Paclitaxel +/- Cetuximab (IND #103444) in Patients With Stage IIIA/IIIB Non-Small Cell Lung Cancer". All analyses and conclusions in this manuscript are the sole responsibility of the authors and do not necessarily reflect the opinions or views of the clinical trial investigators, the NCTN, or the NCI.
This work was funded by the Engineering & Physical Sciences Research Council (EPSRC) (grant reference: EP/S022821/1) and the Beatson Cancer Charity (grant reference: 19-20-043).

References

[1] M. Werner-Wasik, E. Yorke, J. Deasy, et al. "Radiation dose-volume effects in the esophagus". *International Journal of Radiation Oncology* Biology* Physics* 76.3 (2010), S86–S93.

[2] L. J. Isaksson, M. Pepa, M. Zaffaroni, et al. "Machine learning-based models for prediction of toxicity outcomes in radiotherapy". *Frontiers in Oncology* 10 (2020), p. 790.

[3] X. Zheng, W. Guo, Y. Wang, et al. "Multi-omics to predict acute radiation esophagitis in patients with lung cancer treated with intensity-modulated radiation therapy". *European Journal of Medical Research* 28.1 (2023), pp. 1–10.

[4] J. D. Bradley, R. Paulus, R. Komaki, et al. "Standard-dose versus high-dose conformal radiotherapy with concurrent and consolidation carboplatin plus paclitaxel with or without cetuximab for patients with stage IIIA or IIIB non-small-cell lung cancer (RTOG 0617): a randomised, two-by-two factorial phase 3 study". *The lancet oncology* 16.2 (2015), pp. 187–199.

[5] K. Clark, B. Vendt, K. Smith, et al. "The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository". *Journal of digital imaging* 26 (2013), pp. 1045–1057. DOI: 10.1007/s10278-013-9622-7.

[6] J. Bradley and K. Forster. "Data from NSCLC-Cetuximab". *The Cancer Imaging Archive* (2018). DOI: 10.7937/TCIA.2018.jze75u7v.

[7] A. Trotti, A. D. Colevas, A. Setser, et al. "CTCAE v3. 0: development of a comprehensive grading system for the adverse effects of cancer treatment". *Seminars in radiation oncology*. Vol. 13. 3. Elsevier. 2003, pp. 176–181.

[8] T. Hastie, R. Tibshirani, J. H. Friedman, et al. *The elements of statistical learning: data mining, inference, and prediction*. Vol. 2. Springer, 2009.

[9] N. V. Chawla, K. W. Bowyer, L. O. Hall, et al. "SMOTE: synthetic minority over-sampling technique". *Journal of artificial intelligence research* 16 (2002), pp. 321–357.

Improving Dose Volume Histogram (DVH) based analysis of clinical outcomes using modern statistical techniques. A systematic answer to multiple comparisons concerns

Mirek Fatyga[1], Jiuyun Hu[2], and Jing Li[3]

[1]Department of Radiation Oncology, Mayo Clinic Arizona, Phoenix AZ, USA
[2]School of Computing & Augmented Intelligence, Arizona State University, Tempe, AZ
[3]H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta GA, USA

Dose Volume Histogram (DVH) indices remain the backbone of treatment plan optimization in radiation therapy (RT). Optimization constraints are derived from clinical outcomes studies which search for DVH indices that are predictors of toxicity. Conventional studies treat DVH indices as independent candidate predictors, selecting indices that meet the fixed statistical significance threshold. This process leads to inconsistent results among studies and multiple comparisons concerns. We present a more systematic approach to toxicity modelling, using knowledge constrained Generalized Linear Model (KC-GLM). This approach identifies one or more dose thresholds beyond which DVH indices become truly predictive and generates Normal Tissue Complications Probability (NTCP) models which are more interpretable and likely more generizable. We propose a hybrid model for clinical use, by using a single index NTCP model in the predictive dose range to establish optimization constraints, with the full model used for final NTCP verification.

1 Introduction

Dose Volume Histogram (DVH) indices remain a backbone of treatment planning in radiation therapy. Clinical goals of target coverage and Organs at Risk (OAR) protection are expressed mathematically as constraints on DVH indices which define the optimization criteria. DVH constraints must be derived from clinical outcomes studies which seek to identify indices that are predictive for clinical outcomes and build Normal Tissue Complications Probability (NTCP) models based on them. A planning constraint on a DVH index is derived from the NTCP model by imposing a limit on the allowable risk of toxicity. Many outcomes studies treat DVH indices as independent candidate predictors and retain only indices that meet the statistical significance cutoff of p=0.05. Such a procedure raises multiple comparisons concerns (Bonferroni correction) and often leads to inconsistencies among studies which use randomly chosen DVH indices. Inconsistencies among studies can lead to doubts as to the applicability of results to the clinical practice, or even their veracity. In addition, Bonferroni correction significantly increases required sizes of patient cohorts, which is difficult to attain in realistic clinical studies.

Modern Treatment Planning Systems (TPS) retain full dosimetry information in their databases and give users programmatic access to this information through Application Programmers Interfaces (API). APIs allow users to scan the space of DVH indices and search for "optimum" DVH indices with lowest p-values. Such searches would not be feasible, however, if one treated DVH indices as independent predictors subject to Bonferroni correction. New approaches to DVH analysis are needed to enable more systematic analysis of DVH index space. DVH indices are very strongly correlated because of the physics of dose deposition. Thus, a more comprehensive approach needs to employ an array of DVH indices and account for the strong correlations among them. The purpose of this work is to construct a plausible approach and apply it to clinical outcomes data.

2 Materials and Methods

2.1 Patient Outcomes Studies

We used two patient outcomes studies: (1) grade >= 2 acute rectal toxicity in a cohort of 79 patients treated with conventionally fractionated radiation therapy (RT) for prostate cancer, and (2) cardiac toxicity in a cohort of 134 locally advanced, Non-small Cell Lung Cancer (LA-NSCLC) patients treated with conventionally fractionated radiation therapy, with overall survival (OS) as a clinical endpoint. Both cohorts were treated at Mayo Clinic Arizona.

2.2 Dosimetry Data Extraction

All patients had their treatment plans in Eclipse TPS (Varian, Inc). ESAPI interface was used to extract arrays of $V\%_D$ and $D_{V\%}$ indices.

2.3 Initial DVH Index Scanning Investigation

A "common sense" scanning of DVH index space was based on inserting a single DVH index into a logistic regression or a Cox model, fitting a family of models, each model with a different index (index scanning) and examining fit quality for each model. A model with the lowest p-value and/or best AUC could be picked as the "best DVH index". Insights from this investigation were used to construct a more integrative model.

2.4 Base Statistical Model for the Integrative Approach

A Generalized Linear Model (GLM) was used as the base model, with Fused Lasso operator to reduce the dimensionality and to account for strong correlations among indices. The dosimetry was represented by an array:

$$\theta = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

where

$$X_i = V\%_{D_i}$$

which is the percentage of OAR volume exposed to dose D_i or greater. Array Θ was used with the logistic regression model for the study of acute rectal toxicity, and with the proportional hazards Cox model for the study of cardiac toxicity through overall survival. A base other than $V\%_D$ could also have been used to construct the model. However, the $V\%_D$ base is easiest to interpret in the context of radiobiological constraints.

2.5 Knowledge Constraints (KC-Lasso model)

Two additional constraints on parameters β are motivated by radiobiological considerations.

The positivity constraint: $\beta_i \geq 0$

Since the risk of toxicity is monotonically dependent on the value of Θ, positivity constraint holds that none of the contributing DVH indices X_i can decrease the risk of toxicity.

The monotonicity constraint: $\beta_{i+1} \geq \beta_i$

Consider $X_i = V\%_{D_i}$ set to some value (eg. 5%). What will be the contribution of this term to the risk if D_i increases? The monotonicity condition states that the contribution of $X_i$to risk increases or stays the same when D_i increases. This condition can only be enforced by the hierarchy of coefficients β, as the only remaining degree of freedom in the sum θ.

We gave the final model the name of Knowledge Constrained Lasso (**KC-Lasso**) [1],[2].

3 Results

3a. Acute rectal toxicity in RT for prostate cancer.

A "common sense" approach to searching for the most predictive index is shown in Fig.1. A single $D_{V\%}$index was used in logistic regression and a family of models in $D_{90\%} - D_{10\%}$range was fit. The left side of Fig.1 shows p-values for each model. One observes a structure in the plot which indicates that $D_{V\%}$ are not independent predictor candidates. We hypothesize that the region of monotonically declining p-values represents a transition region which is dominated by strong correlations among indices. The true predictive region is approximately $D_{10\%} - D_{25\%}$ while the region $D_{25\%} - D_{55\%}$ is a transition region. This hypothesis is supported by the right-hand plot in Fig.1. We plot logistic regression coefficients on the X-axis, and Spearman correlation coefficients with respect to $D_{20\%}$ on the Y-axis. These two sets of numbers are almost perfectly correlated ($R^2 = 0.995$). A remarkable aspect of this plot is that the X-axis holds all the complexities of the clinical response, while Y-axis is just a "play" on treatment planning numbers. And yet, these two sets of numbers are nearly perfectly correlated.

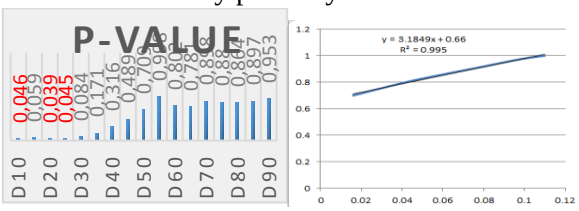


Fig.1 – left side: p-values for logistic regression models using one $D_{V\%}$index at a time.

Fig.1 – right side: a correlation between logistic regression coefficients in the transition region (X-axis) and Spearman correlation coefficient with respect to $D_{20\%}$ (Y-axis).

The KC-Lasso model was used in conjunction with the logistic regression model to detect the dose threshold separating the truly predictive region from the transition region in Fig.1. Two distinct dose thresholds which were predictive for grade 2 acute rectal toxicity were detected: 35Gy and 78Gy (Fig.2). Do note a p-value "break" in the hypothesized predictive region in Fig.1 which is interpreted by the KC-Lasso model as a second dose threshold. It is likely that the high threshold corresponds to rectal bleeding, while the lower threshold corresponds to the remainder of grade 2 toxicity.

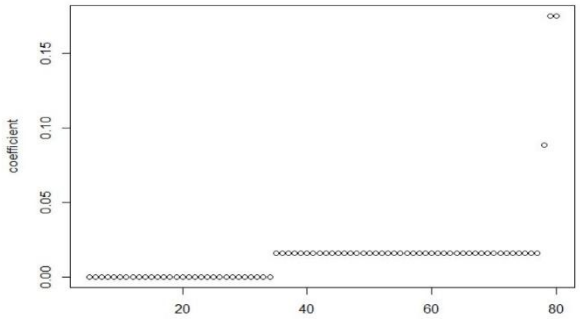


Fig.2 – coefficients β as a function of dose (Gy), acute rectal toxicity in prostate cancer treatments.

3b. Radiation induced cardiac toxicity in RT for LA-NSCLC.

The "common sense" approach is illustrated in Fig.3. A single heart $V\%_D$ index at a time was used in the Cox model, and Fig.3 shows p-values as a function of Dose D. One observes a pronounced minimum at D~55Gy with a monotonic decline in p-values, similar to the transition region in Fig.1. The minimum, rather than a predictive plateau, is caused by the loss of data for doses greater than the prescription dose (60Gy).

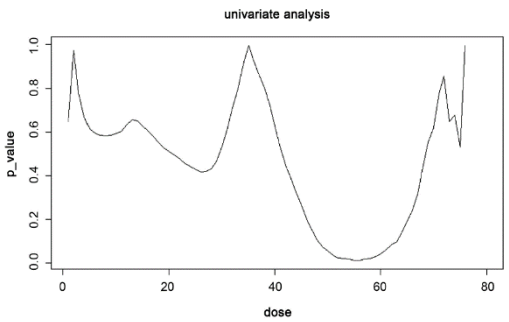


Fig.3 – p-values for a family of Cox models, using one $V\%_D$ index at a time, plotted as a function of Dose D.

The KC-Lasso model was used in conjunction with the proportional hazards Cox model and detected a single predictive threshold of 51Gy (p=0.035) for reduced overall survival due to cardiac irradiation in LA-NSCLC patients.

3c. Effects of knowledge constraints.

In Fig.4 we show results of fitting acute rectal toxicity data with several models with and without knowledge constraints.

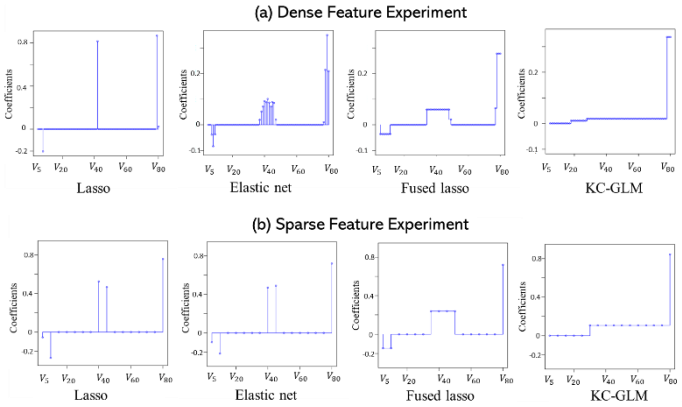


Fig.4 – coefficients β as a function of dose without (Lasso, Elastic Net, Fused Lasso) and with knowledge constraints (KC-GLM). Sparse and dense experiment refers to 5Gy and 1Gy steps. Grade 2 acute rectal toxicity in RT for prostate cancer.

In Fig.5 we show results of fitting Overall Survival against cardiac dosimetry with several models with and without knowledge constraints.

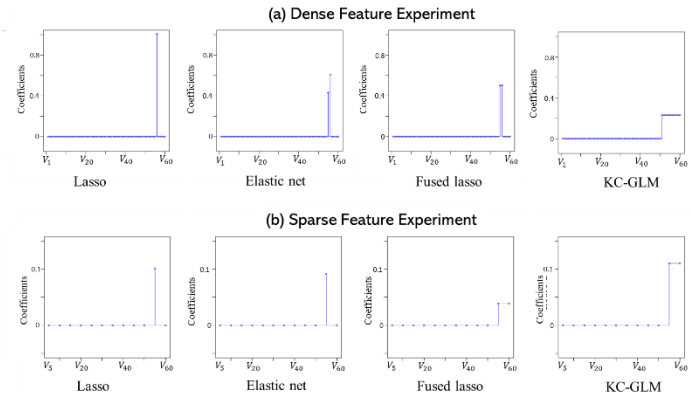


Fig.5 - -coefficients β as a function of cardiac dose without (Lasso, Elastic Net, Fused Lasso) and with knowledge constraints (KC-GLM). Sparse and dense experiment refers to 5Gy and 1Gy step. Overall Survival versus cardiac dosimetry in RT for LA-NSCLC.

Both Fig.4 and Fig.5 show that models without knowledge constraints (first three columns) produce results which are difficult to interpret, with negative and/or isolated coefficients. Such results are difficult to interpret because one would have to argue that one, specific $V\%_D$ index is predictive for the outcome while a neighboring index, at almost the same dose, is not. This result would be contradictory to the basic dose-volume relationship assumptions which form the foundation of treatment planning.

4 Discussion

The model which is presented in this work has been designed to account for an unavoidable intrinsic property of DVH indices, which is strong correlations among them because of the physics of radiation transport. One of the main objectives of this approach is to detect a dose threshold beyond which DVH indices become truly predictive for clinical outcomes. If one treats DVH indices as independent candidate predictors, one could assign significance to indices which are not truly predictive but generate low p-values because of their correlations with the predictive region. In addition, treating DVH indices as independent predictors implies the need for Bonferroni correction which is not applicable to an array of strongly correlated indices. Our work further suggests that models which are purely data driven must be augmented by knowledge constraints. As shown in Fig.4 and Fig.5, purely data driven models generate patterns of negative and isolated coefficients which are difficult to interpret and thus likely not generizable. By adding "knowledge constraints", which are motivated by radiobiology, one stabilizes models and makes them easier to interpret.

The results of KC-Lasso model cannot be used directly in currently available optimization algorithms; hence one needs to use a "hybrid approach" to make the model clinically useful. In a hybrid approach one selects a single DVH index in the predictive region identified by the model and uses this index to construct an NTCP model which is used to compute a DVH constraint, while the full model can be used post optimization to calculate the actual NTCP. While not perfect, this approach provides a path to clinical use. In the case of multiple thresholds (Fig.2), one can include two or more representative indices which correspond to two distinct, predictive regions. It is likely that two thresholds in Fig.2 correspond to two different "classes" of toxicity, namely rectal bleeding at high doses and the remainder of grade 2 toxicities at lower doses. In this case one can construct two approximate NTCP models

with an index in each range, or a single model with two indices.

5 Conclusion

Knowledge Constrained Generalized Linear Models can be used to improve on currently prevailing methodology of toxicity modelling in radiation therapy. Use of such models can address multiple comparison concerns and improve consistency of results from multiple studies. While the GLM cannot be used directly in present day optimization algorithms, a hybrid approach can lead to direct clinical applications. A GLM can identify truly predictive dose range, a single index can be chosen in this range, and be used as an approximate optimization constraint.

References

[1] J. Hu, M. Fatyga, W. Liu, et al. “Radiotherapy toxicity prediction using knowledge-constrained generalized linear model”. IISE Transactions on Healthcare Systems Engineering, 2023 DOI: https://doi.org/10.1080/24725579.2023.2227199

[2] J.R Niska, J. Hu, J. Li, et al. “Using Novel Statistical techniques to Accurately Determine the Predictive Dose Range in a Study of Overall Survival after Definitive Radiotherapy for Stage III Non-Small Cell Lung Cancer in Association with Heart Dose”. Journal of Cancer Therapy 2021 Sep; 12(9):505-529 https://doi.org/10.4236/jct.2021.129044.

On the implementation and evaluation of loss functions for robust multiple anatomy segmentation on CT images

Chengyin Li[1], Rafi Ibn Sultan[1], Hassan Bagher-Ebadian[2], Yao Qiang[1], Kundan Thind[2], Dongxiao Zhu[1], Indrin J. Chetty[3]
1 College of Engineering, Department of Computer Science, Wayne State University, Detroit MI, USA
2 Department of Radiation Oncology, Henry Ford Cancer Institute, Detroit MI, USA
3 Department of Radiation Oncology, Cedars-Sinai Medical Center, Los Angeles, USA

Abstract

In the realm of CT-based medical image segmentation, the selection of a loss function significantly influences deep neural network training. Loss functions, such as Cross Entropy (CE), Dice, Boundary, and Top-K, excel in specific areas but have inherent limitations for tissue segmentation. Our motivation arises from the understanding that relying solely on a single loss function can introduce bias to models. We have implemented a combination of CE, Dice, Boundary, and Top-K loss functions employing both loss-level linear combination and model-level ensemble on two large CT dataset cohorts to investigate model accuracy for auto-segmentation of normal tissues for different anatomic locations. Through extensive experiments on institutional and public datasets, consistent observations emerged: (1) linear combination of loss functions did not exhibit a clear advantage over single loss methods; (2) ensemble-based methods demonstrated a 2% - 7% average increase in DSC scores, and appreciable reductions in Hausdorff distances (HD's) and Average Surface Distances (ASD's) compared to both linear combination or single loss methods; and (3) ensemble approach with optimized weights generally characterized finer details in predicted masks, as evidenced by qualitative analyses. This study provides quantitative results on the accuracy of different loss functions employed in machine learning networks for automatic medical image segmentation. Enhanced accuracy was achieved using an ensemble approach with optimized model weights, compared to the use of a single loss function or a linear combination of loss functions.

1. Introduction

The training of DL models involves iterative parameter adjustments to minimize training errors, typically employing optimization methods such as stochastic gradient descent [1]. To quantify these errors, a loss function, also referred to as a cost function, is utilized to minimize the error between the algorithm prediction and the benchmark and thereby produce a convergent result. In the domain of medical image segmentation on CT images, CE and/or Dice are used commonly as loss functions [2-4]. CE loss [5] gauges the pixel-wise discordance between predicted class probabilities and ground truth labels. However, it tends to assign equal importance to background and foreground regions during training, resulting in potential class imbalance. Conversely, Dice loss [6] evaluates the overlap between predicted and ground truth labels, demonstrating effectiveness in handling class imbalance by emphasizing mask intersection, which is crucial for unbalanced classes. However, Dice as a sole loss function is subject to errors when dealing with small and irregularly shaped organ volumes. To enhance the precision of medical image segmentation, researchers have also introduced alternative loss functions such as Boundary loss [7] and Top-K loss [8]. Boundary Loss proves effective for smaller, imbalanced datasets but lacks global context and may not robustly address class imbalance. It is typically combined with other loss functions for more comprehensive segmentation tasks. Top-K loss aims to steer networks toward challenging samples during training, offering benefits for such classes or when imbalanced class distributions exist in the training dataset. Medical image segmentation relies on a range of loss functions, each with its unique strengths and weaknesses [9]. As such, segmentation based on a single loss model can introduce bias to the model training process. Improvement in loss-driven algorithms and computation power has made it possible to use these models simultaneously to improve robustness in medical image segmentation.

One method of concurrently utilizing multiple segmentation losses involves creating a unified model through a linear combination of various loss functions [9, 10]. This approach calculates a weighted sum of the individual loss function's contributions, with each loss being assigned a weight reflecting its relative significance within the overall loss. Such an approach permits the prioritization of specific aspects of the segmentation task, which can be tailored to domain knowledge, or the specific challenges posed by the dataset. Advanced options like ensemble learning [11] can be appealing for greater flexibility, especially when dealing with variations in architecture, hyperparameters, and training strategies. The ensemble approach frequently leads to enhanced segmentation performance by harnessing diversity among the models. Ensemble models have been extensively employed in medical image classification tasks, often in diverse dataset splitting scenarios and utilizing the CE loss function. However, there is a paucity of studies assessing the impact of ensemble DL models trained with different loss functions for medical image segmentation.

This study contributes to the existing knowledge as follows: (1) Evaluates the impact of unimodal and multimodal loss functions in multi-anatomy CT-based image segmentation; (2) Proposes a learnable ensemble approach that dynamically combines CE, Dice, Boundary, and Top-k loss-induced feature extraction modules to enhance segmentation performance and robustness; (3) Demonstrates the benefits of learnable ensemble models.

2. Materials and Methods

2.2.1. Network architecture

With the introduction of Fully Convolutional Networks (FCNs), the Attention U-Net (AttUNet) [12] stands out as an extension of the U-Net [5] architecture. AttUNet retains the U-Net's encoder-decoder architecture, capturing hierarchical features through generation of a segmentation map. AttUNet uses soft attention gates within skip connections to actively suppress irrelevant regions of the feature maps [12], thereby reducing redundant features and enabling characterization of fine details through the attention mechanism [12]. Recently, vision transformer approaches have emerged with the property of self-attention, which enables encoding of long-range dependencies and highly effective feature representations. Among these techniques, the shifted window UNet Transformer (SwinUNETR [4]) has shown superior accuracy for medical image segmentation. SwinUNETR exhibits several key design features: (a) adopts the Swin Transformer [13] architecture which captures long-range dependencies and contextual information through multi-layer self-attention mechanisms; (b) employs an encoder-decoder structure with the Swin Transformer to produce segmentation maps; (c) incorporates multi-scale processing via multiple encoder levels to handle objects of various sizes effectively; (d) integrates U-Net-style skip connections with residual operations to preserve spatial resolution and fine details; (e) employs patch-based image processing for efficient computations of large image datasets.

Here we employ AttUNet and SwinUNETR as two representative 3D segmentation architectures to investigate the impact of single and multiple loss functions on segmentation accuracy.

2.2.1. CE loss

Cross Entropy (CE) is derived from the Kullback-Leibler (KL) Divergence [14], which measures dissimilarity between two probability distributions, typically denoted as *P* and *Q*.

2.2.2. Dice loss

Dice loss directly optimizes the Dice Similarity Coefficient (DSC), a widely employed metric for segmentation evaluation. In general, two variants of the Dice loss are recognized (Isensee et al. 2021). One of them includes squared terms in the denominator (Milletari et al. 2016)

In the following experiments, the linear (non-squared version) as the default function.

2.2.3. Boundary loss

Boundary loss was introduced with the aim of minimizing dissimilarities between predicted and ground truth segmentations [7].

Here we use boundary loss in conjunction with the distribution-based CE loss:

$$\mathcal{L}_{CE+BD} = \alpha\mathcal{L}_{CE} + (1-\alpha)\mathcal{L}_{BD},$$

where α is a hyperparameter and can be optimized through empirical studies.

2.2.4. Top-K loss

Top-K loss is a variation of cross entropy designed to prioritize challenging samples during training. It retains the *K* percent worst pixels for loss, irrespective of their loss/probability values [8].

2.3.1 Linear combination

As depicted in Figure 1A, the combined loss or total loss is achieved by a linear combination of multiple losses.

$$\mathcal{L}_{Linear} = \lambda_1\mathcal{L}_{Dice} + \lambda_2\mathcal{L}_{CE} + \lambda_3\mathcal{L}_{CE+BD,}$$

where λ_1, λ_2, and λ_3 are the importance/contribution weights for each loss component. We found that when λ_1, λ_2, and λ_3 are set equally to the same weighting of 1.0, the obtained models after training showed better performance compared to other settings. Therefore, we represented the loss function as $\mathcal{L}_{Linear} = \mathcal{L}_{Dice} + \mathcal{L}_{CE} + \mathcal{L}_{CE+BD}$.

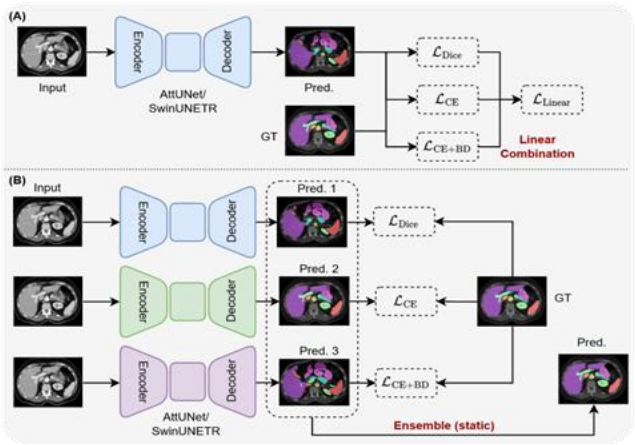


Figure 1: (a) The pipeline for a linear combination of different segmentation loss functions, and (b) the static multiple loss-driven model ensemble model strategy for CT image segmentation.

2.3.2. Static ensemble model

In this scenario (as shown in Figure 1B), we use an averaging ensemble technique by combining the predictions of multiple individual models trained with different loss functions. For instance, we have *K* different versions of trained models $f_{\theta_k} : k \in 1, ..., K$, where θ denotes the parameters of a model. Each of these versions of the model can rely on a specific loss function, e.g., $\mathcal{L}_{CE}$, $\mathcal{L}_{Dice}$, and $\mathcal{L}_{CE+BD}$.

The ensemble probabilistic prediction usually has a threshold of 0.5 to obtain a binary segmentation for each class for medical images. We refer to this approach as the static ensemble model.

2.3.3. Learnable ensemble model

With the learnable ensemble method, feature map extraction process is achieved by the concatenation of three branches of trained models (e.g., from $\mathcal{L}_{Dice}$, $\mathcal{L}_{CE}$, and $\mathcal{L}_{CE+BD}$). Then a few more iterations of training using the

Top-k loss function ($\mathcal{L}_{\text{Top-K}}$) are applied for the final segmentation prediction. We refer to this approach as the learnable ensemble model.

3.1. Datasets

All experiments were implemented on a server equipped with 8 Nvidia A100 GPUs, each with 40 GB of memory. All experiments are conducted in the PyTorch framework in Python 3.8.13, and each model is trained on a single GPU.

3.1.1. An institutional dataset: pelvic organs segmentation dataset

Institutional review board (IRB) approval was obtained for this study. We retrospectively selected planning CT and structure datasets from 300 patients diagnosed with prostate cancer. These 300 cases were randomly divided into three sets: a training set consisting of 225 cases, a validation set with 30 cases, and a testing set containing 45 cases. The testing datasets were held out from the training/validation steps and are considered "unseen". The models were trained for 200 epochs using the AdamW optimizer with an initial learning rate of 5e−4. An exponential learning rate scheduler with a warmup of 5 epochs was applied to the optimizer. Data augmentation was applied 'on the fly' during training. Random flip, rotation, and intensity scaling were used as augmentation transforms, with probabilities of 0.1, 0.1, and 0.2, respectively. We compared the automatic contours generated by our model against ground-truth .

3.1.2. Public dataset: CT organ segmentation dataset (CT-ORG)

A publicly available dataset has been used for evaluation in the CT-ORG task, as described in the reference [15]. The dataset comprises 140 CT scans, each including manual contours of the liver, lung, bladder, kidney, and bones. To maintain consistency with prior work [15], we adopted the same data split for training and testing. Nineteen CT datasets were reserved for testing (holdout), while the remaining 81 scans were used for training.

3.1.3. Ensemble modeling

In the static ensemble strategy, we directly perform averaging over several individually trained models, each utilizing different a loss function, without undergoing any further training. Conversely, in the learning ensemble strategy, we applied an additional training phase encompassing 20 epochs using the Top-k loss function for generating the final segmentation mask. We kept the other feature map generation modules static throughout this phase. We also applied the same data processing and augmentation techniques, albeit with a relatively smaller learning rate set to 1e−4.

We employed the Dice Similarity Coefficient (DSC), the 95th percentile Hausdorff Distance (HD), and Average Surface Distance (ASD) to assess the accuracy of segmentation.

3. Results

Table 1: DSC, HD, and ASD for the different settings using **AttUNet** on the **institutional pelvic dataset**. Each value represents the mean performance (+/- standard deviation). The most accurate results are shown in bold type. * denotes statistically significant differences (p<0.05) between the ensemble method and other non-ensemble methods.

Organ	Method	DSC ↑	HD (mm) ↓	ASD (mm) ↓
Prostate	$\mathcal{L}_{\text{Dice}}$	0.87 ± 0.03	4.66 ± 2.58	1.55 ± 0.53
	$\mathcal{L}_{\text{CE}}$	0.87 ± 0.02	4.48 ± 1.73	1.56 ± 0.42
	$\mathcal{L}_{\text{CE+BD}}$	0.87 ± 0.03	4.55 ± 1.78	1.63 ± 0.46
	$\mathcal{L}_{\text{Linear}}$	0.87 ± 0.03	4.68 ± 1.83	1.61 ± 0.42
	Ensemble (static)	0.88 ± 0.02*	4.39 ± 1.68	0.87 ± 0.03*
	Ensemble (learnable)	**0.89±0.02***	**4.12±1.63***	**0.79±0.06***
Bladder	$\mathcal{L}_{\text{Dice}}$	0.94 ± 0.02	2.62 ± 2.20	0.84 ± 0.45
	$\mathcal{L}_{\text{CE}}$	0.94 ± 0.02	2.70 ± 3.19	0.84 ± 0.63
	$\mathcal{L}_{\text{CE+BD}}$	0.94 ± 0.03	2.81 ± 3.19	0.93 ± 0.64
	$\mathcal{L}_{\text{Linear}}$	0.94 ± 0.02	2.70 ± 3.15	0.88 ± 0.62
	Ensemble (static)	0.95 ± 0.02*	**2.57±2.13***	0.82 ± 0.61*
	Ensemble (learnable)	**0.95±0.02***	2.65 ± 2.15*	**0.79±0.06***
Rectum	$\mathcal{L}_{\text{Dice}}$	0.87 ± 0.03	6.59 ± 5.42	1.66 ± 0.78
	$\mathcal{L}_{\text{CE}}$	0.87 ± 0.03	6.08 ± 4.80	1.47 ± 0.62
	$\mathcal{L}_{\text{CE+BD}}$	0.87 ± 0.03	6.70 ± 5.41	1.67 ± 0.75
	$\mathcal{L}_{\text{Linear}}$	0.87 ± 0.03	6.86 ± 6.08	1.62 ± 0.81
	Ensemble (static)	0.88 ± 0.03*	6.27 ± 5.19*	1.57 ± 0.73
	Ensemble (learnable)	**0.88±0.02***	**6.13±4.65***	**1.42±0.58***

Table 2: DSC, HD, and ASD for the different settings using **SwinUNETR** on the **institutional pelvic dataset**. Each value represents the mean performance (+/- standard deviation). The most accurate results are shown in bold type. * denotes statistically significant differences (p<0.05) between the ensemble method and other non-ensemble methods.

Organ	Method	DSC ↑	HD (mm) ↓	ASD (mm) ↓
Prostate	$\mathcal{L}_{\text{Dice}}$	0.86 ± 0.03	4.60 ± 1.42	1.61 ± 0.45
	$\mathcal{L}_{\text{CE}}$	0.85 ± 0.03	4.89 ± 1.82	1.67 ± 0.48
	$\mathcal{L}_{\text{CE+BD}}$	0.86 ± 0.03	4.81 ± 1.58	1.73 ± 0.43
	$\mathcal{L}_{\text{Linear}}$	0.86 ± 0.03	4.61 ± 1.22	1.62 ± 0.39
	Ensemble (static)	0.87 ± 0.02*	4.55 ± 1.23	1.59 ± 0.41*
	Ensemble (learnable)	**0.88±0.02***	**4.23±1.13***	**1.49±0.40***
Bladder	$\mathcal{L}_{\text{Dice}}$	0.94 ± 0.02	2.71 ± 2.17	0.82 ± 0.23
	$\mathcal{L}_{\text{CE}}$	0.93 ± 0.02	2.68 ± 1.69	0.83 ± 0.21
	$\mathcal{L}_{\text{CE+BD}}$	0.93 ± 0.03	2.97 ± 2.64	0.91 ± 0.31
	$\mathcal{L}_{\text{Linear}}$	0.93 ± 0.03	3.10 ± 3.01	0.89 ± 0.41
	Ensemble (static)	0.94 ± 0.02*	2.60 ± 1.53*	**0.81±0.20***
	Ensemble (learnable)	**0.95±0.02***	**2.53±1.54***	0.78 ± 0.17*
Rectum	$\mathcal{L}_{\text{Dice}}$	0.85 ± 0.05	7.47 ± 6.19	1.78 ± 0.97
	$\mathcal{L}_{\text{CE}}$	0.84 ± 0.04	8.14 ± 6.11	1.76 ± 0.87
	$\mathcal{L}_{\text{CE+BD}}$	0.84 ± 0.05	8.10 ± 6.37	2.17 ± 1.31
	$\mathcal{L}_{\text{Linear}}$	0.86 ± 0.04	8.23 ± 6.84	2.00 ± 1.11
	Ensemble (static)	0.86 ± 0.03*	7.11 ± 5.92*	1.77 ± 0.91
	Ensemble (learnable)	**0.88±0.02***	**6.04±4.21***	**1.36±0.61***

Table 3: DSC, HD, and ASD for the different settings using **SwinUNETR** on the **CT-ORG dataset**. Each value represents the mean performance (+/- standard deviation). The most accurate results are shown in bold type. * denotes statistically significant differences (p<0.05) between the ensemble method and other non-ensemble methods.

Organ	Method	DSC ↑	HD (mm) ↓	ASD (mm) ↓
Lung	$\mathcal{L}_{\text{Dice}}$	0.97 ± 0.02	10.37 ± 12.42	0.15 ± 0.14
	$\mathcal{L}_{\text{CE}}$	0.98 ± 0.01	6.17 ± 11.22	0.14 ± 0.13
	$\mathcal{L}_{\text{CE+BD}}$	0.98 ± 0.01	5.56 ± 7.78	0.15 ± 0.13
	$\mathcal{L}_{\text{Linear}}$	0.98 ± 0.01	7.29 ± 13.99	0.15 ± 0.13
	Ensemble (static)	0.98 ± 0.01*	**4.97±7.61***	0.14 ± 0.13*
	Ensemble (learnable)	**0.98±0.01***	5.23 ± 6.25*	**0.12±0.11***
Liver	$\mathcal{L}_{\text{Dice}}$	0.94 ± 0.02	3.90 ± 3.36	0.65 ± 0.37
	$\mathcal{L}_{\text{CE}}$	0.94 ± 0.01	3.03 ± 3.49	0.59 ± 0.17
	$\mathcal{L}_{\text{CE+BD}}$	0.94 ± 0.02	3.10 ± 1.46	0.72 ± 0.29
	$\mathcal{L}_{\text{Linear}}$	0.95 ± 0.02	2.91 ± 1.83	0.58 ± 0.21
	Ensemble (static)	0.96 ± 0.01*	2.55 ± 1.34*	0.56 ± 0.15*
	Ensemble (learnable)	**0.96±0.01***	**2.12±0.93***	**0.35±0.10***
Kidney	$\mathcal{L}_{\text{Dice}}$	0.92 ± 0.03	3.26 ± 1.11	0.61 ± 0.21
	$\mathcal{L}_{\text{CE}}$	0.91 ± 0.03	6.29 ± 14.08	0.58 ± 0.19
	$\mathcal{L}_{\text{CE+BD}}$	0.91 ± 0.03	6.97 ± 13.76	0.69 ± 0.22
	$\mathcal{L}_{\text{Linear}}$	0.92 ± 0.03	3.03 ± 1.76	0.52 ± 0.17
	Ensemble (static)	0.93 ± 0.03*	3.25 ± 0.96	0.56 ± 0.21*
	Ensemble (learnable)	**0.94±0.02***	**2.21±0.82***	**0.46±0.22***
Bladder	$\mathcal{L}_{\text{Dice}}$	0.86 ± 0.08	9.97 ± 18.31	0.77 ± 0.48
	$\mathcal{L}_{\text{CE}}$	0.83 ± 0.13	7.29 ± 12.28	0.90 ± 0.80
	$\mathcal{L}_{\text{CE+BD}}$	0.83 ± 0.13	5.71 ± 10.73	1.01 ± 0.71
	$\mathcal{L}_{\text{Linear}}$	0.84 ± 0.13	5.86 ± 10.54	0.89 ± 0.67
	Ensemble (static)	0.87 ± 0.09*	4.71 ± 8.90*	0.81 ± 0.61
	Ensemble (learnable)	**0.87±0.08***	**3.56±6.38***	**0.68±0.49***
Bone	$\mathcal{L}_{\text{Dice}}$	0.88 ± 0.04	4.91 ± 3.52	0.76 ± 0.35
	$\mathcal{L}_{\text{CE}}$	0.89 ± 0.05	5.50 ± 4.25	0.98 ± 0.37
	$\mathcal{L}_{\text{CE+BD}}$	0.88 ± 0.05	5.84 ± 3.87	1.20 ± 0.48
	$\mathcal{L}_{\text{Linear}}$	0.89 ± 0.05	4.88 ± 2.82	1.04 ± 0.28
	Ensemble (static)	0.90 ± 0.04*	4.85 ± 3.28*	0.98 ± 0.46

4. Discussion

The learnable ensemble model achieved the most accurate segmentation results across nearly all metrics for the institutional pelvic dataset and the public CT-ORG dataset across multiple organs.

These findings agree with previous studies showing improved accuracy of ensemble models for image analysis tasks, albeit on different network architectures [16, 17]. By combining multiple models trained on different loss functions, the proposed ensemble approach overcomes limitations of individual loss functions and integrates their complementary strengths. For instance, CE loss optimizes pixel-wise probabilities, enabling precise localization but ignoring class imbalances. Dice loss directly maximizes overlap with ground truth, overcoming imbalance but sacrificing localization accuracy. Boundary loss focuses on contour distances, handling smaller objects effectively. Top-K loss concentrates on challenging samples regardless of loss value. No single loss completely addresses all segmentation challenges simultaneously. The ensemble approach enables harnessing of complementary advantages of different losses.

The learnable ensemble model achieves superior performance by increasing model diversity through an additional trainable module, enabling dynamic weighting instead of fixed averaging to aggregate predictions, and allowing targeted optimization of challenging samples. This combination provides greater segmentation accuracy and robustness compared to the static ensemble or individual models with distinct loss functions, as evidenced by the statistically significant improvements in segmentation accuracy across multiple anatomic locations. However, the ensemble approach also has limitations. Training multiple models increases computational requirements for image analysis. Additionally, increasing the sample size will benefit from understanding the tradeoffs between variance and bias in the training datasets. Although we have utilized data augmentation, the influence of augmentation on the training data bias, variance and ultimately the generalization error is not fully understood. Determining the optimal combination of loss functions and models for a given segmentation task is a challenging task and one that is under further investigation. Studies providing greater insight into designing high-performance ensembles tailored to different imaging modalities and anatomical structures would be of benefit to the field.

5. Conclusion

Comprehensive experiments, conducted on institutional and public datasets, provide evidence of the superior performance achieved through a learnable ensemble model compared to results based solely on a single loss function or a linear combination of loss functions. These findings will inform future efforts to develop automated segmentation techniques for significantly enhancing efficiency and accuracy of segmentation of anatomy on CT images.

Acknowledgement: This work was supported in part by a grant from Varian Medical Systems, Siemens Healthineers.

References

[1] Wang L, Yang Y, Min R, et al. Accelerating deep neural network training with inconsistent stochastic gradient descent. Neural Networks. 2017;93:219-29.

[2] Li C, Bagher-Ebadian H, Goddla V, et al. FocalUNETR: A Focal Transformer for Boundary-aware Segmentation of CT Images. arXiv preprint arXiv:221003189. 2022.

[3] Balagopal A, Kazemifar S, Nguyen D, et al. Fully automated organ segmentation in male pelvic CT images. Physics in Medicine & Biology. 2018;63:245015.

[4] Li C, Bagher-Ebadian H, Sultan RI, et al. A new architecture combining convolutional and transformer-based networks for automatic 3D multi-organ segmentation on CT images. Med Phys 50:6990; 2023

[5] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18: Springer; 2015. p. 234-41.

[6] Milletari F, Navab N, Ahmadi S-A. V-net: Fully convolutional neural networks for volumetric medical image segmentation. 2016 fourth international conference on 3D vision (3DV): IEEE; 2016:565-71.

[7] Kervadec H, Bouchtiba J, Desrosiers C, et al. Boundary loss for highly unbalanced segmentation. International conference on medical imaging with deep learning: PMLR; 2019. p. 285-96.

[8] Wu Z, Shen C, Hengel Avd. Bridging category-level and instance-level semantic image segmentation. arXiv:160506885. 2016.

[9] Ma J, Chen J, Ng M, et al. Loss odyssey in medical image segmentation. Medical Image Analysis. 2021;71:102035.

[10] Kendall A, Gal Y, Cipolla R. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. Proceedings of the IEEE conference on computer vision and pattern recognition2018. p. 7482-91.

[11] Dong X, Yu Z, Cao W, et al. A survey on ensemble learning. Frontiers of Computer Science. 2020;14:241-58.

[12] Oktay O, Schlemper J, Folgoc LL, et al. Attention u-net: Learning where to look for the pancreas. arXiv preprint arXiv:180403999. 2018.

[13] Liu Z, Lin Y, Cao Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows. Proceedings of the IEEE/CVF international conference on computer vision2021. p. 10012-22.

[14] Csiszar I. Divergence Geometry of Probability Distributions and Minimization Problems. The Annals of Probability. 1975;3:146-58, 13.

[31] Rister B, Yi D, Shivakumar K, et al. CT-ORG, a new dataset for multiple organ segmentation in computed tomography. Scientific Data. 2020;7:381.

[32] Rajaraman S, Zamzmi G, Antani SK. Novel loss functions for ensemble-based medical image classification. Plos one. 2021;16:e0261307.

[33] Müller D, Soto-Rey I, Kramer F. An analysis on ensemble learning optimized medical image classification with deep convolutional neural networks. Ieee Access. 2022;10:66467-80.

Scripted dosimetric validation of AI auto-contouring in prostate radiotherapy

Daniel R. Warren[1], Adam Chalkley[1], and Ian Stronach[1]

[1] Clinical Computing and Imaging Science Service (CCISS), Department of Medical Physics, University Hospitals Birmingham NHS Foundation Trust, Birmingham, UK

Abstract Artificial Intelligence (AI) contouring tools have the potential to optimise radiotherapy workflows and increase standardisation. Verification of these tools remains an important issue, complicated by rapid development cycles and explainability issues. The dosimetric impact of AI tools is of primary clinical importance, but models have largely been validated via geometric measures. In this study a simple scripted planning process has been applied to fifty prostate cases using the original clinical contours and four AI auto-contouring products, with unedited structures used to produce planning target volumes. Dosimetric clinical goals were assessed for the auto-contoured plans, and compared to standard geometric measures. Commonly used geometric measures, such as Dice, were found to be less useful surrogates for dosimetric measures than a simple metric that took into account the plan geometry (auto-contoured prostate length in z direction), which correlated with a systematic underdose when assessed on the clinical contours. This work suggests that a simple scripted planning approach to assessing auto-contouring output may help identify clinically relevant dosimetric changes.

1 Introduction

The accurate contouring of Organs at Risk (OARs) and target volumes is a key step in the production of radiotherapy treatment plans. Manual contouring is a time-consuming process and inter-observer variation remains a large source of uncertainty in radiotherapy treatments [1]. Many methodologies have been discussed for automating contour generation [2], normally focusing on the delineation of OARs, since tumour volumes show greater variation between patients and require more clinical judgement. Methods using artificial intelligence (AI), particularly deep learning, represent the current state of the art, and they are now being readily adopted into clinical practice [3]. In the UK, the National Institute for Health and Care Excellence (NICE) has recently identified nine commercial applications that have regulatory approval for use within the NHS [4]. In the case of prostate cancer, where a complete organ is typically treated, these auto-contouring systems can potentially outline both target and OAR volumes.

The NICE guidelines highlight a current evidence gap in the dosimetric impact of using AI auto-contouring tools. In the majority of studies, agreement to clinician contours has been assessed using geometric measures [5], but their relationship to dosimetric changes is unclear. To assess dosimetric consequences it is necessary to repeat the treatment planning process on a range of contours, which can require considerable manual effort. This presents a challenge to validating new or updated AI tools and for performing routine Quality Assurance (QA) [3]. NICE guidance, in common with many studies in the literature, recommends manual review of the contours for each patient for QA purposes. This review is an additional time burden, particularly for adaptive radiotherapy, where plans are generated on daily images [6], and may introduce additional error. There are therefore calls for a standardised, automated approach to validation, and interest in using AI tools to facilitate this [3]. Regional or national programmes may be required to avoid the time taken in commissioning new models negating time savings from using them.

The study presented here is development work, intended to inform the methodology for an internal procurement exercise to evaluate commercial AI auto-contouring tools. The primary objective was to automate the generation of plans from manual and automated contours in order to assess the impact of auto-contouring on key clinical dosimetric goals. The automation process was not intended to produce clinical plans, but to provide a consistent level of optimisation, enabling comparison between systems. Additionally, geometric measures were calculated and correlated to the dosimetric goals, to assess whether they could be used as surrogates for QA purposes.

2 Materials and Methods

The local Oncology Information System database was queried to identify a random sample of 50 patients who received radical prostate radiotherapy in the 6-month period starting June 2021 (prior to the introduction of AI contouring systems in our clinic). DICOM datasets comprising CT scans and oncologist-approved manual structure sets were extracted from the clinical image contouring package and de-identified for subsequent processing. The manual structure sets were used as the gold-standard reference in all subsequent analysis.

Four auto-contoured structure sets were generated for each patient by sending the CT scan to three commercial AI contouring systems (all certified medical devices) and one open-source system. Systems will not be identified, since this preliminary methodological work is not sufficient to draw conclusions on their relative performance when commissioned and optimised for clinical use.

Geometric agreement of the manual and auto-contoured structure sets was assessed using scikit-rt v0.3.27 [7], producing standard metrics including Dice coefficients, Haussdorf distance and mean distance to conformity [8].

Dosimetric agreement of radiotherapy treatment plans derived from the manual and auto-contoured structure sets was assessed via an automated planning process. This used

a suite of Python programs interfacing with the RayStation 11B (RaySearch Laboratories, Stockholm, Sweden) scripting API to perform five tasks for each structure set:

1. **Data import:** import structure set and associated CT scan as a new case
2. **Structure name normalisation:** rename structures to match a common naming convention
3. **Create derived ROIs (Regions of Interest):** use Boolean algebra and apply margins to generate Planning Target Volume (PTVs) and optimisation control structures
4. **Beam setup and optimisation:** set up a full single arc VMAT beam, add class solution dose objectives and run optimisation
5. **Plan evaluation:** extract dose-volume statistics for the optimised dose distribution using a) the planning structure set and b) the gold-standard manual set

Automated planning followed the local clinical protocol for intermediate risk prostate cancer (CHHiP volume definitions [9]; 60Gy/20fx prescription) with adaptations to accommodate limitations in the structures output by some packages (OARs used were rectum, bladder and femoral heads only; seminal vesicles were replaced by a proxy structure 2 cm posterosuperior to the prostate).

3 Results and Discussion

	Dice	HD (mm)	MDC (mm)	ΔVol (cm^3)
Sys	***Bladder (mean ref vol. 316 ± 113 cm^3)***			
A	**0.96 ± 0.02**	12.7 ± 4.2	**0.2 ± 0.1**	-7.8 ± 11.4
B	0.96 ± 0.02	**12.0 ± 4.6**	0.2 ± 0.2	-3.9 ± 15.5
C	0.95 ± 0.03	13.7 ± 6.6	0.3 ± 0.3	**-1.0 ± 25.4**
D	0.95 ± 0.03	12.7 ± 6.8	0.3 ± 0.4	-19.5 ± 31.4
Sys	***Femoral Heads (mean ref vol. 144 ± 35 cm^3)***			
A	0.52 ± 0.09	68.0 ± 23.7	60.8 ± 67.9	**213 ± 64**
B	**0.52 ± 0.09**	**63.9 ± 12.2**	**52.8 ± 21.2**	222 ± 65
C	0.41 ± 0.09	152 ± 26	176 ± 111	370 ± 86
D	0.48 ± 0.10	84.8 ± 22.7	75.5 ± 46.8	258 ± 68
Sys	***Prostate (mean ref vol. 51.1 ± 28.2 cm^3)***			
A	**0.82 ± 0.05**	11.8 ± 3.9	**1.0 ± 0.4**	-9.4 ± 7.9
B	0.81 ± 0.06	**11.7 ± 3.8**	1.0 ± 0.5	-12.1 ± 9.4
C	0.78 ± 0.06	12.9 ± 3.6	1.3 ± 0.5	-12.9 ± 9.5
D	0.81 ± 0.06	12.2 ± 2.9	1.1 ± 0.6	**-3.6 ± 8.5**
Sys	***Rectum (mean ref vol. 64.8 ± 20.9 cm^3)***			
A	0.82 ± 0.09	17.4 ± 9.8	2.4 ± 3.3	-2.9 ± 20.3
B	0.84 ± 0.07	17.3 ± 8.2	1.9 ± 2.6	3.6 ± 12.9
C	**0.84 ± 0.05**	**15.6 ± 6.4**	**1.3 ± 1.3**	**0.2 ± 12.0**
D	0.76 ± 0.10	33.8 ± 14.2	7.0 ± 8.2	14.8 ± 20.7

Table 1: Geometric comparison metrics (mean ± standard deviation) for ROIs generated by the four auto-contouring products compared to oncologist-approved manual contours across 50 patients. HD = Hausdorff distance; MDC = Mean distance to conformity; Bold = best score.

Table 1 presents a summary of geometrical measures comparing the primary prostate planning ROIs produced by each auto-contouring system to the reference clinician contours. The product with the closest match to the clinical contours varies by OAR and scoring metric. Score ranges overlap in the majority of cases, suggesting broadly equivalent geometric performance of all systems with a dataset of this size. In general, volume differences between AI and reference ROIs were systematically in the same direction for all products e.g. the prostate and bladder were on average smaller than the references; this is not readily apparent from the unsigned Dice and distance metrics. Femoral heads showed the worst agreement overall, which is due to differences in contouring protocols; products tended to contour the whole femur rather than the head.

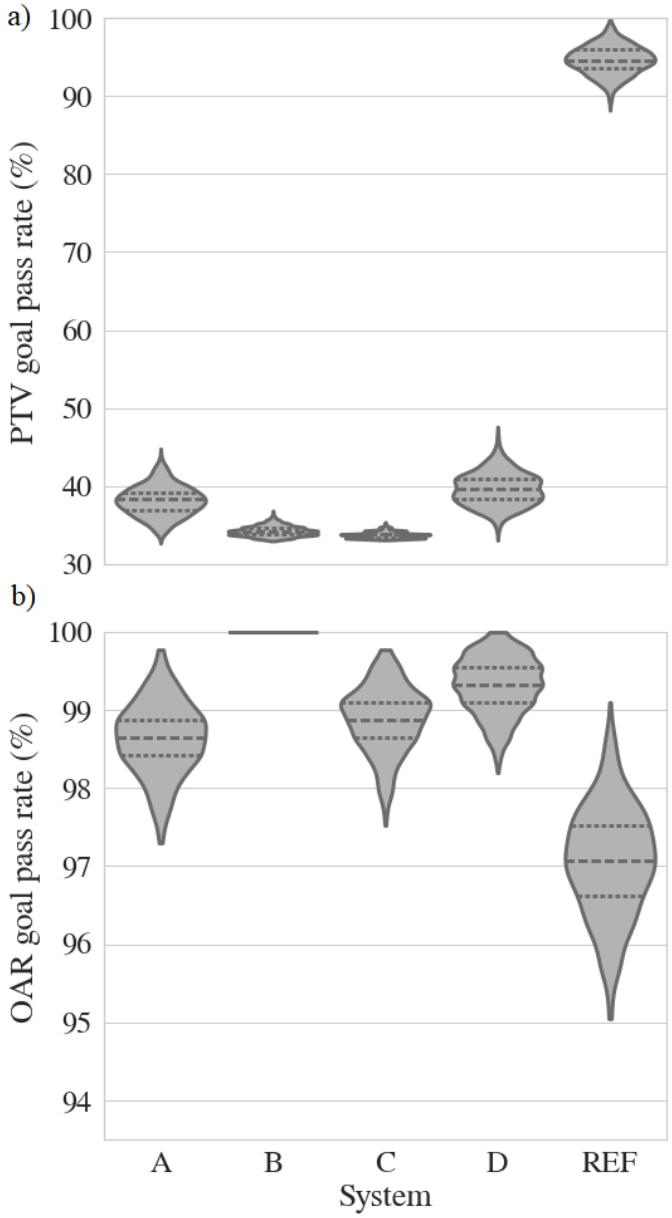


Figure 1: Pass rate for a) PTV and b) OAR clinical goals for automated treatment plans generated using structure sets from each auto-contouring system (A-D) and manual contours (REF) for 37 patients. Goals were assessed using the reference contours. Distributions calculated via bootstrap (n=500). Dashed lines indicate quartiles.

13 patients (26%) were excluded from dosimetric analysis due to missing or overlapping contours in one or more structure sets, which would require manual intervention within the planning pipeline to resolve.

Figure 1 shows the percentage pass rate for all clinical PTV and OAR objectives across all eligible patients when the standardised planning approach is applied to each product's contours, and to the reference clinical contours. Results presented use the gold-standard reference contours to assess goals across all products, since they represent the standard for clinical acceptability.

The PTV goal pass rates are poorer for each of the AI products, representing a reduction in reference target coverage when auto-contoured structures are used to generate the PTV. Conversely, it is notable that the OAR goals for AI contours had a higher pass rate in general compared to those produced on reference contours (for product B the rate was 100% meaning that bootstrapping did not produce a distribution).

The trends in PTV and OAR goal pass rate are partly driven by a general reduction in size of the prostate in auto-contoured cases (Table 1). Prostate ROI size is the main determinant of the size of the PTV in this treatment planning protocol, so coverage would be expected to be poorer in auto-contoured cases assessed using the larger PTVs of the reference dataset. Similarly, OARs will abut the smaller PTVs of the auto-contoured contours to a lesser degree, making it easier for the optimiser to minimise doses in them.

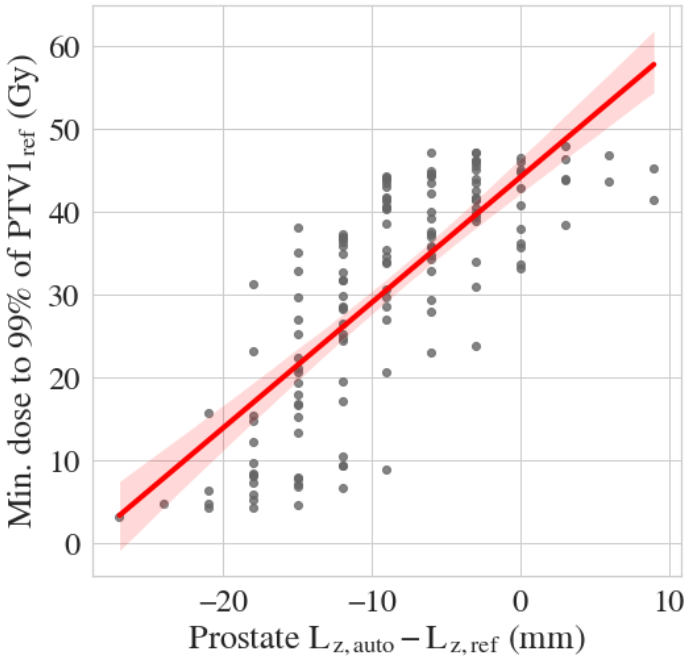


Figure 2: Plot of $D_{99\%}$ assessed on reference PTV1 versus difference in prostate length, along inferior-superior (z) axis, between auto- and reference contours. Shaded area show 95% confidence interval calculated via bootstrap (n=1000). Pearson correlation coefficient r=0.77.

The effect of prostate size was investigated in more detail with respect to the numerical value of the PTV goal $D_{99\%,PTV1,ref}$ (minimum dose to 99% of reference PTV1). Figure 2 shows a plot of this goal vs. difference in prostate length between auto-contoured and reference structures, along the inferior-superior (z) axis. Strong correlation is observed with Pearson correlation coefficient r = 0.77. Equivalent correlations in the x and y directions were considerably lower at r = 0.23 and r = 0.33, as were correlations with all geometric measures displayed in Table 1. These results imply that a shortening in the z-length of prostates was a major determinant of reduced target coverage in auto-contoured plans. This may be explained by the geometry of the single-arc treatment plan, which results in considerably steeper dose gradients along the gantry rotation axis (z) compared to the x-y plane where intensity modulation principally occurs.

5 Conclusion

This work has shown that a simple scripted approach can be applied to assess dosimetric consequences of implementing auto-contouring systems in radiotherapy. Geometric measures that consider treatment plan geometry may be more useful in assessing clinically-relevant dosimetric changes than traditional overlap or distance metrics. Further work to examine how optimisation approach (e.g. scripted class solutions, AI or full manual replanning) affects the discoverability of systematic changes is warranted.

References

[1] Njeh C. Tumor delineation: The weakest link in the search for accuracy in radiotherapy. J Med Phys 2008;33:136. https://doi.org/10.4103/0971-6203.44472

[2] Sharp G, Fritscher KD, Pekar V, et al. Vision 20/20: Perspectives on automated image segmentation for radiotherapy. Med Phys 2014;41:050902. https://doi.org/10.1118/1.4871620

[3] Hindocha S, Zucker K, Jena R, et al. Artificial Intelligence for Radiotherapy Auto-Contouring: Current Use, Perceptions of and Barriers to Implementation. Clinical Oncology 2023;35:219–26. https://doi.org/10.1016/j.clon.2023.01.014

[4] NICE Guidance: Artificial intelligence technologies to aid contouring for radiotherapy treatment planning: early value assessment. National Institute for Clinical Excellence. 2023. Available online: https://www.nice.org.uk/guidance/hte11

[5] Mackay K, Bernstein D, Glocker B,et al. A Review of the Metrics Used to Assess Auto-Contouring Systems in Radiotherapy. Clinical Oncology 2023;35:354–69. https://doi.org/10.1016/j.clon.2023.01.016

[6] Koo J, Caudell J, Latifi K, et al. Essentially unedited deep-learning-based OARs are suitable for rigorous oropharyngeal and laryngeal cancer treatment planning. J Applied Clin Med Phys 2023. https://doi.org/10.1002/acm2.14202

[7] Pullen H, Harrison K. 2023 https://github.com/scikit-rt/scikit-rt

[8] Jena R, Kirkby NF, Burton KE, et al. A novel algorithm for the morphometric assessment of radiotherapy treatment planning volumes. BJR 2010;83:44–51. https://doi.org/10.1259/bjr/27674581

[9] Dearnaley D, Syndikus I, Mossop H, et al. Conventional versus hypofractionated high-dose intensity-modulated radiotherapy for prostate cancer: 5-year outcomes of the randomised, non-inferiority, phase 3 CHHiP trial. The Lancet Oncology 2016;17:1047–60. https://doi.org/10.1016/s1470-2045(16)30102-4

Implementing a simple, dynamic and controllable AI delineation tool in clinical settings

Nis Sarup[1], Maximilian Lukas Konrad[1], Marie Louise Madsen Lange Sevdal[1], Simon Long Krogh[1], Christian Rønn Hansen[1,2], Irene Hazell[1], Ebbe Laugaard Lorenzen[1], Carsten Brink[1]

[1]Laboratory of Radiation Physics, Department of Oncology, Odense University Hospital, Odense, Denmark
[2]Department of Clinical Research, University of Southern Denmark, Odense, Denmark

Abstract This paper describes some of the development, decision-making processes, and outcomes of implementing a locally hosted service for automatic segmentation, using in-house trained models complying with national and international guidelines, of medical scans in cancer radiation therapy. By integrating this service, we significantly reduce the time from scanning to treatment planning, as the system automates tasks traditionally performed manually, thereby reducing the need for manual editing. We detail the service's architectural structure, discussing critical decisions in its design and addressing the constraints posed by the hospital's broader IT infrastructure. Our results highlight that automatic segmentation improves workload, and sheds light on practical considerations and challenges of embedding such technological solutions in a clinical setting.

1 Introduction

Artificial intelligence (AI) in radiotherapy, especially in auto-segmentation, has shown promising advancements in increasing consistency in segmentations compared to manual segmentations[3]. Yet, the practical application of AI in clinical settings lags behind.

To address this, we developed the AI Inference Service (AIS), a system designed to facilitate the use of AI for clinical auto-segmentation in radiotherapy in our department, using AI models tailored to our needs. AIS focuses on robustness, ease of use, and adaptability to various clinical scenarios.

The AIS operates transparently for the end-user, see Figure 1. In a typical workflow, scanners send images directly to the treatment planning system. With AIS, scanners transmit images simultaneously to the treatment planning system and AIS.

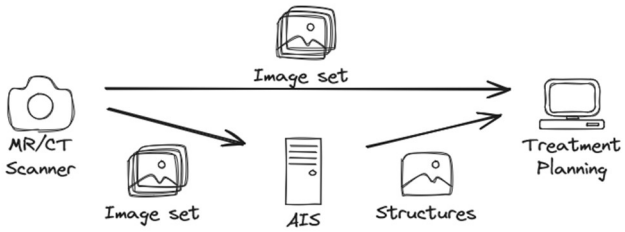


Figure 1: The workflow of the AIS as seen by the end-user.

Modular Service Structure

AIS is composed of four main modules, see Figure 2, each serving a distinct function within the service:

- Gateway: As an extension to a DICOM database, it filters incoming images, determining their suitability for segmentation and the appropriate AI model for processing.
- Watchdog: This module monitors a directory for new image sets, triggering AI inference processes. It ensures timely execution of AI scripts and limits the system to only have as many active inference processes as the available computational resources allow.
- Inference: Manages the conversion from DICOM to Nifti format, runs the nnU-Net[1] for segmentation, post processing, and converts the output back to a DICOM RTStruct file.
- Return service: Tracks both successful and unsuccessful inferences, managing the output RTStructs accordingly. It returns completed segmentations or error-coded RTStructs for failed processes. Currently, both the DICOM protocol and standard file copying are supported.

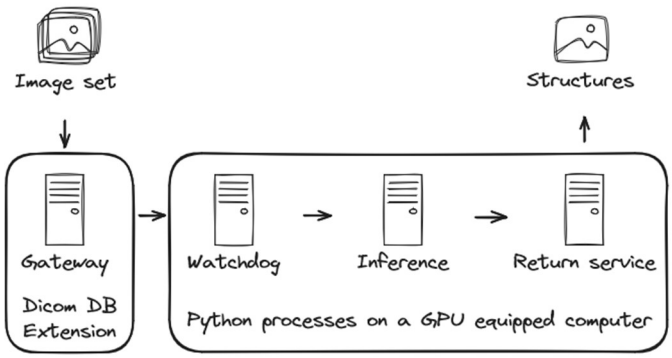


Figure 2: Modular structure of AIS.

Each delineation of an image set is allocated a unique directory on the inference machine, where all files relevant to that image set are stored. The system renames this directory based on the process stage, facilitating easy tracking and logging. In the event of successful segmentation and delivery, the directory is deleted. Otherwise, it is preserved with logs to aid in error identification and resolution.

2 Materials and Methods

System Architecture and Technical Specifications

The AIS is engineered to integrate into the existing IT infrastructure of our radiotherapy department. Image sets are sent directly to the normal recipient systems, while simultaneously being sent to AIS. The system's software components are primarily written in Python and Matlab, chosen for their robustness and compatibility with medical

imaging tools and the AI models produced using the nnU-net[1] architecture.

Ensuring reproducible results

After a model has been trained and validated, a hash for the files containing the trained model weights is calculated. This hash is used to ensure that an image set that is expected to be segmented using a specific model is indeed segmented using that model.

A separate service recomputes the hash for the installed models each night, so any change in the installed models will trigger an error.

Modular Design for Enhanced Functionality

To make the system robust, the four modules have clearly defined and non-overlapping responsibilities. As an image set progresses through the system, only one module handles the data at any given time.

Configurations

The Gateway has a configuration for where it should send files for inference.

Additionally, it has several trigger-configuration files. Each of these has:

- A set of DICOM tags and values that will trigger an inference process.
- Information on where to send the results.
- Information specifying custom changes for the structures needed by the clinicians.
- A "niceness" level. Image sets with lower "niceness" will be processed before image sets with higher values, if multiple image sets are waiting to be processed.
- The hash of the expected model to ensure correct model is used for inference.
- Whether or not to include a structure with metadata on the inference.

To enable the reading and, potentially, changing of these files by technical staff, they are plain-text files with an easy syntax.

We added the “niceness” level to allow for prioritizing inferences, e.g., for online treatments that need the delineations as fast as possible.

The Watchdog module configuration consists of:

- Where it should scan for incoming image sets, and at what interval.
- The command for inference.

The Inference process is configured with:

- System specific nnU-net information
- Structure information, nnU-net arguments for each enabled model

The Return Service mimics the Watchdog service and is configured with:

- Where it should scan for incoming image sets, and at what interval.
- When to clean up image sets that have encountered an error.

Process Management

The Watchdog, Inference and Return service are written in Python 3.11. To ensure robustness, they all implement a main process, which spawns a worker process that does the actual work. The main process monitors and restarts the worker if a crash or any other unrecoverable error happens.

Workflow

In AIS, each image set is processed in a unique, dedicated directory on the inference machine. This directory acts as a central node for all files related to the inference, including configurations, logs, and outputs. The system employs a dynamic directory renaming strategy to reflect the current stage of the inference process, enhancing transparency and trackability. Upon successful completion of the segmentation, the directory and its contents is automatically deleted to maintain system efficiency. In contrast, any failures in the process result in the retention of the directory, along with detailed logs, to facilitate error analysis and resolution.

The naming of the directories follows the format of "STATE_UID" where "STATE" is one of the following: *receiving*, *ready*, *active*, *inferred*, *error* or *handled_error*, and UID is a unique ID for that image set. Each module will only work on a directory in a subset of the possible states, and the renaming happens as the absolute last step of each module.

The representation of state via directory naming was chosen for its simplicity and ease of understanding. This approach ensures that even individuals with limited familiarity with AIS can easily understand and navigate the system's status.

Integration with Clinical Workflow

AIS's design emphasizes minimal disruption to the existing clinical workflow. Automated processes and streamlined data handling ensure that clinicians can access AI-aided segmentations without deviations from standard operational procedures. The system's adaptability allows for easy incorporation of future advancements in AI segmentation algorithms and models.

Communicating metadata to the end-user

AIS employs a simple approach to address the challenge of effectively communicating inference process details to the end-user, especially considering the limitations in DICOM for such metadata transmission. We utilize an empty structure in the generated DICOM-file, which is named to convey information about the AI segmentation process

despite being devoid of segmentation data. This naming convention serves two purposes:

- **Successful Inference**: In cases of successful AI segmentations, the empty struct's name includes the AI model used and the number of DICOM images used for inference. This provides end-users with accessible information about segmentations they are reviewing.
- **Unsuccessful Inference**: If an error occurs during segmentation, the empty structure is named with an error code and a concise description of the issue. This approach allows for rapid identification and understanding of errors, facilitating efficient troubleshooting and resolution without overwhelming the user with technical details.

Including the image count in the metadata is a critical feature, designed to address scenarios where an incomplete image set might be processed due to DICOM transfer failures. By detailing the number of slices in the metadata, AIS empowers users to identify and address such discrepancies quickly, ensuring the integrity of the treatment planning process.

Furthermore, AIS defaults to appending "_AI" to the names of all AI-generated structures. This simple yet effective naming convention allows clinicians to distinguish between manually created and AI-generated structures, thus maintaining clarity in treatment planning and review processes. This feature is configurable and naming can be changed per structure in individual workflows.

Image set filtering

The filtering in the Gateway module allows AIS to accept all image sets, but will only send image sets matching the filters to be inferred. This ensures that a model will only be used to delineate images of the same type as the images it has been trained and validated with. A model trained on CT-images may not be validated on MR-images and should not be used for those. Likewise, a model trained to delineate the male pelvis, should not be used on images of the female pelvis.

3 Results

AIS Deployment

The deployment of the AI Inference Service (AIS) within our radiotherapy department was a smooth process. The primary adjustments post-deployment involved fine-tuning the system's logging capabilities.

Inference Time Efficiency

Depending on the specific image set and model used, the inference process, which starts when an image set has been received and ends when the delineated structures have been delivered, typically takes between 2 and 3.5 minutes, even on consumer-grade GPUs like the NVIDIA 3090. This duration demonstrates the system's efficiency in handling inference within a reasonable timeframe, fitting well within the clinical workflow.

User Feedback and Workflow Impact

The reception of AIS among clinicians has been positive, even though the system is relatively new in production. While formal data collection and analysis is on-going but not yet finalized, initial user feedback suggests an estimated 30-50% increase in workflow efficiency when using AIS compared to not using AIS.

Challenges in Metadata Communication

A challenge encountered involves the transmission of metadata to end-users. Some treatment planning systems do not support the import of empty structures, hindering the intended communication method. We included a minimal geometric structure within the 'empty' structure to address this. This workaround ensures the conveyance of metadata.

4 Discussion

Extensibility of AIS with External Models

A key aspect of the AIS is its flexibility in integrating AI models. Currently, AIS provides delineation using four models:

- Male Pelvis Model[2].
- Head and Neck model[5].
- Heart model[4].
- Brachytherapy model for cervical cancer.

However, its architecture is designed for easy extension to include externally developed models as well. This adaptability significantly broadens the potential applications of AIS.

Windows-Based System: Constraints and Benefits

Operating AIS on Windows does impose certain constraints, particularly in terms of model compatibility, as nnU-net version 1 models do not run on Windows. Despite this, we believe the benefits, notably the access to comprehensive support from the hospital's larger IT system, outweigh the limitations. This support is crucial for maintaining system stability and reliability, which are paramount in a clinical setting.

DICOM database as a Potential Bottleneck

The integration of part of AIS as an extension in a DICOM database, while beneficial for data handling and processing, has been identified as a bottleneck in terms of total inference time. This may be a hindrance for using AIS in time-critical workflows. Exploring the possibility of excluding the DICOM database from the process could yield improvements in inference speed, enhancing patient experience and treatment efficiency.

Research AIS
A non-clinical variant of AIS, characterized by more flexible or entirely absent filters, is used by researchers and developers to test and validate new models before clinical implementation.

Future Directions
In addition to technical advancements, a significant focus is on optimizing AIS for faster inference times. This is especially critical in real-time treatment scenarios, where prompt decision-making is needed.

5 Conclusion
The implementation of the AIS in our department marks a significant step. AIS has demonstrated its potential to streamline workflows, enhance treatment planning efficiency, and provide high-quality auto-segmentations, proving to be a valuable asset in our radiotherapy treatments.

Key Achievements and Impact:
- AIS's smooth deployment and stable performance highlight its reliability in a clinical setting.
- The system's ability to process image sets efficiently, even with consumer-grade GPUs, illustrates its practicality and scalability.
- Positive feedback from clinicians, suggesting a 50% increase in workflow efficiency, underscores the system's effectiveness and user-friendliness.

References
[1] Isensee, F., Jaeger, P.F., Kohl, S.A.A. et al. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat Methods 18, 203–211 (2021). DOI: 10.1038/s41592-020-01008-z.

[2] Lorenzen EL, Celik B, Sarup N, Dysager L, Christiansen RL, Bertelsen AS, Bernchou U, Agergaard SN, Konrad ML, Brink C, Mahmood F, Schytte T and Nyborg CJ (2023) An open-source nnU-net algorithm for automatic segmentation of MRI scans in the male pelvis for adaptive radiotherapy. Front. Oncol. 13:1285725. doi: 10.3389/fonc.2023.1285725.

[3] Nielsen, Camilla Panduro, et al. "Consistency in contouring of organs at risk by artificial intelligence vs oncologists in head and neck cancer patients." Acta Oncologica 62.11 (2023): 1418-1425. DOI: 10.1080/0284186X.2023.2256958

[4] Olloni A, Lorenzen EL, Jeppesen SS, Diederichsen A, Finnegan R, Hoffmann L, Kristiansen C, Knap M, Milo MLH, Møller DS, Pøhl M, Persson G, Sand HMB, Sarup N, Thing RS, Brink C, Schytte T. An open source auto-segmentation algorithm for delineating heart and substructures - Development and validation within a multicenter lung cancer cohort. Radiother Oncol. 2023 Dec 19;191:110065. doi: 10.1016/j.radonc.2023.110065. Epub ahead of print. PMID: 38122851.

[5] Lorenzen, EL, Zukauskaite, R, Kyndt, M, Eriksen, JG, Sarup, N, Johansen, J, Maare, C, Primdahl, H, Bratland, Å, Kristensen, CA, Andersen, M, Overgaard, J, Brink, C & Hansen, CR 2023, 'First results on DAHANCA automatic segmentation algorithms of organs at risk', Radiotherapy and Oncology, bind 182, nr. Suppl. 1, OC-0118, s. S79-S80. DOI: 10.1016/S0167-8140(23)08532-8

A robust auto-segmentation approach to accelerate adaptive MRI-guided radiotherapy for pancreatic cancer patients

Mehdi Shojaei[1], Björn Eiben[1], Jamie McClelland[2], Simeon Nill[1], Alex Dunlop[1], and Uwe Oelfke[1]

[1]Joint Department of Physics, The Institute of Cancer Research and The Royal Marsden NHS Foundation Trust, London, UK
[2]Centre for Medical Image Computing and The Wellcome/EPSRC Centre for Interventional and Surgical Sciences, Department of Medical Physics and Biomedical Engineering, University College London, London, UK

Abstract
Pancreatic cancer is challenging to treat with radiotherapy due to anatomical changes during different fractions and its proximity to critical organs at risk (OARs). MR-Linacs facilitate treatment plan adaptation to the current patient anatomy based on daily session MR images. Target and OAR contours on the session images are a prerequisite for daily plan creation and currently involve labor-intensive manual contouring. Therefore, auto-contouring is crucial for an online adaptive MR-Linac workflow. One major challenge in training a robust auto-segmentation model is the availability of the substantial amount of data required. This study investigates an auto-segmentation approach using nnU-Net, in conjunction with generative model augmentations to improve performance. Balanced 3DVane images from 9 patients (7 train/2 validation) from multiple fractions were used to train our auto-segmentation model. An abdominal CT dataset was utilized for generative model inference. Compared to conventional augmentation, the generative network-based augmentation increased the average Dice score from 0.86 to 0.88 and reduced the average surface distance from 4.46mm to 2.78mm across 8 OARs. Our method can be integrated into an MR-Linac adaptive workflow to address time-intensive manual contouring.

1 Introduction

Pancreatic cancer is one of the most challenging cancers in oncology, marked by its aggressive nature and often limited treatment options. Radiotherapy plays a pivotal role as a treatment strategy for pancreatic cancer. However, the effective delivery of a sufficiently high radiation dose is challenging, due to the proximity of critical organs, such as the duodenum and stomach, which are highly sensitive to radiation-induced toxicity.
Recent advances like MR-Linac machines may increase the precision of radiation treatment delivery by enabling MRI visualization before/during treatment for target validation and healthy tissue sparing with their excellent soft-tissue contrast. MR-Linacs can acquire session images before each fraction to manually adapt contours to anatomical changes. Offline manual contouring of all OARs and the radiation target can take up to 3-4 hours; however, dedicating this time before each fraction, in an online workflow, is impractical. Therefore, image registration is employed to propagate structures from pre-treatment CT onto the session MR images, and then, re-contouring is selectively performed on specific areas, facilitating adaptation to changes. Hence, in the online workflow, the manual re-contouring process is streamlined to only 10-20 minutes and focuses solely on structures within the margin of the Planning Target Volume plus 2 cm (PTV+2cm), leading to inaccuracies beyond this margin.
Developing reliable auto-segmentation methods to speed up this contouring process is therefore highly desirable to improve efficiency in the MR-Linac workflow [1]. Various artificial intelligence (AI) methods with different architectures and techniques have been introduced to improve segmentation performance. Although auto-segmentation algorithms have seen significant advancements, unresolved challenges remain due to complexities from tumor and organ size, shape, and position variations, contributing to segmentation inaccuracies that impact radiotherapy precision [2]. Additionally, the diverse range of imaging modalities and contrasts necessitates the development of specific models tailored to each. In our investigation, we selected a set of MR images acquired using the 3DVane technique (Philips Co.) from MR-Linac machines. While these images offer robustness to motion [3], they may lack optimal soft-tissue contrast due to this trade-off. To the best of our knowledge, no previous studies have explored multi-organ auto-contouring for 3DVane images.
Limitations also exist regarding data availability for training an auto-segmentation model, particularly with scarce datasets of 3DVanes which pose challenges for training a deep-learning model. Additionally, in adaptation, only organs within the PTV+2cm margin are manually re-contoured, so outside this region instead of a robust ground-truth only propagated CT contours exist. Data augmentation has become crucial to address ground-truth data scarcity, allowing for the best use of the available data [4]. Nonetheless, conventional data augmentation methods show limited impact as they are usually applied globally or unrealistically.
Altogether, these pitfalls underscore the need for more robust and automated segmentation techniques and approaches with a prior realistic data augmentation pipeline. Therefore this study investigates to which extent an advanced, deep-learning-based augmentation method can enhance a previously published auto-segmentation approach.

2 Materials and Methods

We employed 3D nnU-Net [5], a U-Net-based framework with automated configuration and data augmentation for med-

ical image segmentation to investigate multi-organ segmentation in abdominal MR images. Following the preprocessing phase, this study bifurcated into two approaches. The first approach implements conventional augmentations, such as rotation, scaling, and using various filters. The second approach utilizes generative adversarial networks (GAN) and image registration.

2.1 Data Selection

The dataset for the auto-segmentation model contained various MR/CT images from the MOMENTUM study [6], but only patients who also had 3DVane images were used as they are the preferred acquisition technique at our institute. The selection of image contrast depends on clinicians' preferences. However, due to the availability of more images with balanced contrast, we chose balanced 3DVanes (b3DVanes). Hence, b3DVane session MR and pre-treatment CT images with structures from 9 patients (5-6 fractions each, 50 total) over multiple sessions were obtained.
After consultation with clinicians and physicists at our institutes and reference to prior research [7, 8], the dataset was consolidated into 8 structures, which include the duodenum, liver, left kidney, right kidney, spleen, stomach, spinal cord, and small bowel.
Another dataset, a publicly available pediatric CT dataset with expert contours for 29 structures and 325 images, was utilized to synthetically generate b3DVane images via generative networks. This dataset was selected because it contained all 8 structures [9].

2.2 Image Preprocessing

First, foreground organ regions were cropped to remove background objects. Some of the labels had overlaps. Overlapping labels were separated by assigning each voxel the label of its closest structure boundary, performed via a SciPy (v1.9.3) distance transform providing voxel-wise Euclidean distances. To reduce intensity inhomogeneities, artifacts, and noise in MR images, N4 bias field correction [10] and anisotropic diffusion filtering [11] were applied using SimpleITK (Ver. 2.1.1.2). The parameters for these tasks were taken from Arabi et al [12]. After basic preprocessing, the images/labels were augmented to account for data scarcity.

2.3 Augmentation Approaches

2.3.1 Approach I - Conventional Augmentation

The nnU-Net model uses spatial and intensity-based augmentations including rotation, scaling, noise, blurring, brightness/contrast changes, simulated low resolution, and gamma augmentation. These are applied randomly with varying magnitudes to introduce realistic variations and improve generalizability. In this study, these are referred to as conventional augmentation applied before training the nnU-Net.

2.3.2 Approach II - Synthetic MR Images

This approach was implemented to synthesize realistic MR images from CT images with ground-truth contours.
Registration - To enhance alignment and facilitate training convergence for generating realistic medical images, particularly since the initial image count was insufficient, initial registration of CT and MR images was performed. The registration was executed in two sequential steps. First, an affine registration for global alignment and then a B-spline deformable registration to account for local deformations were implemented using the robust Elastix toolbox integrated with the SimpleITK Python library [13].
The Model Configuration - After registration, images were locally deformed using the Platipy library (Ver. 0.6.0), increasing the initial 50 images to 244. Then, a cycleGAN featuring a patchGAN discriminator [14] was implemented. The input image was resized to 360x360x128 voxels, with a voxel spacing of 0.78x0.78x1.5 mm^3. Both CT and MR images underwent histogram normalization, and then intensity normalization (Z-score normalization) was performed using the 1st and 99th percentiles, scaling values within this range from -1 to 1.
Loss Functions - The generator and discriminator loss functions were calculated using mean squared error (MSE). The cycle consistency loss, which measures the error between the reconstructed and the original images, was computed using a weighted average of one minus the Structural Similarity Index Measure (SSIM) and the Mean Absolute Error (MAE). The identity loss, which measures how well the model preserves the input image when cycled through the generators, was also calculated using MAE.
Optimizer - Model optimization was performed using the Adam algorithm, with a learning rate of 10^{-4} and weight decay of 10^{-5} to regularize and prevent overfitting. To dynamically adjust the learning rate during training, a Cosine Annealing Learning Rate Scheduler was employed [15].
Training - The training process took approximately 5 days on a 24Gb NVIDIA RTX A5000 GPU.
Inference - After training the model, in the inference step, the pediatric abdominal CT images were employed to generate synthetic b3DVane images. As a result, 325 images were synthesized.

2.4 Training the nnU-Net

For approach I, the train and validation sets consisted of images from 7 and 2 patients respectively, with multiple sessions per patient. Training took approximately 16 hours. For approach II, the model was first trained on 325 images over 16 hours, and then the weights were updated by retraining on the 7 train/2 validation split for an additional 4 hours. All computations were performed on a 24Gb NVIDIA RTX GPU.

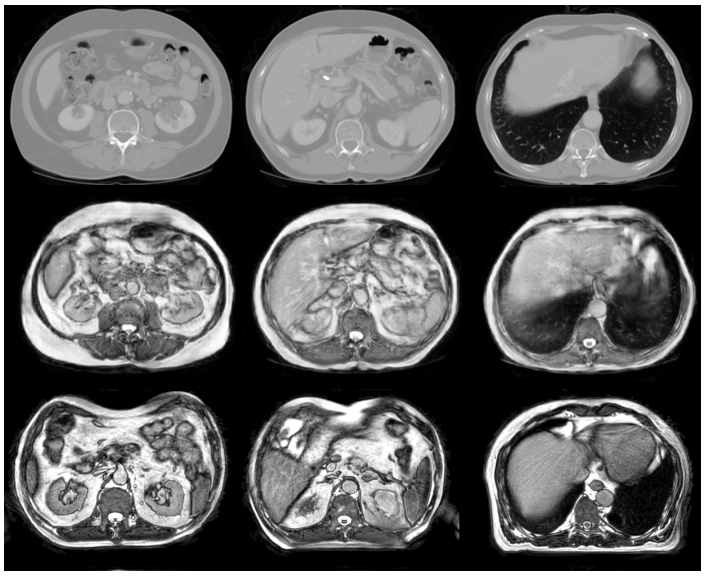


Figure 1: Pediatric CT images (top), generated b3DVane images (middle), and real b3DVane images (bottom).

3 Results

3.1 Image Synthesis

The cycleGAN employed in this study produced 325 b3DVane images resembling MR-Linac output. An example set of resulting images was verified by a clinician regarding its realism of appearance (Figure 1).

3.2 Auto-segmentation

Figure 2 compares auto-contour results between both approaches for the patient with a median Dice score. The model showed the highest Dice performance for the liver and the lowest for the small bowel with both approaches. However, in terms of Average Surface Distance (ASD), the spinal cord had the lowest error and the small bowel the highest. Figure 3 shows the case with the lowest Dice score versus the propagated ground-truth structures, highlighting some ground-truth issues by comparing prediction alignment with organ margins. Table 1 presents the auto-segmentation results for the 2 validation patients over multiple sessions, comparing the augmentation approaches. Our approach, from preprocessing to prediction, required approximately 30 seconds per patient on average.

4 Discussion

Precision radiotherapy delivery for pancreas cancer faces challenges due to anatomical changes that can occur throughout a treatment course. Adaptive radiotherapy workflows aim to account for these variations but are resource-intensive due to the time-consuming manual segmentation required on each treatment day. Overcoming these barriers is crucial to enable the full potential of adaptive radiotherapy.
This study investigated using 3D nnU-Net on 50 3DVane images from 9 patients over multiple sessions, improving performance via our customized auto-segmentation model to reduce manual delineation workload. The initial generative model training without prior registration and localized deformations with Platipy failed to converge due to insufficient training data. Altogether, 325 synthetic b3DVane images were generated from pediatric abdominal CTs with expert contours to address ground-truth inaccuracies and provide increased training data to improve model generalization. Our approach was significantly faster than 10-20 minute partial manual re-contouring for adaptation, without sacrificing accuracy while being comparable to inter-observer variation [16] despite 3DVane's lower soft-tissue contrast. As our model only auto-contours OARs, the target still needs clinician delineation. However, when integrated into an online adaptive workflow, we still expect substantial time-savings. One limitation was the availability of abdominal CTs with OAR contours, confining selection to pediatric CTs which were the only suitable contoured dataset. Our customized approach demonstrated that with only 9 patients, it is possible to produce good segmentation results compared to previous studies on abdominal multi-organ auto-segmentation even when using images with lower soft-tissue contrast and less available data [7]. However, to validate the generalizability of our model, testing on a larger set of patients is required.

Segmented Organs	ASD (mm)		Dice	
	Approach I	Approach II	Approach I	Approach II
Liver	1.45	**1.45**	0.93	**0.96**
Duodenum	3.03	**2.79**	**0.78**	**0.78**
L. Kidney	1.17	**1.05**	**0.93**	**0.93**
R. Kidney	1.18	**1.03**	0.92	**0.93**
Small Bowel	19.03	**8.85**	0.72	**0.75**
Spinal Cord	**0.66**	0.73	0.88	**0.90**
Spleen	1.39	**1.22**	0.91	**0.93**
Stomach	9.20	**4.86**	0.79	**0.82**
Mean	4.46	**2.78**	0.86	**0.88**

Table 1: Comparison between approach I and approach II. The highest Dice scores and lowest ASDs are highlighted.

5 Conclusion

Our approach coupled with nnU-Net demonstrated promising results even with low-contrast images from only a few patients and can be integrated with MR-Linac adaptive workflows to address time-intensive and inaccurate marginally-limited manual contouring.

Acknowledgements

We gratefully acknowledge the data provided by MOMENTUM (MR-Linac Consortium) as well as funding and support from CRUK (C33589/A28284 and C33589/A19727) and the NIHR Biomedical Research Centre at the RMH and ICR.

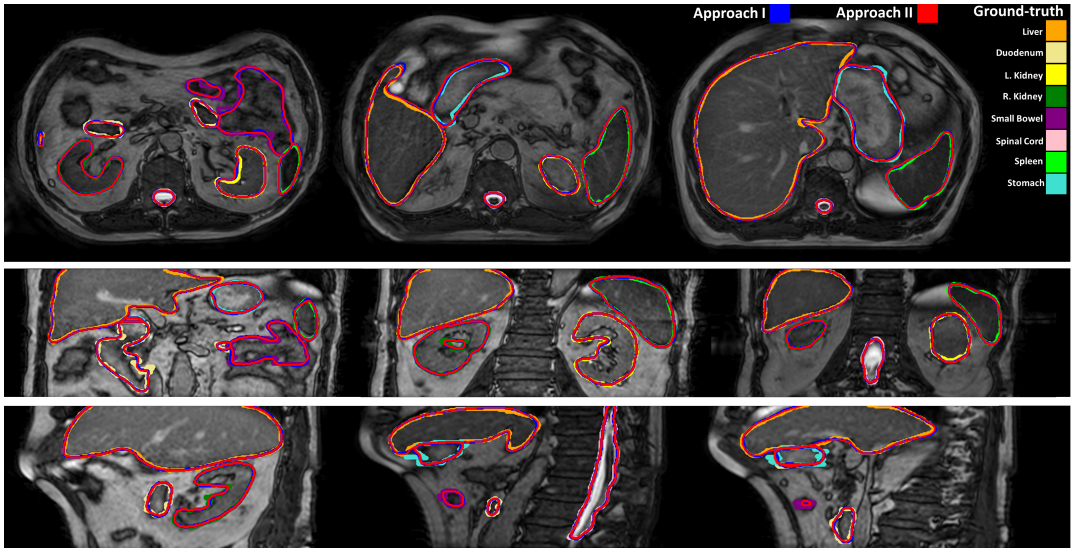


Figure 2: Visual comparison between the 2 approaches on axial (top row), coronal (middle row), and sagittal (bottom row) views. The ground-truth contours are shown in different colors other than blue (approach I) and red (approach II). Approach II, the GAN-based approach, covers the ground-truth contours across organs slightly better than approach I.

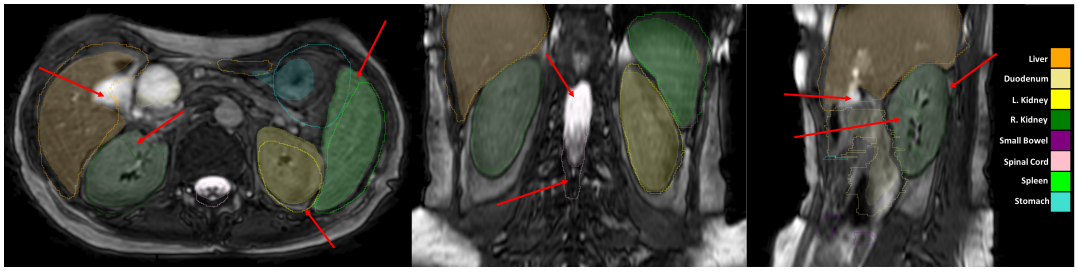


Figure 3: Comparison of model predictions (filled colored shapes) to ground-truth contours (contour lines). Red arrows indicate that the model predictions align better with the organs in the MR image than the propagated CT ground-truth contours.

References

[1] R. Zukauskaite, C. N. Rumley, C. R. Hansen, et al. “Delineation uncertainties of tumour volumes on MRI of head and neck cancer patients”. *ctRO* 36 (2022), pp. 121–126.

[2] H. Ghorpade, J. Jagtap, S. Patil, et al. “Automatic Segmentation of Pancreas and Pancreatic Tumor: A Review of a Decade of Research”. *IEEE Access* (2023).

[3] H. Chandarana, T. K. Block, A. B. Rosenkrantz, et al. “Free-breathing radial 3D fat-suppressed T1-weighted gradient echo sequence: a viable alternative for contrast-enhanced liver imaging in patients unable to suspend respiration”. *Invest Radiol.* 46.10 (2011), pp. 648–653.

[4] Z. Fabian, R. Heckel, and M. Soltanolkotabi. “Data augmentation for deep learning based accelerated MRI reconstruction with limited data”. *ICML*. PMLR. 2021, pp. 3057–3067.

[5] F. Isensee, J. Petersen, A. Klein, et al. “nnu-net: Self-adapting framework for u-net-based medical image segmentation”. *arXiv preprint arXiv:1809.10486* (2018).

[6] S. R. de Mol van Otterloo, J. P. Christodouleas, E. L. Blezer, et al. “The MOMENTUM study: an international registry for the evidence-based introduction of MR-guided adaptive therapy”. *Frontiers in Oncology* 10 (2020), p. 1328.

[7] Y. Chen, D. Ruan, J. Xiao, et al. “Fully automated multiorgan segmentation in abdominal magnetic resonance imaging with deep neural networks”. *J. Med. Phys.* 47.10 (2020), pp. 4971–4982.

[8] J. Lukovic, L. Henke, C. Gani, et al. “MRI-based upper abdominal organs-at-risk atlas for radiation oncology”. *IJROBP* 106.4 (2020), pp. 743–753.

[9] P. Jordan, P. M. Adamson, V. Bhattbhatt, et al. “Pediatric chest-abdomen-pelvis and abdomen-pelvis CT images with expert organ contours”. *J. Med. Phys.* 49.5 (2022), pp. 3523–3528.

[10] N. J. Tustison, B. B. Avants, P. A. Cook, et al. “N4ITK: improved N3 bias correction”. *IEEE Trans. Med. Imaging* 29.6 (2010), pp. 1310–1320.

[11] P. Perona and J. Malik. “Scale-space and edge detection using anisotropic diffusion”. *IEEE PAMI* 12.7 (1990), pp. 629–639.

[12] H. Arabi and H. Zaidi. “One registration multi-atlas-based pseudo-CT generation for attenuation correction in PET/MRI”. *EJNMMI* 43 (2016), pp. 2021–2035.

[13] S. Klein, M. Staring, K. Murphy, et al. “Elastix: a toolbox for intensity-based medical image registration”. *IEEE Trans. Med. Imaging.* 29.1 (2009), pp. 196–205.

[14] J.-Y. Zhu, T. Park, P. Isola, et al. “Unpaired image-to-image translation using cycle-consistent adversarial networks”. *Proceedings of the IEEE ICCV*. 2017, pp. 2223–2232.

[15] I. Loshchilov and F. Hutter. “Stochastic gradient descent with warm restarts”. *Proceedings of the 5th Int. Conf. Learning Representations*, pp. 1–16.

[16] G. Chlebus, H. Meine, S. Thoduka, et al. “Reducing inter-observer variability and interaction time of MR liver volumetry by combining automatic CNN-based liver segmentation and manual corrections”. *PloS one* 14.5 (2019), e0217228.

A national infrastructure for sharing AI segmentation tools

Simon Long Krogh[1], Ruta Zukauskaite[2], Ebbe Lorenzen[1], Nis Sarup[1], Maximilian Konrad[1], Janne Nørlykke Drudgaard[1], Mette Klüver-Kristensen[1], Eva Samsøe[3], Mohammad Farhadi[3], Thomas Hinz-Berg Johansen[3], Carsten Brink[1], Christian Rønn Hansen[1,4,5]

[1]Laboratory of Radiation Physics, Department of Oncology, Odense University Hospital, Odense, Denmark
[2] Department of Oncology, Odense University Hospital, Odense, Denmark
[3] Department of Oncology, Radiotherapy, Zealand University Hospital, Naestved, Denmark
[4] Danish Centre for Particle Therapy, Aarhus, Denmark
[5] Department of Clinical Research, University of Southern Denmark, Odense, Denmark

Abstract *Introduction*: The development of new, often AI-based, automatic contouring solutions has escalated rapidly in recent years. Implementing these new technologies requires resources and specialised skills from the local staff, limiting access to them for smaller or resource sparse sites. This work explores a standardised infrastructure for sharing automated delineation technologies among centres. The infrastructure should be so easy to integrate that any site can access it, broadening the field of sites gaining the advantages of the technologies. *Materials and methods*: The proposed solution is based on the national RT dose plan storage infrastructure, DcmCollab, making an nnUnet-based head and neck OAR model available to a smaller RT centre. *Results*: A set of extensions to the DcmCollab system were implemented, and the test setup was evaluated over the course of seven months, processing 116 patients. The participating clinicians reported a potential time saving of 30%-40%. *Conclusion*: The proposed infrastructure was successfully implemented and is proven to be a viable means of sharing an automated delineation solution. Before a full clinical integration, the legal aspects must be considered, especially MDR compliance in the EU.

1 Introduction

In recent years, great advances have been made in the development of automated processes to improve the consistency and speed of routine clinical workflows in radiotherapy (RT). Many of these rely on artificial intelligence (AI) based contouring of medical image sets. Depending strongly on the treatment site, manual delineation for one RT treatment will typically require between 30 minutes and several hours per delination set [1,2]. This process can be sped up significantly by supplementing it with novel, automated solutions [1,2], and thus releasing precious workforce resources for other tasks.

However, institutions with many resources can often invest more in implementing new technologies for their workflows than those with sparse resources. This can have the unfortunate effect of excluding the clinics that need the automated solutions the most from gaining their benefits.

In this work, we explore a standardised national infrastructure allowing any centre with an automated delineation model to share this implementation with other clinics. The goal is that the standardisation of the infrastructure, along with the centralisation of the automated functions, will increase the ease of implementation of modern automation and, presumably, the homogeneity of the delineations and the conformance to national guidelines across centres as exemplified by Nielsen et al [3].

2 Materials and Methods

The national RT dose plan storage infrastructure, DcmCollab, serves as a basis for this work, providing both the DICOM network connections between all RT centres in Denmark and the mechanisms for handling the specific DICOM tags that trigger the processing of submitted data [4,5].

The automated delineation model used is an nnUNet-based Head&Neck (H&N) organ at risk (OAR) model [6] supplemented with locally developed pre- and postprocessors whose main functions are converting data from DICOM to NIfTI and back.

After implementing the necessary modules to support the proposed infrastructure, a testing agreement was made with an RT institution with limited resources for devoloping AI models locally. Given the limitation in available clinical resources at the testing site, the degree of detailed reporting and documentation possible was limited to subjective evaluations and rough estimates of potential timesaving in clinical workflows.

3 Results

An extension of the DcmCollab system was necessary for the proposed functionality. DcmCollab already supports and often uses the concept of features in DICOM data triggering functions within the system. These triggering features were limited to delineation structures in RTSTRUCT datasets with zero contours and a pre-defined description tag. Using these trigger structures, users can include the submitted data in a research protocol defined in DcmCollab, or trigger the automatic forwarding of a data set to a selected external destination. In order to support sharing automated delineation solutions, these mechanisms needed to be expanded with the following features: (1) Triggering should be possible on any combination of tags in the submitted DICOM data. Modality (0008,0060) and series description (0008,103E) were chosen for the test setup. For each shared delineation solution, a sufficiently specific set of identifying DICOM

tags must be defined to ensure that only the intended datasets are processed. (2) The creation of a derived trigger to handle the data returned from the third-party system should be added to the set of possible triggered actions. This will allow DcmCollab to recognize data returned from the AI delineation inference machine and forward it to the originating centre. (3) It should be possible to define a trigger to be used only once, both to avoid the risk of unexpected triggered actions and to minimise the number of active triggers the system needs to check for in all submitted data. This type of trigger is called a "one-shot trigger" henceforth.

Two new services were added to the DcmCollab system servers to handle the necessary new functions: One to assess submitted image data and estimate whether all data in a series was received. All non-image data, eg. RTSTRUCT, RTDOSE, and RTPLAN, is assumed to be complete as soon as it is received and analysed. The second service checks for and acts upon triggering features in the submitted data.

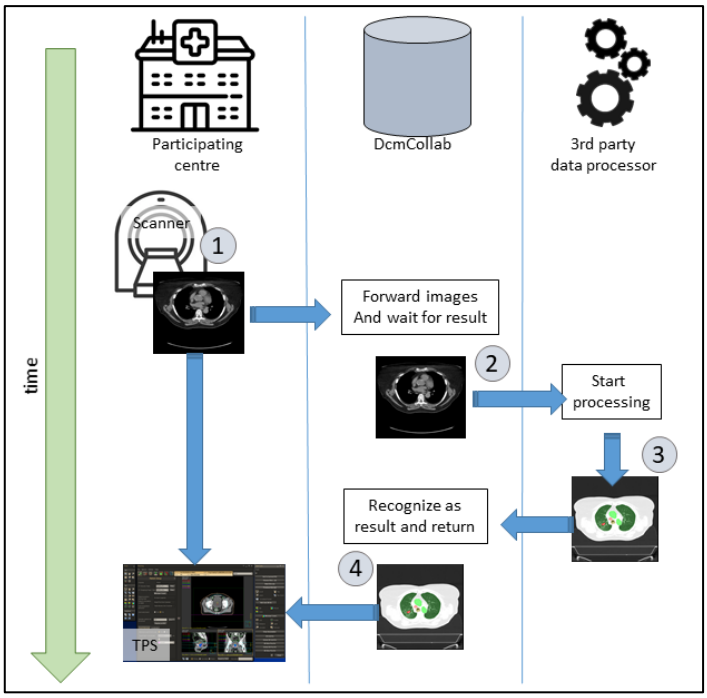


Figure 1 - The data flow of the proposed infrastructure: (1) An image set with a pre-defined text string in its DICOM Series Description is generated at the participating centre and exported both to DcmCollab and to the local treatment planning system. (2) The image set is received by DcmCollab and its triggering DICOM features are recognised. DcmCollab then prepares to receive the result from the 3rd party data processor by creating a derived trigger and forwards the image set to it. (3) The 3rd party data processor creates the AI-generated delineations in a DICOM RTSTRUCT SOP instance and transmits it back to DcmCollab. (4) DcmCollab recognises the returned data with its derived trigger and forwards it to the treatment planning system at the participating centre.

The final workflow is visualised in Figure 1. At the scanner in the participating centre, the DICOM image series has a pre-defined text string encoded in its Series Description DICOM tag and is sent using DICOM communication to both the DcmCollab system and the local treatment planning system (TPS). At DcmCollab, the image data is received. Once the received data is recognised as complete and containing the triggering features, the image series is forwarded to the third-party data processor - in the test case, an nnUnet inference machine. Furthermore, a derived one-shot trigger is created to handle the resulting delineations once they are returned from the inference machine. The one-shot trigger responds to a DICOM modality of RTSTRUCT and the Referenced Frame Of Reference Sequence (3006,0010) ‣ Referenced Study Sequence (3006,0012) ‣ Referenced Series Sequence (3006,0014) tag corresponding to the Series Instance UID (0020,000E) tag of the original image set. The latter tag indicates that the received structure set is generated from the submitted image set. The inference machine is configured to export its results to DcmCollab, where the one-shot trigger initiates the final export of the generated delineations to the submitting centre. After processing the one-shot trigger, DcmCollab disables it to avoid the set of active triggering features that must be checked for each received dataset to grow uninhibited.

In practice, the time elapsed from an image set is submitted from the scanner until the resulting delineations are received at the local TPS is in the range of 10-15 minutes, with the larger part being spent by DcmCollab assessing the completeness of the submitted image set. This duration is close to the minimum time passing from an image set is acquired until manual delineation is performed in the current clinical workflow.

The tested workflow was made available via the DcmCollab infrastructure in May of 2023 and has been used for 116 CT image sets from the participating institution. As mentioned, the participating institution did not have the resources for detailed individual registration of potential time saving for each patient. An overall subjective assessment was provided after four months of testing. The H&N team estimated 30%-40% potential time saving in OAR delineation with the implemented model. This measure relies mostly on the shared delineation model which is not he focus of this work. More importantly, the infrastructure used to utilize the model did not inhibit the benefit it introduced. Furthermore, the participating centre reports that they would not have been able to implement a local AI-based automatic delineation setup in the testing period.

4 Discussion

Given the subjectivity of the reported results, this work should mainly be perceived as a feasibility project. It is possible to implement an infrastructure in which a central entity provides participants with automated delineation functions and to embed such a solution in a clinical workflow. Furthermore, there are strong indications that doing so will result in significant timesaving and harmonisation of OAR segmentation. The H&N OAR segmentation model strictly follows the Danish Head and Neck Cancer Groups (DAHANCA) guidelines [7]. The clinical implementation and fine-tuning of the OAR segmentations can be validated using the DcmCollab system by comparing the raw AI segmentation with the clinical OAR segmentation. The process of fine-tuning can identify deficiencies in several aspects: It could be the local centre, which has a historical tradition for correcting specific structures, the AI segmentation model may be insufficiently accurate for specific areas, or it could be due to a need for an update of the national guidelines. In all three cases, the information would be found by comparing the final clinical delineation after manual correction and the raw AI segmentations. These observations are not specific to the H&N treatment sites, and it is not bold to assume that similar ones could be made for other RT treatment sites.

This work is strictly technical, and has not focused on the legal and licensing issues that should be addressed if an infrastructure of this nature is implemented in the clinical routine. The nnU-net framework is open-source, and the model used in the test setup described was developed locally, so in this specific case, licensing is trivial. If proprietary solutions such as MIM [8] or Siemens Healthineers Autocontouring [9] were to be shared, proper licensing would become a more complicated and costly issue to solve.

Legally, the solutions provided and the infrastructure supporting them should adhere to the European Medical Device Regulations (MDR) or similar. The provided solution should be tested and documented by the providing instance and the site using it. To minimise the resource cost in testing and documenting an implementation of an externally hosted solution in a clinic, it would be advisable to create a set of standard documents covering the majority of models and sites with a minimum of adjustments. These templates would then be provided to each participating centre along with access to the corresponding automated solutions.

To ensure traceability, each model made available should undergo strict version control, while each product of a model should be clearly marked with information about both that they are indeed the product of an automated solution and the name as well as the version of the model used.

5 Conclusion

This work has broken the first steps of a path to making cutting-edge solutions more broadly available in RT. Other paths towards the same goal exist; e.g. making the setup and installation process of new solutions as seamless as possible. Unfortunately, history shows that these aspects rarely are a priority when new technologies are developed. Furthermore, easily installed local solutions would not provide the added benefit of nationally homogenised uses of delineation models. Another approach could be a web-based solution, though such an implementation would inevitably be less integrated into an existing clinical workflow. The proposed infrastructure can be introduced with close to zero overhead in terms of human resources at the participating centre, and provides the added advantage of supporting the homogenisation of guideline interpretation and national consensus on which models to use in clinical practice.

To implement the proposed solution fully, some effort should be made to address the legal aspects of providing clinical services to external collaborators. The handling of such aspects has only recently reached maturity, specifically for AI solutions internally in an organisation, so one should expect to spend some resources on expanding this to cross-center collaboration. Finally, licensing must also be considered if proprietary solutions are to be shared in the framework.

In this work, a single AI delineation model was implemented to prove the efficacy of the concept. In larger scale use, considerations regarding computer resources should be made; should each model run on their own hardware, or should hardware be shared and access to it be managed by a queuing system? No investigation into these issues was made in the current context.

References

[1] Kosmin M, Ledsam J, Romera-Paredes B, et al. Rapid advances in auto-segmentation of organs at risk and target volumes in head and neck cancer. Radiother Oncol. 2019 Jun;135:130-140. DOI:10.1016/j.radonc.2019.03.004

[2] van der Veen J, Willems S, Deschuymer S, et al. Benefits of deep learning for delineation of organs at risk in head and neck cancer. Radiother Oncol. 2019 Sep;138:68-74. DOI:10.1016/j.radonc.2019.05.010

[3] Nielsen CP, Lorenzen EL, Jensen K, et al. Consistency in contouring of organs at risk by artificial intelligence vs oncologists in head and neck cancer patients. Acta Oncol. 2023 Nov;62(11):1418-1425. DOI:10.1080/0284186X.2023.2256958

[4] Krogh SL, Brink C, Lorenzen EL, et al. A national repository of complete radiotherapy plans: design, Results, and experiences. Acta Oncologica. 2023:1-8. DOI:10.1080/0284186X.2023.2270143

[5] Westberg J, Krogh SL, Brink C, et al. A DICOM based radiotherapy plan database for research collaboration and reporting. J Phys: Conf Ser. 2014;489. DOI:doi:10.1088/1742-6596/489/1/012100

[6] Isensee F, Jaeger PF, Kohl SAA, et al. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat Methods. 2021 Feb;18(2):203-211. DOI:10.1038/s41592-020-01008-z

[7] Jensen K, Friborg J, Hansen CR, et al. The Danish Head and Neck Cancer Group (DAHANCA) 2020 radiotherapy guidelines. Radiother Oncol. 2020 Oct;151:149-151. DOI:10.1016/j.radonc.2020.07.037

[8] Inc. MS. Contour ProtégeAI+ 2024. Available from: https://www.mimsoftware.com/radiation-oncology/contour-protegeai-plus

[9] Healthineers S. Deep learning-based autocontouring. Available from: https://www.siemens-healthineers.com/dk/radiotherapy/software-solutions/autocontouring

Use of complementary multispectral data for improving the accuracy of deep learning solution for bioluminescence tomography reconstruction in glioblastoma tumor models

Behzad Rezaeifar[1,2], Cecile J.A. Wolfs[1], Natasja G. Lieuwes[3], Rianne Biemans[3], Brigitte Reniers[2], Ludwig J. Dubois[3], Frank Verhaegen[1]

[1] Department of Radiation Oncology (Maastro), GROW School for Oncology and Reproduction, Maastricht University Medical Centre+, Maastricht, The Netherlands

[2] Research group NuTeC, Centre for Environmental Sciences, Hasselt University, Diepenbeek, Belgium

[3] The M-Lab, Department of Precision Medicine, GROW - School for Oncology and Reproduction, Maastricht University, Maastricht, The Netherlands

Abstract We previously proposed an AI-based solution for obtaining the location and shape of a glioblastoma tumor in a preclinical rat model using information obtained from the bioluminescence imaging (BLI). The utilization of spectral data in bioluminescence tomography (BLT) reconstruction through conventional model-based algorithms has demonstrated notable advantages. Therefore, in this study we investigate how the spectral data can be employed in the deep-learning approach. For this purpose, a novel multi-channel 3D U-Net deep learning model is proposed, concurrently incorporating bioluminescence skin fluence (BSF) and two additional input channels comprising spectral data obtained by applying specific optical filters to the detected BSF. Results indicate enhanced performance, as evidenced by a 74 ± 12% median Dice Similarity Coefficient (DSC) between the predicted tumor and the ground-truth CT-based tumor contour, compared to the previous full optical spectrum solution's 61%, affirming the efficacy of incorporating spectral data in the deep-learning approach.

1 Introduction

Bioluminescence imaging (BLI) is a powerful tool to monitor tumor growth in vivo in pre-clinical cancer research using luciferase-enabled light-emitting small animal tumor models [1], [2]. In its 2D form, BLI can merely provide relative tumor growth information from the integrated light intensity and an inaccurate 2D radiotherapy target region [3], [4], [5]. However, in order to monitor the tumor's volumetric information more effectively and enable quantitative radiation targeting, a 3D representation of the tumor is required, which can be achieved by reconstructing the 3D bioluminescence tomography (BLT) [6], [7].

In our previous publications, two distinct AI-based methods were introduced for BLI-based tumor targeting [8], [9]. Our first approach was based on predicting the central location of the tumor using a deep convolutional neural network (CNN) and deriving a spherical envelope around the predicted location to encompass the tumor [8]. The next step was accomplished by capturing shape information together with the location of the tumor by using a fully convolutional 3D ResNet model, which enabled 3D tumor mask predictions from the bioluminescence skin fluence (BSF), i.e. the sum of all the BLI images acquired at different camera viewing angles projected on the skin of the animal [9].

This study investigated the added benefit of using multispectral information, i.e. the resulting BLI images acquired with various optical filters, to improve the AI-based method for BLI-based tumor targeting. Utilizing such multispectral information has been shown to be informative in mathematically-driven BLT reconstruction solutions [10]. The utility of multispectral information in BLT reconstruction lies in its capacity to exploit wavelength-dependent variations in photon absorption and scattering, wherein photons of distinct wavelengths exhibit differential magnitudes of scattering and absorption, thereby providing valuable insights into the precise depth of tumor localization within the tissue. Therefore, this study aimed to implement a spectrally-driven deep learning solution for the BLT reconstruction problem that can enhance the quality of the previously proposed models. In order to obtain the multispectral data required for training and validation, in this study, a Monte Carlo simulation (MCS) environment is utilized, as described in our previous publications [8], [9]. MCS are often considered as the gold standard method for photon transport and therefore can be considered a good candidate to build a high quality training database for the AI solution.

2 Materials and Methods

The starting point for this study is a database of 57 contrast-enhanced cone-beam CT (CE-CBCT) scans of orthotopic glioblastoma model in rats, described fully in our previous publication [5]. In this database, hereafter called F98 database, various anatomical structures (tumor, brain, and eyes) in the CT images are manually contoured by expert biologists for brain and tumor contours. Thereafter, a MCS geometry is constructed for each case within the F98 database, excluding the tumors smaller than 10 mm^3, by assigning various optical properties to different tissues in the image according to the literature. Then, an optical light emission according to the emission spectrum of the firefly luciferase is assigned to the tumor and for each voxel inside the tumor 5×10^6 photons are generated.

In this study, photon transport is performed by using the Geant4 application for tomographic emission (GATE)[11]. More specifically, the photons generated from the tumor are tracked using the fluence actor inside previously described voxelized MCS geometry. This resulted in a 3D volumetric image with the same size as the MCS geometry with values equal to number photons passing each voxel. Therefore, to obtain the BSF using the MCS output a skin mask is applied

to remove all the values not visible to the optical camera. Additionally, separate MCS is performed for each spectral portion of the emission spectrum. In other words, due to the implementation issues with the MCS environment, the filtration is not simulated on the detection side, as it is performed in the real BLI measurements by blocking the photons reaching the detector outside the corresponding pass-band wavelength spectrum. In contrast, the optical filter is applied on the tumor emission spectrum, limiting the generated photons to the wavelength range of the detection filter. Although this might seem inaccurate, considering that only elastic scattering of optical photons is modeled in the MCS, which is the dominant type of photon scattering at this wavelength, this will result in the same output as photon filtration upon detection. Therefore, for each case in the MCS database, new MC simulations were performed for each optical filter by blocking different parts of the emission spectrum. Figure 1 represents the optical filters used in this study which match those in our small animal irradiator (X-RAD 225Cx, Precision X-ray Inc., North Branford, CT, USA). Furthermore, the number of photons emitted was kept to the original 5,000,000 photons per unit voxel in the tumor to provide good statistics in the output and to match the original MCS.

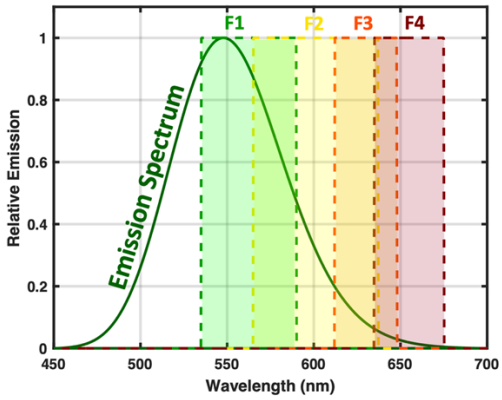


Figure 1: Bioluminescence emission spectrum of firefly luciferase and filters (F1-F4) used for modeling the spectrally driven measurements

Once the new multispectral MCS database was generated for all the samples, a new training database was obtained by stacking the 3D BSFs resulting from the two furthest spectral filters, depicted in **Error! Reference source not found.** as F1 and F4, together with the no-filter original MCS, hereafter called open filter, in the fourth dimension. In other words, the input of the Multispectral network is analogous to an RGB-colored image, which stacks red, green, and blue input channels in the fourth dimension for a 3D volumetric image. Conversely, here the three different spectral channels, each consisting of a 3D BSF image for different wavelength ranges, are stacked in the fourth dimension.

The multispectral MCS database is then utilized as the input of a DL network with a similar architecture to the previously proposed network[9]. The DL network comprises of a 3D U-Net architecture, namely the RatLesNet [12], with a slight modification in the number of input channels to accommodate the multichannel nature of the input data. Furthermore, the hyperparameters of the DL network, including the optimizer and its learning rate, loss function, and number of filters, were automatically optimized, using Optuna [13], to ensure the best optimized hyperparameters for the newly generated input.

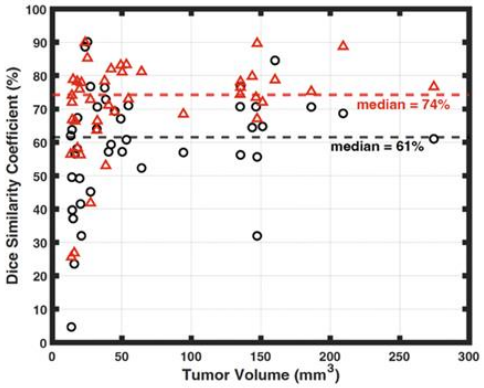


Figure 2: performance evaluation of the proposed multispectral solution (red) compared to the open filter counterpart (black).

3 Results

Training a deep learning solution with the multispectral BSF images increased the performance for most cases in the database over open filter BSF images, presented in our previous publication [9]. The overall performance comparison between the multispectral network and its open filter counterpart is depicted in Figure 2 and further visualized for a few cases in Figure 3. As presented in **Error! Reference source not found.**, the median DSC between the predicted and ground truth tumor contours increased from 61%, for the open filter BSF to 74% for the multispectral BSF. In other words, in 75% of the cases within the database, inclusion of the complementary multispectral information resulted in a better performance compared to the open filter counterpart.

4 Discussion

It has been shown that the complementary information obtained using different optical filters can be used to distinguish the shape variations for the majority of the cases, as presented in Figure 2. This resulted in a 13 percentage point median DSC increase over the entire database. However, it is important to note that memory limitations during AI model training can be a bottleneck for using more and more multispectral information. Hence, in future investigations, optimizing the quantity of complementary data and refining its incorporation into the deep learning model can prove advantageous for

maximizing the utilization of limited hardware capacity in a more efficient manner.

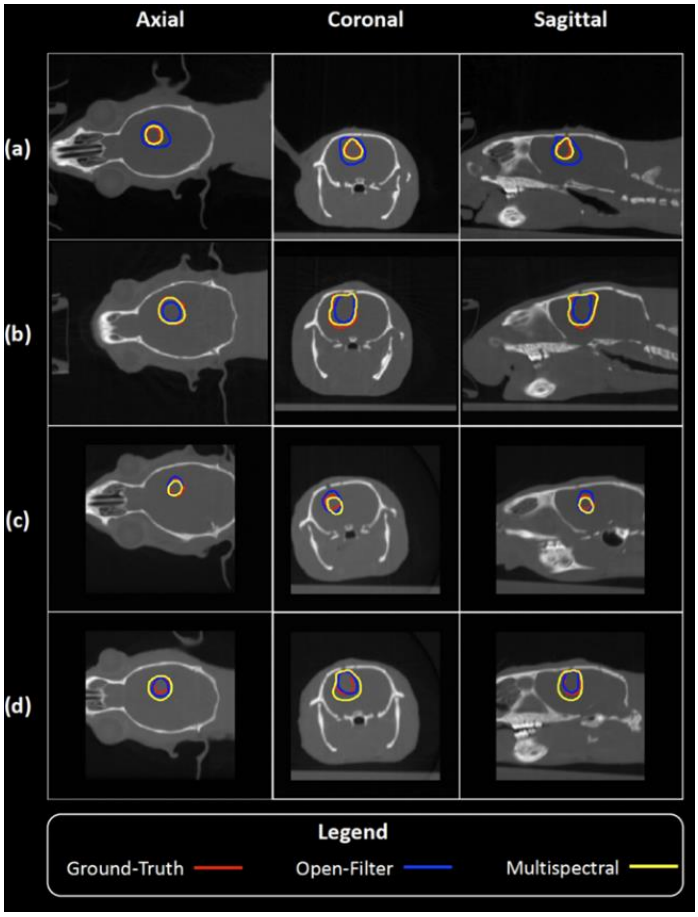


Figure 3: Tumor contour visualization for the multispectral network (yellow) compared to the open filter network (blue) and the ground truth CT-based contour (red): (a,b) present the best cases in the database with improved performance from the added information and (c,d) depict the cases with reduced performance.

The proposed BLI-based solution in our previous publication fails to capture the correct volume for some cases, such as flat tumors located close to the skull. This study provides proof that the multispectral measurements can help alleviate this issue since each spectral bandwidth includes a different portion of the absorption coefficient and can help distinguish smaller flat tumors close to the skull from larger deep-seated tumors. To clarify this further, assume the two different cases where a flat surface tumor and a deeper larger tumor result in approximately the same number of photons reaching the skin. Consequently, these two tumors are considered to have similar volumes just based on the amount of open filter bioluminescence light exiting from the skin. However, the multispectral information holds important details that differentiate the two cases since photons emitted from the larger tumor have to transport in a larger portion of the biological tissue. Therefore such a tumor gains more photon intensity since a different photon spectrum, particularly the near-infrared spectrum, passes through the tissue more easily and reaches the surface due to differential absorption in different wavelengths. This study shows that the provided multi-channel AI model can use this additional information to improve the performance of the BLI-based tumor segmentation.

In future work, a combination of the AI models with physics models of the light propagation, in the form of novel physics-informed AI models, can help to improve the results even further. These models can incorporate the flexibility of AI solutions with the explainability of physics models. Furthermore, such models can alleviate the uncertainties in AI models caused by small-sized databases and perform better with data limitations in mind.

5 Conclusion

This study presents a comprehensive examination of a multi-channel deep learning framework designed to exploit the complementary multispectral characteristics inherent in luciferase emission, with the aim of refining the precision of Bioluminescence Tomography (BLT) reconstruction in the context of an orthotopic glioblastoma rat model. This comparative assessment underscores the superior efficacy of the proposed multi-channel deep learning approach. The results highlight the significance of integrating multispectral data within the luciferase emission spectrum, showcasing its potential to elevate the accuracy and reliability of BLT reconstructions within the specific domain of orthotopic glioblastoma rat models..

References

[1] J. Cosette, R. Ben Abdelwahed, S. Donnou-Triffault, C. Sautès-Fridman, P. Flaud, and S. Fisson, "Bioluminescence-Based Tumor Quantification Method for Monitoring Tumor Progression and Treatment Effects in Mouse Lymphoma Models," *J. Vis. Exp. JoVE*, no. 113, p. 53609, Jul. 2016, doi: 10.3791/53609.

[2] F. Verhaegen, P. Granton, and E. Tryggestad, "Small animal radiotherapy research platforms," *Phys. Med. Biol.*, vol. 56, no. 12, pp. R55-83, Jun. 2011, doi: 10.1088/0031-9155/56/12/R01.

[3] R. A. Weersink *et al.*, "Integration of optical imaging with a small animal irradiator," *Med. Phys.*, vol. 41, no. 10, p. 102701, 2014, doi: 10.1118/1.4894730.

[4] S. Yahyanejad *et al.*, "Complementary use of bioluminescence imaging and contrast-enhanced micro-computed tomography in an orthotopic brain tumor model," *Mol. Imaging*, vol. 13, 2014, doi: 10.2310/7290.2014.00038.

[5] A. M. Mowday *et al.*, "Use of a Luciferase-Expressing Orthotopic Rat Brain Tumor Model to Optimize a Targeted Irradiation Strategy for Efficacy Testing with Temozolomide," *Cancers*, vol. 12, no. 6, Art. no. 6, Jun. 2020, doi: 10.3390/cancers12061585.

[6] S. Rapic, T. Samuel, P. E. Lindsay, S. Ansell, R. A. Weersink, and R. S. DaCosta, "Assessing the Accuracy of Bioluminescence Image-Guided Stereotactic Body Radiation Therapy of Orthotopic Pancreatic Tumors Using a Small Animal Irradiator," *Radiat. Res.*, Feb. 2022, doi: 10.1667/RADE-21-00161.1.

[7] Y. Gao, K. Wang, Y. An, S. Jiang, H. Meng, and J. Tian, "Nonmodel-based bioluminescence tomography using a machine-learning reconstruction strategy," *Optica*, vol. 5, no. 11, pp. 1451–1454, Nov. 2018, doi: 10.1364/OPTICA.5.001451.

[8] B. Rezaeifar *et al.*, "A deep learning and Monte Carlo based framework for bioluminescence imaging center of mass-guided glioblastoma targeting," *Phys. Med. Ampmathsemicolon Biol.*, vol. 67, no. 14, p. 144003, Jul. 2022, doi: 10.1088/1361-6560/ac79f8.

[9] B. Rezaeifar *et al.*, "A deep-learning assisted bioluminescence tomography method to enable radiation targeting in rat glioblastoma," *Phys. Med. Biol.*, 2023, doi: 10.1088/1361-6560/ace308.

[10] H. Dehghani, J. A. Guggenheim, S. L. Taylor, X. Xu, and K. K.-H. Wang, "Quantitative bioluminescence tomography using spectral derivative data," *Biomed. Opt. Express*, vol. 9, no. 9, pp. 4163–4174, Aug. 2018, doi: 10.1364/BOE.9.004163.

[11] V. Cuplov, I. Buvat, F. Pain, and S. Jan, "Extension of the GATE Monte-Carlo simulation package to model bioluminescence and fluorescence imaging," *J. Biomed. Opt.*, vol. 19, no. 2, p. 026004, Feb. 2014, doi: 10.1117/1.JBO.19.2.026004.

[12] J. M. Valverde, A. Shatillo, R. De Feo, O. Gröhn, A. Sierra, and J. Tohka, "RatLesNetv2: A Fully Convolutional Network for Rodent Brain Lesion Segmentation," *Front. Neurosci.*, vol. 14, p. 610239, Dec. 2020, doi: 10.3389/fnins.2020.610239.

[13] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A Next-generation Hyperparameter Optimization Framework," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, in KDD '19. New York, NY, USA: Association for Computing Machinery, Jul. 2019, pp. 2623–2631. doi: 10.1145/3292500.3330701.

Monte Carlo photons simulation from alpha-emitter radionuclides

David Sarrut[1], Ane Etxebeste[1], and Jean-Michel Létang[1]

[1]Université de Lyon, CREATIS; CNRS UMR5220; Inserm U1044; INSA-Lyon; Université Lyon 1; Centre Léon Bérard, France.

Abstract

We present a Monte Carlo virtual source model for generating photons emitted during the decay chain of alpha-emitter radionuclides. The model, integrated into GATE 10, utilizes Geant4 databases to extract photon emission lines from decaying daughters. Employing Bateman equations for decay chain abundances and activities within a specified time range, it generates photons' energy and temporal distribution. By avoiding the need to simulate the entire decay chain, it reduces the computation time compared to reference ion source. The source model can be applied to any alpha-emitter radionuclide in the Geant4 database. It facilitates the investigation of (SPECT) imaging photon emission from alpha-emitters.

1 Introduction

Targeted Alpha Therapy (TAT) is a type of Radiopharmaceutical therapy (RPT) that exploits the short range and high linear energy transfer (LET) of alpha particles (^{4}He) to enhance targeted tissue dosages [1]. Examples of radionuclides for TAT include ^{223}Ra (Xofigo™) for castration-resistant prostate cancer (CRPC) with metastatic bone disease [2] and, more recently, ^{225}Ac (Ac-225-PSMA-617) combined with PSMA for metastatic prostate cancer [3]. Investigation of other alpha emitters such as ^{149}Tb, ^{211}At, ^{212}Pb, ^{213}Bi, ^{223}Ra, and ^{227}Th is underway. The LET of alpha particles, approximately 50-200 keV/µm, is about three orders of magnitude larger [4] than the LET of β^- particles emitted by conventional radionuclides such as ^{177}Lu or ^{90}Y. The alpha particle's range, around 50 to 100 µm, is close to the diameter of a few cells [5], leading to high tumor damage, but also with potential toxicity risks.

The assessment of the dose spatial biodistribution and temporal biokinetics is critical to mitigate these risks. In-vivo imaging is feasible for certain alpha emitters using emitted photons in the decay chain, detectable by SPECT systems [6]. For instance, imaging feasibility has been reported for ^{212}Pb [7] (phantoms) and ^{225}Ac [8] (patients).

However, each radionuclide's decay chain produces distinct photons with different energy lines. SPECT acquisition must be tailored by selecting appropriate energy windowing, scatter correction, and reconstruction methods based on the emitted photons' energy ranges and quantities. For high-energy photons (>400-450 keV), challenging for Anger camera-based SPECT systems, alternative imaging systems like Compton Camera may be considered [9]. Monte Carlo simulation is an important method for investigating SPECT image acquisition parameters and advancing new imaging systems [10, 11]. In this study, we propose an efficient source model generating gamma and X-ray emissions from alpha emitters, encompassing associated energy spectra, intensity, and timing.

2 Materials and Methods

The proposed photon source model is based on two primary components. Firstly, it involves identifying and compiling all photon lines, incorporating their energies and intensities, from decay daughters and processes into a list of energy-intensity pairs. The origin of these photons is attributed to (1) isomeric transitions and (2) atomic relaxation. Bremsstrahlung radiation, depending on the media, is neglected here. Secondly, the model automatically computes and utilizes Time Activity Curves (TAC) to express the relative abundances of decaying daughters as a function of time. During simulation runtime, the Geant4 engine sequentially generates primary photon events, considering their relative intensities, energy levels, and associated timestamps.

Gamma emission data from isomeric transitions (IT) is extracted from the Geant4 database, and relative intensities are computed recursively for all decay channels and energy levels. Initial examination of a radionuclide's decay channels involves referencing the decay table of the `G4RadioactiveDecay` process object. The relevant data, including excitation energies and deexcitation channel details, are stored in files (e.g., `z89a225` for ^{225}Ac). Each deexcitation channel provides gamma transition energy, relative intensity, and the total internal conversion coefficient. Recursive computation of the final gamma emission rate considers branching ratios, gamma emission probabilities, total transition probabilities, and recursively accumulated channel probabilities. This process is performed for common radionuclides (^{149}Tb, ^{211}At, ^{212}Pb, ^{213}Bi, ^{223}Ra, ^{227}Th, etc.) and their decay daughters.

Photons from atomic relaxation (AR) are extracted from the IAEA website using the provided web API[1] (ENSDF Evaluated Nuclear Structure Data Files). Similar to the isomeric transition process, the obtained data are stored in a file, as lists of photon emission energies and associated probabilities. Geant4 incorporates the Bearden database, aligned with the IAEA website and activated using `/process/em/fluoBearden true`. Considering the

[1]API documentation: https://nds.iaea.org/relnsd/vcharthtml/api_v0_guide.html

nanosecond timescale of atomic relaxation decay, characteristic X-ray emission is treated as prompt, excluding delayed fluorescence [12].

The photon lines with their respective energy and probabilities extracted from IT and AR are not sufficient, we also need the relative abundance of each daughter radionuclide. This is done by solving Bateman equations that compute the time-dependent decay of radioactive nuclides and their daughter products, accounting for changes in population over time due to parent nuclide decay and subsequent daughter decay. These equations yield to TAC for all daughters based on specified time points.

In summary, the proposed source model automatically retrieves comprehensive information (daughters, half-life, TAC, etc.) for a given radionuclide. It incorporates relative activity from computed TAC, coupled with emitted photons from isomeric transitions and atomic relaxation, to selectively generate primary photons with accurate energy, intensity, and timing within the specified timeframe.

3 Experiments and results

We compared the photons emitted from a reference Geant4 simulation and the proposed model. The simulation used a ^{225}Ac point-like source of 1000 Bq, in a water box, during an acquisition time frame of 1000 seconds starting 14 hours after injection of the radionuclide. Photons from the reference simulation were extracted from a phase space that contains all particles that have been created and tracked together with the kinetic energy, their Geant4 creator process, and their associated timestamps. At the end of the simulation, the phase space was processed to extract the photons whose creator process is IR or AR and whose timestamps are between 14 hours and 14 hours plus 1000 seconds. The resulting reference photon energy spectrum was compared to the one obtained from a second simulation in the same condition but using the proposed source model. Figure 1 shows very good concordance between most of the photon lines except for the very low energy (<20 keV).

We also simulate a projection from Siemens Intevo Bold SPECT/CT device equipped with High Energy collimator (8000 holes of 4 mm, 2 mm septal thickness) coupled to 9.5 mm thickness NaI crystal. We simulated the NEMA IEC body phantom with six spheres filled with 100 kBq/mL ^{225}Ac (around 0.5 MBq total) and no background activity. The acquisition was one single-angle orientation with a time frame of 210 seconds starting at t=24h like in [8]. We compared three types of sources: (1) the reference ^{225}Ac ion source, (2) a simplified dual gamma source of 218.12 keV (11.4%) and 440.45 keV (25.9%), like the one in [13], and (3) the proposed source model. Figure 2 shows the obtained projections and associated profiles. The proposed source model is in good agreement with the reference profile, while the simplified model depicts larger discrepancies.

Using the reference ion source involved the tracking of a lot of different particles and required to start from t=0, and filter the desired photons. If we consider an acquisition of 200 seconds with a ^{225}Ac source of 3 kBq, about 15.8×10^6 particles will be tracked for only 2.7×10^5 photons that should be simulated (32 seconds vs 1.4 seconds). In the previous experiment, computation time was 28 hours for the reference source, 1.6 hours for the simplified two peaks source and 3.7 hours for the proposed one.

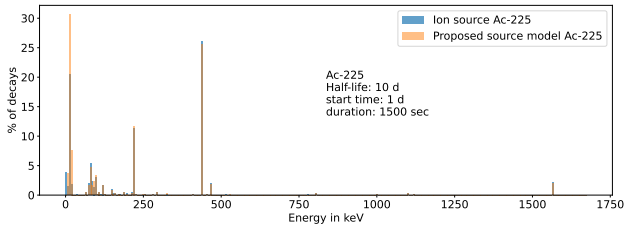


Figure 1: Comparison of emitted photons from the reference ion source and the proposed source model for ^{225}Ac.

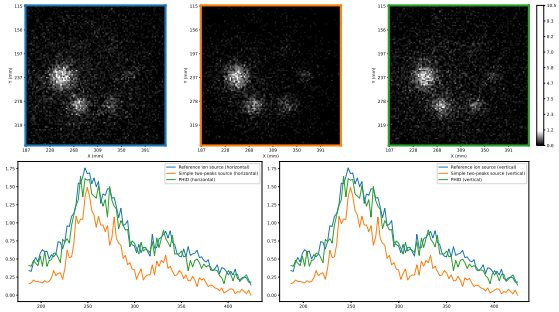


Figure 2: Top: ^{225}Ac SPECT projections of the IEC phantom for an energy window around 218 keV (energy window equal to 196-240 keV), with the reference ion source (left), the simplified two-peaks model (middle), and the proposed source model (right).

4 Conclusion

We presented and evaluated a specialized Monte Carlo source model for alpha emitters, specifically designed for the generation of photons within a specified timeframe. This approach provides computational efficiency benefits compared to conventional ion sources that require the generation and tracking of all decay products. Obtained photons are in good agreement with the reference ones except for the low energy range below 20 keV. The proposed source model is versatile and has been applied to various other alpha emitters. It can be useful for optimizing acquisition protocols, refining scatter correction parameters, and exploring advanced reconstruction methods. The source model is open source and available in GATE 10.

Acknowledgments

This work was performed within the framework of the MOCAMED project (ANR-20-CE45-0025), the SIRIC LYriCAN+ Grant INCa-DGOS-INSERM-ITMO cancer18003. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research. This work was granted access to the HPC resources of IDRIS under the allocation 2019-101203 made by GENCI (Jean Zay computing center).

References

[1] S. Tronchin, J. C. Forster, K. Hickson, et al. "Dosimetry in Targeted Alpha Therapy. A Systematic Review: Current Findings and What Is Needed". *Phys Med Biol* 67.9 (Apr. 2022), 09TR01. DOI: 10.1088/1361-6560/ac5fe0.

[2] P. G. Kluetz, W. Pierce, V. E. Maher, et al. "Radium Ra 223 Dichloride Injection: U.S. Food and Drug Administration Drug Approval Summary". *Clinical Cancer Research* 20.1 (Jan. 2014), pp. 9–14. DOI: 10.1158/1078-0432.CCR-13-2665.

[3] C. Kratochwil, F. Bruchertseifer, F. L. Giesel, et al. "225Ac-PSMA-617 for PSMA-Targeted α-Radiation Therapy of Metastatic Castration-Resistant Prostate Cancer". *Journal of Nuclear Medicine: Official Publication, Society of Nuclear Medicine* 57.12 (Dec. 2016), pp. 1941–1944. DOI: 10.2967/jnumed.116.178673.

[4] A. I. Kassis. "Therapeutic Radionuclides: Biophysical and Radiobiologic Principles." *Seminars in nuclear medicine* 38.5 (Sept. 2008), pp. 358–366. DOI: 10.1053/j.semnuclmed.2008.05.002.

[5] J.-P. Pouget, C. Lozza, E. Deshayes, et al. "Introduction to Radiobiology of Targeted Radionuclide Therapy." *Frontiers in medicine* 2 (2015), p. 12. DOI: 10.3389/fmed.2015.00012.

[6] Y. Seo. "Quantitative Imaging of Alpha-Emitting Therapeutic Radiopharmaceuticals". *Nuclear Medicine and Molecular Imaging* (2019), pp. 1–7. DOI: 10.1007/s13139-019-00589-8.

[7] M. Kvassheim, M.-E. R. Revheim, and C. Stokke. "Quantitative SPECT/CT Imaging of Lead-212: A Phantom Study". *EJNMMI Physics* 9.1 (Aug. 2022), p. 52. DOI: 10.1186/s40658-022-00481-z.

[8] A. Gosewisch, M. Schleske, F. J. Gildehaus, et al. "Image-Based Dosimetry for 225Ac-PSMA-I&T Therapy Using Quantitative SPECT". *European Journal of Nuclear Medicine and Molecular Imaging* 48.4 (Apr. 2021), pp. 1260–1261. DOI: 10.1007/s00259-020-05024-1.

[9] M. Fontana, D. Dauvergne, J. M. Létang, et al. "Compton Camera Study for High Efficiency SPECT and Benchmark with Anger System". *Physics in Medicine & Biology* 62.23 (Nov. 2017), pp. 8794–8812. DOI: 10/ggjtkf.

[10] N. Benabdallah, W. Scheve, N. Dunn, et al. "Practical Considerations for Quantitative Clinical SPECT/CT Imaging of Alpha Particle Emitting Radioisotopes". *Theranostics* 11.20 (Sept. 2021), pp. 9721–9737. DOI: 10.7150/thno.63860.

[11] D. Sarrut, A. Etxebeste, N. Krah, et al. "Modeling Complex Particles Phase Space with GAN for Monte Carlo SPECT Simulations: A Proof of Concept". *Physics in Medicine and Biology* 66.5 (2021). DOI: 10.1088/1361-6560/abde9a.

[12] G. F. Knoll. *Radiation Detection and Measurement*. 4th Edition. Hoboken, N.J: John Wiley & Sons, Sept. 2010.

[13] T. Lee, M. Kim, W. Lee, et al. "Performance Evaluation of a Compton SPECT Imager for Determining the Position and Distribution of 225Ac in Targeted Alpha Therapy: A Monte Carlo Simulation Based Phantom Study". *Applied Radiation and Isotopes* 154 (Dec. 2019), p. 108893. DOI: 10.1016/j.apradiso.2019.108893.

A full GATE/Geant4 model of a helical radiotherapy device

Sheraz Blas[1], Luc Simon[1,2]

[1]Centre de Recherches en Cancérologie de Toulouse, UMR1037 INSERM, Université Toulouse III, ERL5294 CNRS Oncopole, 31037 Toulouse, France

[2]Department of Medical Physics, Oncopole Claudius Regaud Institut Universitaire du Cancer de Toulouse, 31059 Toulouse, France

Abstract Radiation treatment relies on Treatment Planning Systems (TPS) for precise dose calculations. However, challenging cases, like strong heterogeneities or complex target movements, needs reference calculations such as Monte Carlo model. This study introduces a Monte Carlo model of a helical radiotherapy device (Accuray Radixact). GATE platform v9.1 simulates the linac head using data from the manufacturer. The model is validated by comparing measured and simulated dose distributions. A distinctive feature is its consideration of all accelerator element movements including individual Leaf Open Time. The model will find application in diverse contexts.

1 Introduction

In France, External Beam Radiotherapy (EBRT) is used to treat more than 60% of cancer patients (1). In routine, the absorbed dose is calculated by commercial algorithms that are fast and accurate and that give satisfactory results in most clinical cases (2). However, in complex situations, these algorithms, based on simplifications, show their limits. In these cases (inorganic material, dose to skin, out of beam doses) Monte-Carlo methods are considered as the reference computation (3). They can also provide basic data of dose detectors (e.g. output correction factors or OCF) for the determination of absolute dose (4). These methods cannot be implemented in clinical routine for several reasons: complexity of the tools (such as Geant4 (5)), long time computation and a lack of certification.

After have proposed a full open-source model of the TrueBeam linear accelerator (Varian, Palo Alto, CA) (6) based on GATE (7) and showed its potential applications (8) our team is implementing a model of an helical linac i.e. a Tomotherapy unit (Radixact, Accuray, Madison, WI). The complexity of this modeling in GATE/GEANT4, is the way that the dose is dynamically delivered by the Radixact Unit. The patient couch is moving continuously, while the gantry is rotating and, above all, during each control points on the helical-shape trajectory of the gantry, each of the 64 leaves of the Multi-leaf Collimator (MLC) has a very short (<300-400 ms) and unique Leaf Open Time (LOT). The leaves are fully open or fully close on these machines (with a very short opening/closing time).

This present work describes the implementation of the model, including the source tuning, the geometry, and the management of the time slices to model the MLC motion. Then some validation results based on a comparison to measurements are presented, including basic measurements in water for an open-field fixed beam gantry but also results for a fully dynamic irradiation based on a plan provided by the real machine (i.e. DICOM RT Plan). A confidentiality agreement between the manufacturer and our team is signed. Thus, we cannot provide exact figures of the geometry in this article.

2 Materials and Methods

A. General modelling

A comprehensive modeling of the accelerator head was conducted, considering the electron beam, target, primary collimator, jaws, multi-leaf collimator (MLC), and the associated movements of each component. All the elements contained in the head of accelerator are shown in the Figure 2. All complex shapes of the linac head (MLC, jaws, etc.) are reproduced using Blender. The produced .stl file is accepted by GATE.

B. Monte Carlo Model

The model was developed using the GATE Monte Carlo open-source software (version 9.1) associated with Geant4 (version 10.7.2). The physics list is the standard option 3. Cut-off ranges equivalent to an energy of 1 keV for photons and 10 keV for e−/e+ were applied. The dose calculation was performed using the regional high-performance computing, Olympe. Simulations was computed on 6 nodes, each with more than 5×10^9particles, allowing to keep statistical error under 1%

Source

To implement the device, the electron beam was directed towards a tungsten target that has been designed, based on relevant scientific literature (9,10) and adjusted to reproduce Percentage Depth Dose (PDD) profile in water (see hereafter).

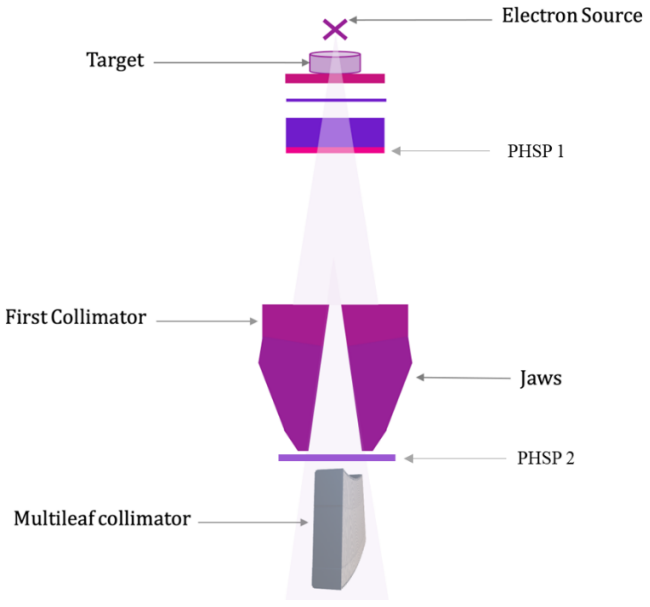


Figure 1 – Schematic representation of the elements of the head

Collimation system

The energy of the electron beam was represented by a Gaussian distribution centered at 6 ± 0.5 MeV energy. To reduce simulation time, intermediate phase spaces files (in format .IAEAphsp) were used. Two phase space files were stored: the first immediately following the target (to avoid simulating the interactions in the target at each simulation) and the second after the jaws (one for each of the three possible jaws opening).

The collimation system included a primary collimator, a set of jaws and a multileaf collimator. Three longitudinal distances can be adjusted using the jaws: 1, 2.5 and 5 cm.

The multi-leaf collimator is composed of 64 binaries leaves, which allows to modulate the field in transverse direction. When all the leaves are open, the field opening is 40 cm at the isocenter.

Dynamics fields

In GATE, the movement is generated with an update of geometries, using *time slices*. To model the modulation, plans are generated with the Treatment Planning System (TPS) and were exported as DICOM RTPLAN object. A locally developed Python code is used to retrieve all the data concerning the movement (gantry, jaws, couch and MLC) and stored in *placement files* in the GATE format. For gantry rotation, couch translation and for each moving leaf, one *placement file* was created. Figure 2 shows the principle of the different modeled time slices. For each CPI, a first *time slice* (corresponding to the first geometry) is generated, with the duration of the shortest LOT. Thus, a second time slice is added with this first leaf closed, and a duration corresponding to the difference between the two shortest LOT, and so on. Finally, the number of created time slices corresponds to the number of open leaves of the CPI. For Radixact unit, the MLC has 64 leaves and there are 51 CPI/rotation. The number of generated time slices may be considerable.

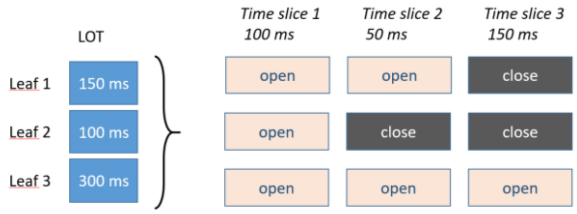


Figure 2 – Principle of the creation of time slices for a single CPI (with only 3 leaves in this example).

C. Validation of the system

Several validation steps were achieved. To assess the radiation source and, by extension, the geometry of the collimator jaws, depth dose distributions and dose profiles in a water tank were realized. For the three field sizes, simulations and measurements were conducted on depth dose distributions and dose profiles. The measurements were performed using the A1SL ionization chamber within a water phantom. PDD, longitudinal (LPD) and transverse (TPD) dose profiles were measured. The GATE simulations were computed under the same conditions with a size of voxels of 1.25^3 mm^3. The dose ratio D20/D10 was calculated on PDD for both measurements and computation. Attention was given to full width at half-maximum (FWHM) and penumbra on LPD, while the TPD was assessed using the width at 25%.

The geometry of the MLC was verified with an simple MLC static pattern (alternating sequences of 4 open and 4 closed leaves). Computation results were compared to a measurement with a radiochromic film. In the context of dynamic field simulations, model validation was established using a TPS plan. A gamma index was employed to compare the dose distributions between the simulation and the TPS.

3 Results

A. Static Mode

Measurements and GATE results of PDD, LPD and TPD in water are represented in Figure 3.

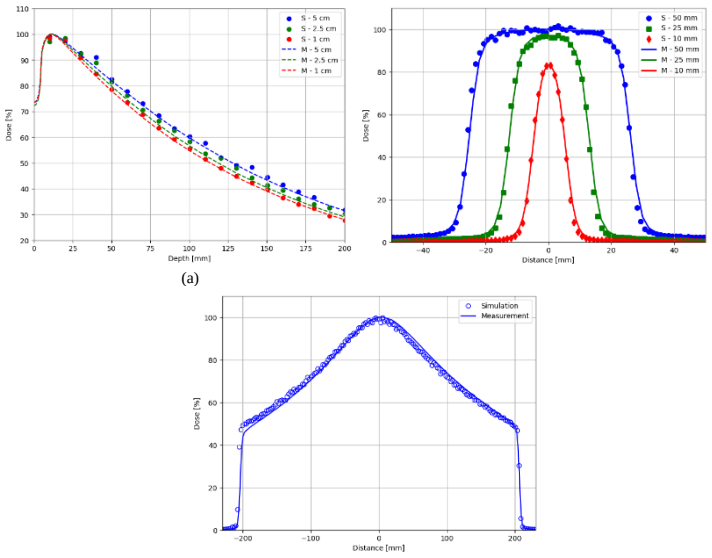


Figure 3 – PDDs (a) LPDs at 1.5 cm depths (b) and TPDs (c) at 1.5 cm depth obtained by simulations and measurements. Curves are normalized at maximum dose (PDD) and on central axis of the largest opening field (LPD, TPD)

Using 2%/2mm gamma index analysis, the passing rate for PDD ranged between 96.35%-99.6% between simulation and measurement curves. D20/D10 for measurements and GATE were 0.527 and 0.531 respectively (<0.7% difference). For LPD, the passing rate with the same criteria (2%/2mm) ranged between 98.14-99.9%. The difference between the values of the half-height width and the penumbra values from the simulation and the measurement was less than 1 mm. For TPD, gamma rates were over 95% and the differences between the values of the width at 25% of the dose are below 2 mm for the three fields.

For the static MLC pattern the comparison between simulation and measurement is shown in Figure 4.

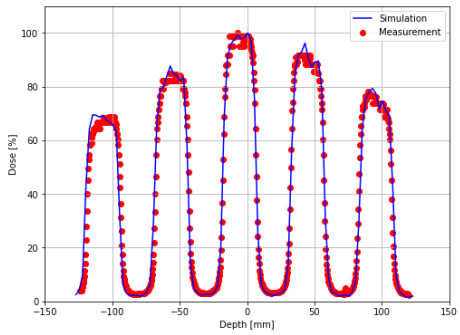


Figure 4 - Comparison of an alternating 4 open/closed leaves between GATE and Gafchromic film measurements

The gamma index pass rate between the simulation and measurement value (3%/2mm) was 95.15% for this static MLC pattern confirming the consistency of leaves modeling.

B. Dynamic acquisition

A plan was developed to compare the dose distributions from the simulation and those determined by the Treatment Planning System (TPS).

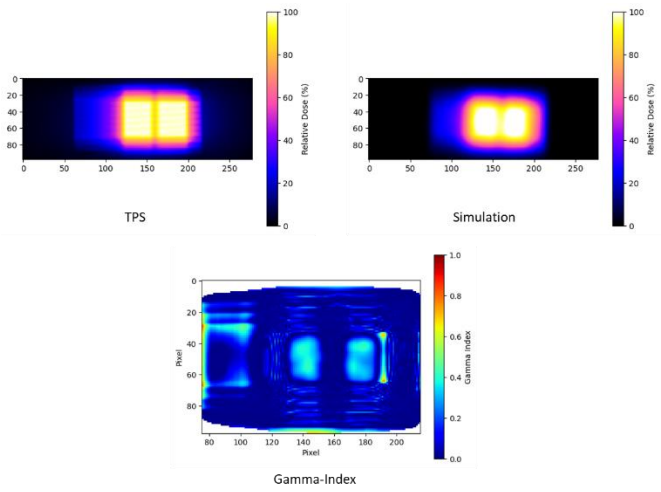


Figure 5 - Dose distribution of a dynamic plan, comparison between TPS and simulation with gamma-index

The plans obtained with both the simulation (smooth image) and the TPS are shown in the following Figure 5. The gamma index at 3%/3mm has been calculated between the two dose images.

4 Discussion

The aim of this work was to validate our GATE/Geant4 model of the Accuray Radixact unit. Other Monte-Carlo models of older Tomotherapy units (e.g. Tomo HD) were published, in particular based on Penelope code (11) or MCNP (9). However, this is the first model of the RADIXACT, to our knowledge, that is Geant4-based and that models each individual LOT. To achieve this goal, the GATE simulation is divided in a large number of time slices (up to 10k).

The first step of the validation was to find the optimal target geometry to obtain the Radixact energy spectrum. The final spectrum was slightly different than previously published model of Tomo HD spectrum (10) (not presented here). But the simulated PDDs in water were consistent with the measurements (in particular the D20/D10 index). Moreover, the consistency of LPD/TPD and of the static MLC pattern allow validating the general geometry of the model. Finally, the dynamic simulations were achieved to verify that all the system movements (couch, gantry, jaws, MLC), extracted from RT Plan file, were correctly reproduced using our *time slices* generator. Further, more complex plans with clinical cases will need to be used later on.

The next step of this project is to use the model for different needs in the field of dose computation: presence of inorganic materials in the patient body, determination of the dose in the regions where TPS results are poor (skin, out-of-field zone), or e.g. to determine OCF of different dose detectors. As these simulations are still very computing time consuming, we are currently working on the validation of time-reduced simulations with some tools to reduce the noise or retrieve the signal from our data.

5 Conclusion

The aim of this study was to obtain a full GATE model of a helical radiotherapy device. All the elements of the accelerator were modeled and validated. In the future, the model will be used for several applications such as reference computation beyond the limits of commercial TPS.

References

1. Chiffres clés de l'observatoire national de la radiothérapie entre 2017 et 2021 - Radiothérapie [Internet]. [cited 2024 Feb 4]. Available from: https://www.e-cancer.fr/Professionnels-de-sante/Les-traitements/Radiotherapie/Chiffres-cles-de-la-radiotherapie
2. Chetty IJ, Rosu M, Kessler ML, Fraass BA, Ten Haken RK, Kong FM (Spring), et al. Reporting and analyzing statistical uncertainties in Monte Carlo-based treatment planning. Int J Radiat Oncol Biol Phys. 2006;65(4):1249–59.
3. Rogers DWO. Fifty years of Monte Carlo simulations for medical physics*. Phys Med Biol [Internet]. 2006 Jun 20 [cited 2024 Feb 4];51(13):R287. Available from: https://iopscience.iop.org/article/10.1088/0031-9155/51/13/R17
4. Palmans H, Andreo P, Huq MS, Seuntjens J, Christaki KE, Meghzifene A. Dosimetry of small static fields used in external photon beam radiotherapy: Summary of TRS-483, the IAEA–AAPM international Code of Practice for reference and relative dose determination. Med Phys [Internet]. 2018 Nov 1 [cited 2024 Feb 4];45(11):e1123–45. Available from: https://onlinelibrary.wiley.com/doi/full/10.1002/mp.13208

5. Agostinelli S, Allison J, Amako K, Apostolakis J, Araujo H, Arce P, et al. Geant4—a simulation toolkit. Nucl Instrum Methods Phys Res A. 2003 Jul 1;506(3):250–303.
6. Leste J, Younes T, Chauvin M, Franceries X, Delbaere A, Vieillevigne L, et al. Technical note: GAMMORA, a free, open-source, and validated GATE-based model for Monte-Carlo simulations of the Varian TrueBeam. Physica Medica. 2021 Sep 1;89:211–8.
7. Sarrut D, Bardiès M, Boussion N, Freud N, Jan S, Létang JM, et al. A review of the use and potential of the GATE Monte Carlo simulation code for radiation therapy and dosimetry applications. Med Phys. 2014;41(6).
8. J L, I M, FX A, R F, X F, M B, et al. A study of the interplay effect for VMAT SBRT using a four-axes motion phantom. J Appl Clin Med Phys [Internet]. 2020 Aug 1 [cited 2021 Oct 6];21(8):208–15. Available from: https://pubmed.ncbi.nlm.nih.gov/32573908/
9. Robert Jeraj, Thomas R Mackie, John Balog, Gustavo Olivera, Dave Pearson, Jeff Kapatoes, Ken Ruchala PR. Radiation characteristics of helical tomotherapy. Med Phys [Internet]. 2004; Available from: https://aapm.onlinelibrary.wiley.com/doi/abs/10.1118/1.1639148
10. Sterpin E, Salvat F, Cravens R, Ruchala K, Olivera GH, Vynckier S. Monte Carlo simulation of helical tomotherapy with PENELOPE. Phys Med Biol. 2008;53(8):2161–80.
11. Sterpin E, Chen Y, Chen Q, Lu W, Mackie TR, Vynckier S. Monte Carlo-based simulation of dynamic jaws tomotherapy. Med Phys [Internet]. 2011;38(9):5230–8. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21978067

Impact of Temporal structure of Electron, Proton and Photon on (H2O2) Concentration using Montecarlo Simulation toolkit (TOPAS-nBio)

Chaoui Mustapha[1] Bouhali Othmane[3], and Tayalati Yahya[1,2]

1Department of Physics, University of Mohammed V Rabat, Rabat, Morocco
2Department of Applied Physics, University of Mohammed 6 P Ben guerir, Morocco
3Department of Physics, A&M University at Qatar, Doha, Qatar

Abstract:

FLASH radiation therapy is an innovative modality that relies on the delivery of ultra-high dose rates (UHDR), threshold of 40 Gy/s in a pulsed beam structure. This modality reduces normal tissue complications without compromising tumor control, termed as "FLASH effect". However, the mechanism governing this effect remains elusive, prompting ongoing debates on the plausible mechanisms. This effect might depend upon the early physical-chemical events and biochemical events. The main aim of this study is to track the yields of water radiolysis, particularly the toxic reactive oxygen species Hydrogenperoxide (H2O2). TOPAS-nBio Monte Carlo track structure simulations were conducted across diverse conditions, such as dose rates, number of pulses, and particle beam energy. The yields of (H2O2) exhibited a notable linear increase with the delivered dose, and the dose rates were significantly impacting the (H2O2) across all three particle beams. Whereas these yields (H2O2) can be ranked: FLASHphoton > UHDRproton > UHDRelectron > CONVelectron.

1 Introduction

FLASH Radiation Therapy (FLASH-RT) emerges as an innovative approach in radiation oncology. Its innovation hinges on an ultra-high dose rate (UHDR) delivery, this accomplished by administering radiation under UHDR (≥ 100Gy/s), with a single or a series of pulses, within intra-pulse dose rate in the range (10^4 − 10^9 Gy/s) [1]. FLASH-RT demonstrates exceptional effectiveness in sparing normal tissues. This effect demonstrates an in vivo response wherein mitigating normal tissue toxicities is observed, while maintaining equivalent efficacy in controlling tumor growth comparable to conventional radiotherapy (CONV-RT). Initially, it was observed in various experimental animal models and in different organs, [2,8]. In the realm of FLASH-RT, the concentration of chemical species yields is expected to be altered. Ultimately, this would change the diffusion and kinetics of different radiation chemical reactions. Consequently, these changes would influence the cellular response to irradiation. In this context, hydrogen peroxide plays a pivotal role in water radiolysis. Contributes significantly to DNA damage, primarily inducing single-strand breaks, by its interaction with cellular complexes results in the formation of hydroxyl radicals (OH−) and solvated electrons, augmenting oxidative stress within the biological system, in addition, it diffuses across cellular compartments and participates in Fenton 1 electron chemistry, oxidizing Fe2+ to Fe3+ and generating the hydroxyl radical.[3]. The purpose of this research is to investigate the yields of water radiolysis, with particular attention given to Hydrogen Peroxide (H2O2). Tracking the yields of H2O2 for prolonged time scales till the homogeneous chemistry stage. Under varied particle beams (i.e. Electron, Protons, and Phtons), different pulse beam structures, dose rates. The study involved different proton beam energies, including oxygen concentration, and specified a neutral pH medium. In addition to that, the study encompassed the consideration of temperature effects [4], investigating its impact under CONV versus UHDR. This exploration aimed to gain further insights into its pivotal role in driving reactions and molecular diffusion under both modalities. Moreover, it sought to establish a fair comparison between the MC simulation and the experimental outcomes.

2 Materials and Methods

In the context of FLASH-RT, to simulate water radiolysis, it requires the following chemical species beyond 1us [5]. To this end, the scoring example performed by the TOPAS-nBio collaboration data base, was executed in a manner where two methodologies were employed: independent histories and accumulated histories. The latter method was utilized to consider the interaction of multiple histories within a single pulse within a width 1-2μs. A "history" was defined as the interaction of the primary particle and all its secondaries with the medium, which spanned the subsequent production of all chemical species present after the nonhomogeneous chemistry stage. The methodology to simulate this temporal pulse shape has previously been elucidated by Ramos-Mendez et al. [6]. The mean time was determined based on pulse number and dose rate. The independent reaction time (IRT) algorithm was employed to simulate each pulse independently, under the assumption of no interpulse contributions. Within TOPAS-nBio toolkit, a physics module (TsEmDNAPhysics) is employed, specifically tailored to simulate the complete range of interaction events involving charged particles within water detailed in [7]. To activate the chemical models, we utilized the TsEmDNAChemistry model within TOPAS-nBio. The reactions involving all chemical species studied in this research are detailed in Table1 within [8]. Notably, dissolved oxygen with a concentration of 52μM (4%), is incorporated in this simulation, however, it was represented as a continuous background. Figure 1 shows the full geometry set ups used in this study.

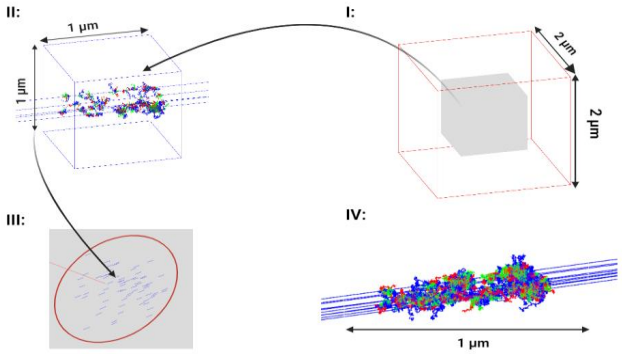


Figure 1: The illustration of the geometry employed in this study :(I) The entire geometry, encompassing the outer box (2 μm^3), constituting the world volume filled with G4−Water. The inner box (1 μm^3), represents the scoring volume, filled with G4−Water. (II) Depicts the scoring volume, showcasing the trajectory of primary particles in the Z direction and the diffusion of chemical species.

3 Results

The evolution of H2O2 concentration as a function of time was simulated with TOPAS-nBio, for various pulsed electron beam structures within 1 MeV energy. The left panel (Figure 2) exhibits the temporal evolution of [H2O2] for FLASH-RT dose rates (500Gy/s). Across the entire range of total doses simulated, the temporal behavior of [H2O2] consistently displays a stable pattern, for all total doses.

For instance, when a total dose of 10 Gy was delivered in three squared pulse beams, each with a mean width of 1μs, it resulted in a concentration of [H2O2] equal to 0.9μM, within an irradiation time up to 1ms. To administer a total dose of 20Gy, an increased number of pulses up to six was used. This adjustment doubled the [H2O2] yield to 1.74μM. Meanwhile, a similar behavior was observed with other total doses (30 − 60Gy). Furthermore, in the medium panel (Figure2), the depiction illustrates the progression of [H2O2] under CONV-RT conditions, which characterized by dose rates of 2.8 Gy/s. Consistently observed across the entire spectrum of simulated total doses, the temporal trends of [H2O2] exhibit persistent stability, even amidst a substantial number of simulated pulses and escalations in irradiation duration up to seconds. Delivering a total dose of 10 Gy by CONV-RT (2.8Gy/s) scenario requires involvement up to three hundred square pulse beams, each spanning a mean width of 1μs, culminating in an [H2O2] concentration of 0.77μM within an irradiation period of up to 13.5 s. Scaling

the total dose to 20Gy and 30Gy or even more, which, doubling the resulting [H2O2] yield to 1.45µM or tripling to 2.25µM , respectively. This trend persisted across varied total doses (30 − 60Gy), manifesting a linear correlation between the total dose and [H2O2] concentration at the end of irradiation period as depicted in left panel (Figure2). In addition to that, dose rates significantly increased the yield of H2O2.

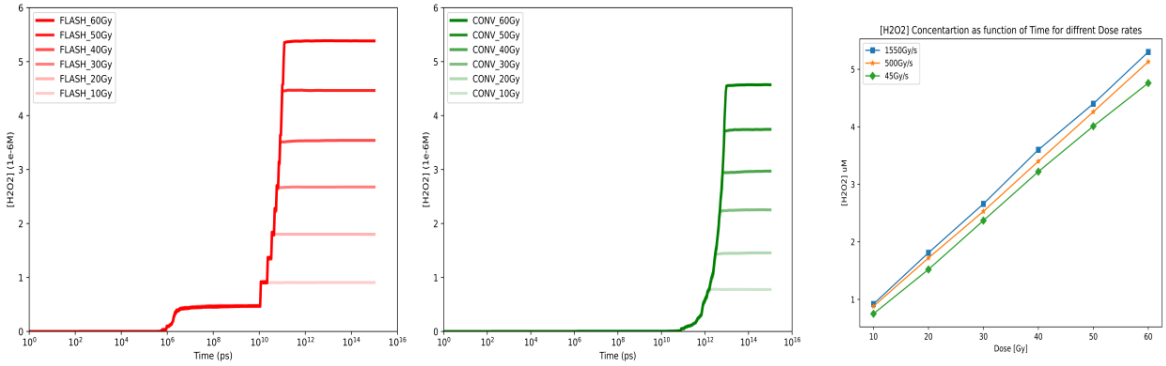


Figure 2: HydrogenPeroxide Concentrations ([H2O2] µM) as function of (log10(T ime)), under UHDR and CONV-RT pulsed electron beams, oxygen concentration 52µM considered. The Total dose delievred was increased by an increment of 10 units up to 60 Gray, as described in the legend by different opacity.

The demonstrated promising outcomes associated with Proton beams, as highlighted in contrast to electrons and photons [9]. Alongside, the mitigation of proton-induced effects on ROS yields, prompt an inquiry, within the framework of MC-Simulation, a systematic exploration encompassing a comparative analysis of particle beams on [H2O2] concentration is depicted in Figure3.

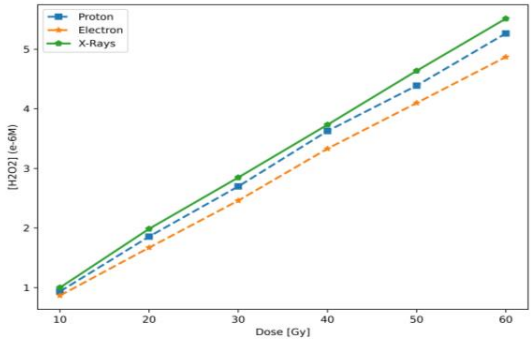


Figure 3: H2O2 concentration as function of total dose, for electron, proton and x-rays. Within a ultrahigh dose rates value 500 Gy/s with same previous pulse beam structure.

Whereas the same beam structure was used to achieve dose rates of 500 Gy/s, protons were chosen with an energy of 300 MeV to have 0.3 keV/µm of (LET), and electrons adjusted to an energy of 0.5 MeV, and X-rays with 250 keV. It's clear from (Figure 3) that X-rays produced more H2O2 compared to other particle beams, and for proton compared to electron across all total doses, where the [H2O2] response to total dose was maintained (increases linearly).

4 Discussion

In the context of FLASH-RT, simulating water radiolysis necessitates consideration of chemical species beyond 1 microsecond. Leveraging our Monte Carlo (MC) simulation methodology, we successfully addressed ultra-high dose rate values akin to those employed in biological experiments. Additionally, the pulse width in our simulations was aligned with those applied in clinical applications. Our findings align consistently with prior MC investigations documented in the literature [8]. However, the generation of

reactive oxygen species (ROS) is notably sensitive to factors such as the concentration of dissolved oxygen in water, (LET), chemical reactions, reaction rates and temperature. In addition to other parameters related to UHDR modality, like number of pulses, dose pre pulse, and mean dose rates.

The findings reported in [8] were obtained by implemented an algorithm called Gellespie which takes the simulation to very long period, whereas some additional reaction contributing to H2O2 generation occurred, which is not in our case. Thus, make some changes in term of (H2O2) amplitude for same total dose, and dose rates. However, 90 % of (H2O2) yields are from self rection of OH (OH+OH->H2O2).

On other hand, biological experiments are expecting (H2O2) to be decreased when dose rates are increased [9], which is not pronounced with MC simulation findings. However, to compare the biological outcomes and MC simulations findings in a rigorous manner, it is worth noting the multitude of parameters that significantly influence water radiolysis, particularly (H2O2) yields. Including LET, which emerges as a crucial determinant, as demonstrated in Figures 2 and 3. Furthermore, the time window during which (H2O2) was calculated and the choice of organic samples also play pivotal roles, given their heightened sensitivity to (H2O2). Further investigation is needed to explore the mechanisms behind the FLASH effect. Water radiolysis alone is not sufficient; it is imperative to consider other biological moieties and plans for future studies involving radical scavengers are underway.

5 Conclusion

In summary, our observations underscore the significant impact of dose rates on H2O2 yields, in addition to the influence of LET. Increasing dose rates and utilizing different particle types resulted in elevated H2O2 yields. Conversely, a high number of pulses led to a substantial reduction in H2O2 production. Our findings shed light on the necessity of understanding the early chemistry stages and their driving factors, which might contribute to the FLASH effect.

References

[1] V. Favaudon, L. Caplier, V. Monceau, F. Pouzoulet, M. Sayarath, C. Fouillade, M. F.Poupon, I. Brito, P. Hupé, J. Bourhis, J. Hall, J. J. Fontaine, M. C. Vozenin, Ultrahigh dose-rate flash irradiation increases the differential response between normal and tumor tissue in mice, Science Translational Medicine 6 (2014). Doi: https://doi.org/10.1126/scitranslmed.3008973

[2] M. C. Vozenin, P. D. Fornel, K. Petersson, V. Favaudon, M. Jaccard, J. F. Germond, B. Petit, M. Burki, G. Ferrand, D. Patin, H. Bouchaab, M. Ozsahin, F. Bochud, C. Bailat, P. Devauchelle, J. Bourhis, The advantage of flash radiotherapy confirmed in mini-pig and cat-cancer patients, Clinical Cancer Research 25 (2019) Doi: https://doi.org/10.1158/1078-0432.CCR-17-3375.

[3] R. Ameziane-El-Hassani, M. Boufraqech, O. Lagente-Chevallier, U. Weyemi, M. Talbot, D.M etivier, F. Courtin, J. M.Bidart, M. E. Mzibri, M. Schlumberger, C. Dupuy, Role of h2o2 in ret/ptc1 chromosomal rearrangement produced by ionizing radiation in human thyroid cells, Cancer Research 70 (2010). Doi: https://doi.org/10.1158/0008-5472.CAN-09-4336

[4] J. Ramos-Méndez, O. Garcia-Garcia, J. Dominguez-Kondo, J. A. LaVerne, J. Schuemann, E. Moreno-Barbosa, and B. Faddegon,TOPAS-nBio simulation of temperature dependent indirect DNA strand break yields, Physics in Medicine & Biology 67, 145007 (2022). DOI 10.1088/1361-6560/ac79f9.

[5] F. Chappuis, H. N. Tran, S. A. Zein, C. Bailat, S. Incerti, F. Bochud, L. Desorgher, The general-purpose geant4 monte carlo toolkit and its geant4-dna extension to investigate mechanisms underlying the flash effect in radiotherapy: Current status and challenges (2023). https://doi.org/10.1016/j.ejmp.2023.102601

[6] J. Ramos-Mendez, N. Dominguezz-Kondo, J. Schuemann, A. McNamara, E. Moreno Barbosa, and B. Faddegon, LET-dependent intertrack yields in proton irradiation at ultra-high dose rates relevant for FLASH therapy, Radiation research 194, 351–362 (2020). https://doi.org/10.1667/rade-20-00084.1

[7] J. Ramos-Méndez et al., TOPAS-nBio validation for simulating water radiolysis and DNA damage under low-LET irradiation, Physics in Medicine & Biology 66, 175026 (2021). https://doi.org/10.1088/1361-6560/ac1f39

[8] D-Kondo JN, Garcia-Garcia OR, LaVerne JA, Faddegon B, Schuemann J, Shin WG, Ramos-Méndez J. An integrated Monte Carlo track-structure simulation framework for modeling inter and intra-track effects on homogenous chemistry. Phys Med Biol. 2023 Jun 9;68(12): doi: 10.1088/1361-6560/acd6d0.

[9] Kacem, Houda, et al. "Comparing radiolytic production of H2O2 and development of Zebrafish embryos after ultra high dose rate exposure with electron and transmission proton beams." *Radiotherapy and oncology* 175 (2022): 197 202. https://doi.org/10.1016/j.radonc.2022.07.011

Log file-based patient specific QA using GPU-accelerated Monte Carlo simulation for proton pencil beam line scanning therapy

Chanil Jeon[1], Sung Hwan Ahn[2], Jungwook Shin[3], Hoyeon Lee[4], Youngyih Han[1,5]

[1]Department of Heal Sciences and Technology, SAIHST, Sungkyunkwan University, Seoul, Republic of Korea
[2]Department of Radiation Oncology, Samsung Medical Center, Seoul, Republic of Korea
[3]Devision of Cancer Epidemiology and Genetics, National Cancer Institute, National Institute of Health, RockVille, Maryland, USA
[4]Department of Radiation Oncology, Massachusetts General Hospital and Havard Medical School, Boston, MA 02114, USA
[5]Department of Radiation Oncology, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, Republic of Korea

Abstract

Monte Carlo (MC) simulation is one of the accurate dose calculation methods in radiation therapy applications. However, long calculation time for simulation is major drawback makes it difficult to utilize MC simulation in practical applications. In this study, we developed log file-based MC calculation application using GPU for proton pencil beam line scanning therapy. The beam parameters of the simulation were determined using beam measurement data for clinically used energy range. To validate the log file-based GPU MC method, eight patient-specific quality assurance plans was used. The two-dimensional absorbed dose at a depth 2cm of water phantom from log file and GPU-based MC was compared with that of the log file and CPU-based MC calculation with gamma evaluation using 2% / 2mm and 5% high dose threshold criterion. The average, maximum, minimum, and standard deviation of 2% / 2mm gamma evaluation were 99.6, 100, 97.9 and 0.55%, respectively.

1 Introduction

Monte-Carlo (MC) simulation is one of the accurate dose calculation methods in radiation therapy applications. However, most of currently developed MC codes are severely time-consuming to calculate high-resolution volumetric doses due to less efficient calculation of CPU. Therefore, it is difficult to utilize the MC simulation deliberately for quality assurance (QA) procedure in most of hospital environments.

Recently, some studies developed GPU-based MC methods for various radiation therapies. For proton therapy, Lee H. et al developed GPU-based proton dose calculation toolkit named Monte Carlo code for quick proton doe calculation (MOQUI) that uses DICOM RT format files [1]. They benchmarked against Tool for particle simulation (TOPAS) MC simulation [2], which was based on Geant4 simulation toolkit [3]. The speed of the MOQUI was increased with simplified physics model and techniques of general-purpose computing on graphics processing units. However, the method cannot simulate proton pencil beam line scanning treatment plan because a DICOM RT-ion plan of the line scanning plan consists of line segment points, dissimilar to the spot scanning plan. Furthermore, the DICOM-based dose calculation does not reflect realistic beam property from daily treatment machine output variation. Thus, the actual beam delivery of treatment plan cannot be accurately accounted for practical environments. In this regard, we developed a log file-based dose calculation method using MOQUI MC simulation in this study. The log file of line scanning has spot-type data with very short time resolution (60 μs) to monitor continuous beam irradiation, which can be adapted to MOQUI. In addition, log file data reflects daily condition of treatment machine and deliverability of treatment plan. To validate log-file and GPU-accelerated MC simulation method, dose calculation results of our previous study [4], the log file-based patient-specific QA method based on TOPAS MC simulation toolkit was used.

2 Materials and Methods

To convert log files to readable information format for the MC code, in-house MC code generation graphic user interface program was developed (Figure 1). The program reads calculated isocenter position of beam and dose count value (MU count) which is proportional to MU from flatness monitor and dose monitor of proton pencil beam line scanning nozzle, respectively. The beam information was reconstructed to plan type files that can be utilized for MC simulation using TOPAS and MOQUI.

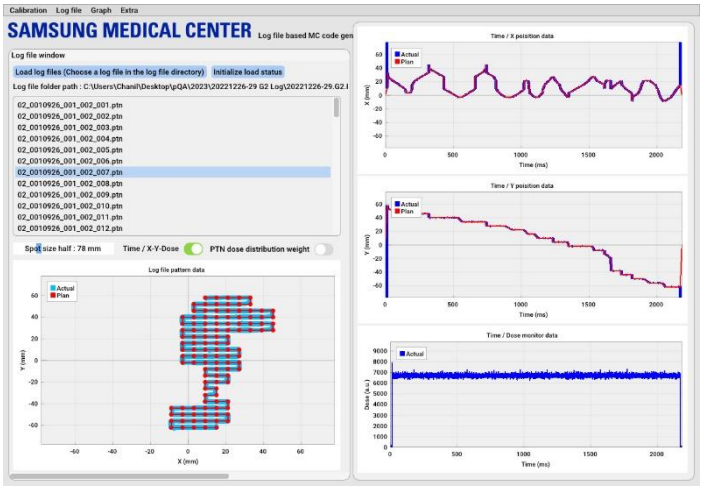


Figure 1. In-house Monte Carlo simulation code generation program

The measured beam data of spot size, angular spread and energy spread was used to determine beam parameters of both MC code for beam energy range of 70 to 230 MeV. The conversion of log file MU count to number of particles in MC simulation was determined by Equation 1, which was validated in the previous research using TOPAS MC toolkit [4]. $MU\ count_{log}$ is MU count of log file, $ED_{Dose/MU\ count}$ and $ED_{Protons/Dose}$ are energy dependent relationships for dose (Gy) per unit MU and number of protons per unit dose (Gy), respectively. Those relationships are determined using MC

simulation and measurement with an ionization chamber array (Octavius 1500, PTW, Freiburg, Germany). R_{DM} is dose monitor (DM) range value saved separately in treatment system console database. Owing to numeral limitation of MU count in a log file, the DM records overflowed MU count from high intensity beam as smaller MU count, with dividing DM range value. The one-level higher DM range value is three times higher than lower-level DM range value.

$$N_{Protons} = MU\ count_{\log} \times ED_{Dose/MU\ count} \times ED_{Protons/Dose} \times R_{DM} \quad (1)$$

The total number of generated particles using Equation 1 for each QA plan was ranged from 1×10^8 to 8×10^8. The GPU computing node for the simulation consisted of Intel Xeon (R) Gold 6326 CPU (2.9GHz) and Nvidia RTX A5000 24GB GPU.

To validate log file-based MOQUI against log file-based TOPAS MC, eight patient-specific QA plan consists of the lung, esophagus, liver, abdomen, brain, and craniospinal irradiation (CSI) for water phantom geometry, were used to calculate three-dimensional (3D) absorbed dose and compared two-dimensional (2D) dose at a depth of 2 cm with that of TOPAS MC.

The calculated MOQUI 2D dose distribution was compared with that of TOPAS using gamma evaluation of 2%/2mm, suppressing dose below 5% of maximum dose of calculated volume.

3 Results

The run-time of the simulation was ranged from 20 to 100 minutes for each patient case to calculate 3D dose of the water phantom using log file-based MOQUI, while the log file-based TOPAS required 8 to 40 hours to simulate for 2D dose at depth 2 cm. The total scorer volume of water phantom using MOQUI was $400 \times 400 \times 400$ cm^3 with $1 \times 1 \times 1$ mm^3 resolution, and that of TOPAS was $400 \times 400 \times 2$ mm^3 with $1 \times 1 \times 2$ mm^3 resolution. The number of energy layers and beam energy for each QA plan field were ranged from 9 to 56 and 70.0 to 171.6 MeV, respectively.

For eight patient QA plans with total 25 treatment fields, the average, maximum, minimum, and standard deviation of 2% / 2mm gamma evaluation were 99.6, 100, 97.9 and 0.55%, respectively.

4 Discussion

In this study, we developed patient specific QA method utilizing log files and GPU-accelerated MC simulation. The accuracy of the proposed method in water phantom condition was validated against the results of previously developed log file and CPU calculation-based MC simulation method. The gamma evaluation for 2D dose distribution was performed with 2% / 2mm and 5% high dose threshold, the acceptable assessment criterion for clinical environments.

Owing to faster dose calculation speed and reflection of practical beam conditions due to log files, the efficient three-dimensional dose calculation and gamma evaluation for patient specific QA are feasible with the proposed method. Furthermore, four-dimensional dose calculation attract attention recently attributing to feasibility of investigation on patient's responsiveness from diverse treatment planning strategies. Using the actual beam irradiation time records in a log file, the synchronization between beam irradiation and position of internal target and critical organ is more practical than the DICOM RT-ion plan computed on planning CT and breathing information. In future studies, the proposed system will be validated against diverse treatment plans and complex geometry. The comparison study of 3D gamma against 2D gamma criteria is underway.

5 Conclusion

We successfully developed log file and GPU-accelerated MC-based dose calculation method and demonstrated the potential of utilization for the clinical applications. In future studies, we aim to optimize MOQUI code for faster calculation speed to use efficiently and assess proposed method with a patient geometry to evaluate the accuracy of the method in inhomogeneous environments.

Acknowledgements

This work was supported by the National Research Foundation of Korea grant funded by the Korean government. (No. 2021ME2E8A1048108).

References

[1] Lee, H., Shin, J., Verburg, J. M., Bobić, M., Winey, B., Schuemann, J., & Paganetti, H. (2022). MOQUI: an open-source GPU-based Monte Carlo code for proton dose calculation with efficient data structure. Physics in medicine and biology, 67(17), 10.1088/1361-6560/ac8716. https://doi.org/10.1088/1361-6560/ac8716

[2] Faddegon, B., Ramos-Méndez, J., Schuemann, J., McNamara, A., Shin, J., Perl, J., & Paganetti, H. (2020). The TOPAS tool for particle simulation, a Monte Carlo simulation tool for physics, biology and clinical research. Physica medica : PM : an international journal devoted to the applications of physics to medicine and biology : official journal of the Italian Association of Biomedical Physics (AIFB), 72, 114–121. https://doi.org/10.1016/j.ejmp.2020.03.019

[3] Agostinelli, S., Allison, J., Amako, K. A., Apostolakis, J., Araujo, H., Arce, P., ... & Geant4 Collaboration. (2003). GEANT4—a simulation toolkit. Nuclear instruments and methods in physics research section A: Accelerators, Spectrometers, Detectors and Associated Equipment, 506(3), 250-303. https://doi.org/10.1016/S0168-9002(03)01368-8

[4] Jeon C, Lee J, Shin J, et al. Monte Carlo simulation-based patient-specific QA using machine log files for line-scanning proton radiation therapy. Med Phys. 2023; 50: 7139 – 7153. https://doi.org/10.1002/mp.16747

A fast 3D range-modulator delivery approach: sensitivity of the resulting dose distribution on potential setup misalignments

Yuri Simeonov[1,4], Miriam Krieger[5], Michael Folkerts[5], Gerard Paquet[5], Pierre Lansonneur[5], Petar Penchev[1], Klemens Zink[1,3], Uli Weber[2] and Christoph Schuy[2]

[1]University of Applied Sciences, Institute of Medical Physics and Radiation Protection, Giessen, Germany

[2]GSI Helmholtzzentrum für Schwerionenforschung GmbH, Biophysics division, Darmstadt, Germany

[3]University Hospital Giessen-Marburg, Department of Radiotherapy and Radiooncology, Marburg, Germany

[4]Philipps University Marburg, Department of Medicine, Marburg, Germany

[5]Varian Medical Systems

1 Introduction

Particle therapy has established clinically in the last decades as it can apply the dose to the target with high precision and conformity and has proven to be particularly effective for certain types of cancer. Moreover, in recent years the ultra-high dose rate "FLASH" irradiation has gained a lot of research interest, as it is thought to reduce the side effects in healthy tissue and increase the therapeutic window [1–3].

However, moving tumors pose a challenge to the conventional multi-energy raster scanning method due to the relatively long irradiation time and the resulting "interplay" effects between the scanned beam and the target motion, which can lead to strong dose inhomogeneities [4,5]. Furthermore, it is difficult to achieve ultra-high dose rate FLASH irradiation due to the switching time between the individual energy layers.

3D range-modulators (3DRM), optimized for a single energy and individual tumor shape are a promising and viable solution for FLASH dose delivery in particle therapy. In combination with a high-intensity accelerator, irradiation time can be reduced tremendously [6,7]. However, a 3DRM is an additional beam modifying device in the beam line, and as such it is important to understand and analyze the impact of potential misalignments on the dose distribution.

2 Materials and Methods

A 6 cm side-length cube, rotated twice at 45°, was used in Eclipse as a target shape to develop and optimize a 3DRM for 250 MeV proton-beam. The rotated cube form ensures relatively sharp proximal and distal edges of the 3DRM and consequently of the dose distribution and together with the maximum pin length of approx. 8 cm can be considered close to a "worst-case" scenario. The dose distribution (reference dose) in water behind the 3DRM and a 17 cm thick PMMA absorber was simulated with the FLUKA Monte Carlo package [8–10]. Subsequently, 5 different modified simulations were performed and compared with the reference dose:

- Raster plan shifted **1** mm in **X**
- 3DRM shifted **1.5** mm in **X**
- 3DRM shifted **1.5** mm in both **X** and **Y**
- 3DRM rotated **0.5**° around **Y**
- 3DRM rotated **1.5**° around **Y** (Figure 1)

The rotation point used was X=20 cm, Z=18 cm in front of the 3DRM and relative to its own isocenter as it approximately represents the mechanical mounting point, which will carry the weight of the 3DRM and the PMMA in the snout and might be susceptible to a weight-dependent tilt.

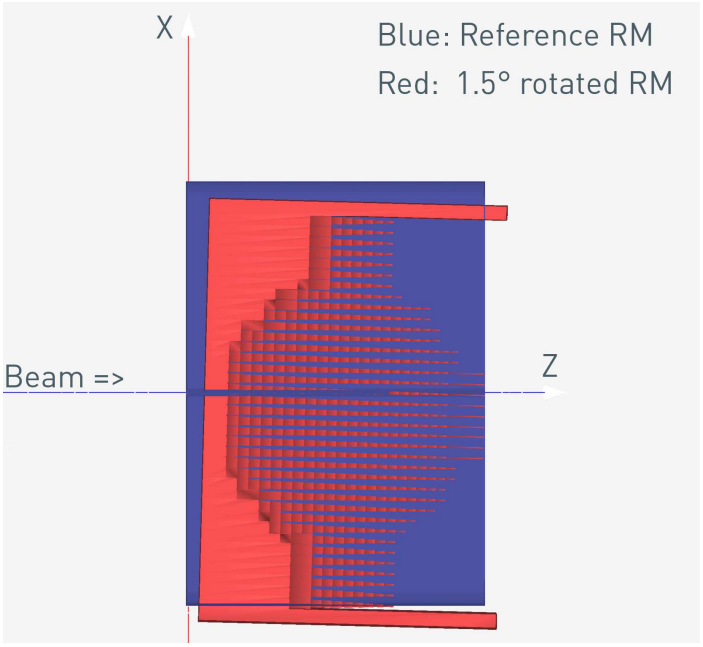


Figure 1: The reference (schematic view) and the rotated (cross section view) 3DRM

3 Results

Figure 2 shows the 2D Gamma Index (GI) for the different scenarios for one proximal (11 cm depth) and one distal (16 cm depth) slice. Additionally, a dose comparison for the 0.5° and 1.5° angular deflection case are plotted (Figure 3 - Figure 4).

Shift/Rotation	Gamma Index (2%/2mm, local)	
	Proximal (11cm)	Distal (16cm)
Raster plan: **1**mm shift in **X**	100%	100%
3DRM: **1.5**mm shift in **X**	100%	100%
3DRM: **1.5**mm shift in both **X/Y**	99%	73%
3DRM: **0.5°** angular deflection	99%	86%
3DRM: **1.5°** angular deflection	77%	23%

Figure 2: 2D Gamma Index for one proximal and one distal depth slice for the different scenarios

4 Discussion

The SOBP transitions to a broad, Gaussian-like distribution with increasing RM rotation, whereas the sharp distal dose is laterally shifted.

5 Conclusion

A shift/rotation of the 3DRM shifts the distal dose. The dose deviations depend thereby on the specific 3DRM and its shape. Precise positioning is important. However, increasing the pin base is expected to mitigate the SOBP dose deterioration to a large extent. It can be concluded that even complex 3D modulators with long pins are robust against a slight misalignment of 0.5° rotation or 1.5 mm shift.

References

[1] V. Favaudon *et al.*, Ultrahigh dose-rate FLASH irradiation increases the differential response between normal and tumor tissue in mice, Sci. Transl. Med., vol. 6, no. 245, pp. 245ra93-245ra93, Jul. 2014, doi: 10.1126/scitranslmed.3008973.

[2] M.-C. Vozenin, J. H. Hendry, and C. L. Limoli, Biological Benefits of Ultra-high Dose Rate FLASH Radiotherapy: Sleeping Beauty Awoken, Clin. Oncol., vol. 31, no. 7, pp. 407–415, Jul. 2019, doi: 10.1016/j.clon.2019.04.001.

[3] E. S. Diffenderfer, B. S. Sørensen, A. Mazal, and D. J. Carlson, The current status of preclinical proton FLASH radiation and future directions, Med. Phys., vol. 49, no. 3, pp. 2039–2054, 2022, doi: 10.1002/mp.15276.

[4] J. Lambert, N. Suchowerska, D. R. McKenzie, and M. Jackson, Intrafractional motion during proton beam scanning, Phys. Med. Biol., vol. 50, no. 20, p. 4853, 2005, doi: 10.1088/0031-9155/50/20/008.

[5] C. Bert, S. O. Grözinger, and E. Rietzel, Quantification of interplay effects of scanned particle beams and moving targets, Phys. Med. Biol., vol. 53, no. 9, p. 2253, 2008, doi: 10.1088/0031-9155/53/9/003.

[6] Y. Simeonov *et al.*, 3D range-modulator for scanned particle therapy: development, Monte Carlo simulations and experimental evaluation, Phys. Med. Biol., vol. 62, no. 17, pp. 7075–7096, Aug. 2017, doi: 10.1088/1361-6560/aa81f4.

[7] Y. Simeonov *et al.*, Development, Monte Carlo simulations and experimental evaluation of a 3D range-modulator for a complex target in scanned proton therapy, Biomed. Phys. Eng. Express, vol. 8, no. 3, p. 035006, Mar. 2022, doi: 10.1088/2057-1976/ac5937.

[8] A. Ferrari, P. R. Sala, A. Fassò, and J. Ranft, *FLUKA: A multi-particle transport code (program version 2005)*. CERN, 2005.

[9] T. T. Böhlen *et al.*, The FLUKA Code: Developments and Challenges for High Energy and Medical Applications, Nucl. Data Sheets, vol. 120, pp. 211–214, Jun. 2014, doi: 10.1016/j.nds.2014.07.049.

[10] G. Battistoni *et al.*, The FLUKA Code: An Accurate Simulation Tool for Particle Therapy, Front. Oncol., vol. 6, May 2016, doi: 10.3389/fonc.2016.00116.

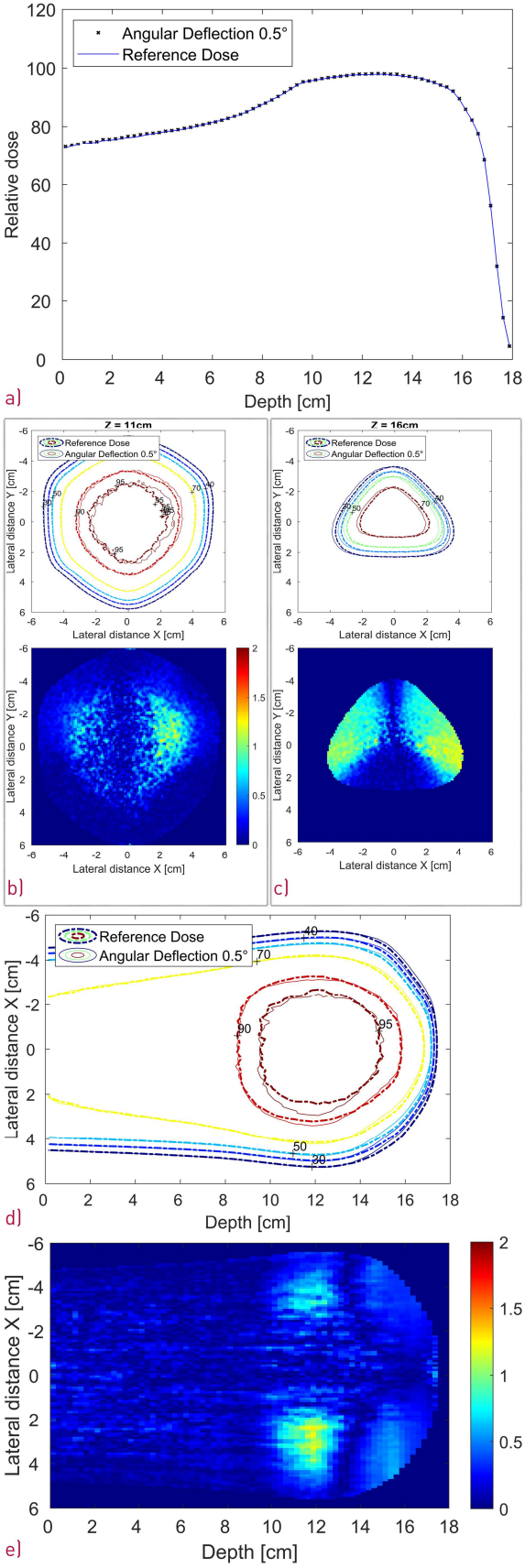


*Figure 3: **0.5**° RM Rotation: Comparison of the centre line depth dose distribution (a); X-Y isodose comparison with the corresponding 2D GI for one proximal (Z = 11 cm) and one distal (Z = 16 cm) depth slice (b-c); X-Z isodose comparison (middle of dose distribution) with the corresponding 2D GI (d-e)*

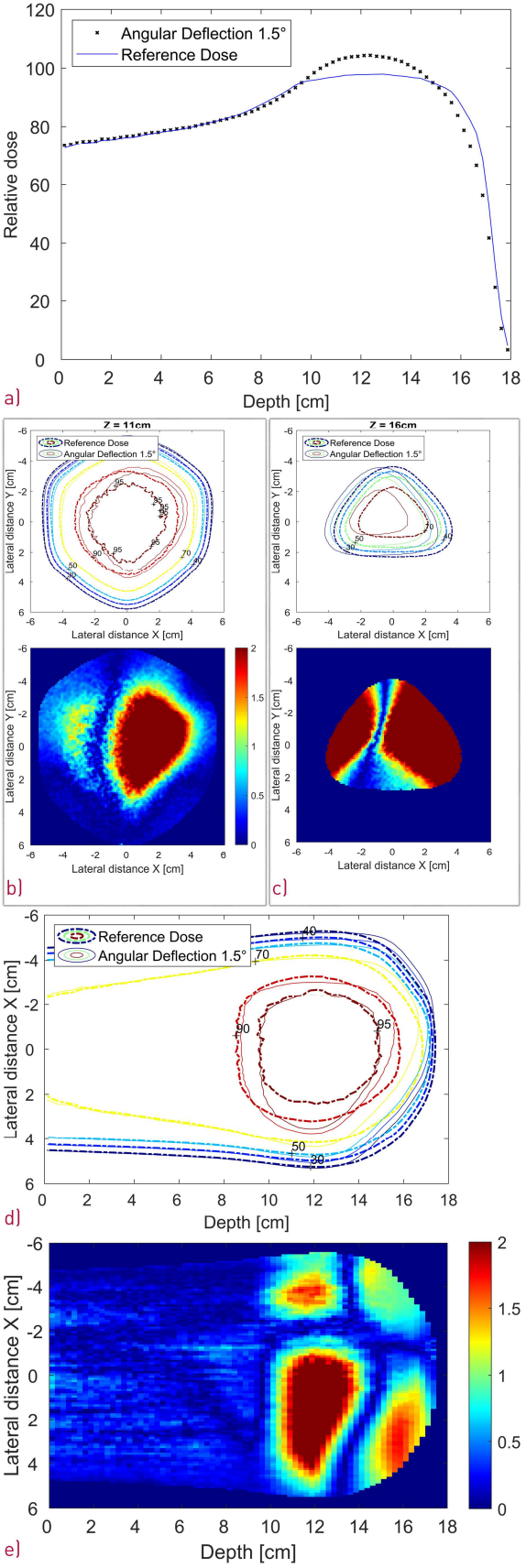


*Figure 4: **1.5**° RM Rotation: Comparison of the centre line depth dose distribution (a); X-Y isodose comparison with the corresponding 2D GI for one proximal (Z = 11 cm) and one distal (Z = 16 cm) depth slice (b-c); X-Z isodose comparison (middle of dose distribution) with the corresponding 2D GI (d-e)*

Benchmark of GEANT4 hadronic models for carbon-therapy within the CLINM project

Lévana Gesson[1], Claire Reibel[1], Nicolas Arbor[1], Aurélia Arnone[1], Christian Finck[1], Catherine Galindo[1], Stéphane Higueret[1], Thé-Duc Lê[1], Arshiya Sood[1], Quentin Raffy[1], Philippe Peaupardin[1], Marco Pullia[2], and Marie Vanstalle[1]

[1]Université de Strasbourg, CNRS, IPHC UMR 7178, F-67000 Strasbourg, France
[2]Centro Nazionale di Adroterapia Oncologica (CNAO), Pavia, Italy

Abstract Optimization of hadrontherapy treatments implies careful consideration of the secondary particles produced by the ion beam fragmentation in the patient. Unfortunately, this is still far from being possible in the software currently used for treatment planning, and research projects are needed to improve the multi-scale calculations of secondary particle production and associated radiobiological effects.
In order to accurately replicate biological responses to ionizing radiation, understanding both the physical and chemical effects of ions is crucial. The CLINM (Cross-Sections of Light Ions and Neutron Measurements) project, aims to comprehensively characterize secondary particles from ion fragmentation on tissues and their associated chemical effects, comparing both simulation and experimental results.
To investigate both chemical and physical effects, radiolysis measurements are conducted alongside the analysis of the secondary particles physical properties (energy, charge) using a ΔE-E telescope. Experiments were carried out at CNAO hadrontherapy center. Water samples were placed after a tissue-equivalent target, irradiated with ^{12}C ions at various energies, in order to study water radiolysis generated by the beam and its secondary particles.
This work will focus on the physical part of the experiment: a ΔE-E telescope, consisting of a thin plastic scintillator (2 mm) and a cerium bromide crystal scintillator ($CeBr_3$), was employed at two different angles (0 and 5°) behind the target, in order to characterize the secondary particles field. Experimental data were compared to simulated results obtained with Geant4.

1 Introduction

During a cancer treatment with particle therapy (heavy ion or proton beam), nuclear reactions of the primary beam with the targeted volume need to be quantified in order to calculate exactly the dose received by the patient. Charged particle therapy (CPT) permits, compared to therapy using X-rays, to deliver a more conformal dose to the tumor while sparing healthy tissues. However, the primary beam and target fragmentation lead to the production of lighter fragments, which may contribute to undesired dose in healthy tissues. Dose calculations in particle therapy rely on high-performance algorithms which include both physical and biological processes. These calculations are based on data provided by Monte Carlo simulation codes. However, there is currently a lack of experimental data regarding nuclear reactions in particle therapy, which can introduce inaccuracies in dose calculations.
In order to accurately reproduce biological phenomena occurring in case of irradiation of living cells with ionizing radiation, it is necessary to properly reproduce the physical interactions of this radiation with matter, as well as its chemical effects, such as the radiolysis of the different molecules composing the cells (water, proteins, DNA…). While several studies have addressed the question of physical properties of secondary particles produced during CPT [1–3], the crucial chemical step, that makes the link between physical and biological effects of radiations, is often neglected. The CLINM (for Cross-Sections of Light Ions and Neutron Measurements) project aims at proposing a global characterization of the secondary particles produced by fragmentation processes of ions on tissues, and of their chemical effects.
These chemical effects are studied through the measurement of the radiolysis products generated by the secondary particles, while the physical properties (energy, charge) of these particles is performed thanks to a ΔE-E telescope. A first experimental campaign of the CLINM project was performed at the CNAO hadrontherapy center (Pavia, Italy), and the achieved results of this campaign will be presented in this work. This abstract especially focus on the preliminary comparison between the obtained physical data with simulations performed with the Geant4 Monte Carlo code.

2 Materials and Methods

During this experiment, the secondary particles produced by different carbon ions energies on different thicknesses of RW3 slabs phantom were measured. Different thicknesses of RW3 were used: 5 cm and 23 cm, with ^{12}C energies of 200 MeV/u and 400 MeV/u respectively. The goal was to achieve a ^{12}C energy of 120 MeV/u exiting the target in all configurations.

The setup to measure the physical characteristics of the secondary particles (charge, energy, yields) is presented on Figure 1(a-b). It was made of a ΔE-E telescope, with one plastic scintillator of 2-mm thickness (ΔE) and one cerium bromide crystal of 5 cm thickness (E). The $CeBr_3$ detector is a cylindrical 2×2 inches crystal scintillator, from Advatech UK Ltd [**Advatech**], coupled to a R6231-100 Hamamatsu photomultiplier (PMT). An entrance window in front of the crystal inside the $CeBr_3$ is made of 400 μm of aluminum and 1 mm of polytetrafluoroethylene (PTFE), used as a reflector. In order to be able to detect ions such as ^{12}C without saturating the system, the high-voltage on the PMT

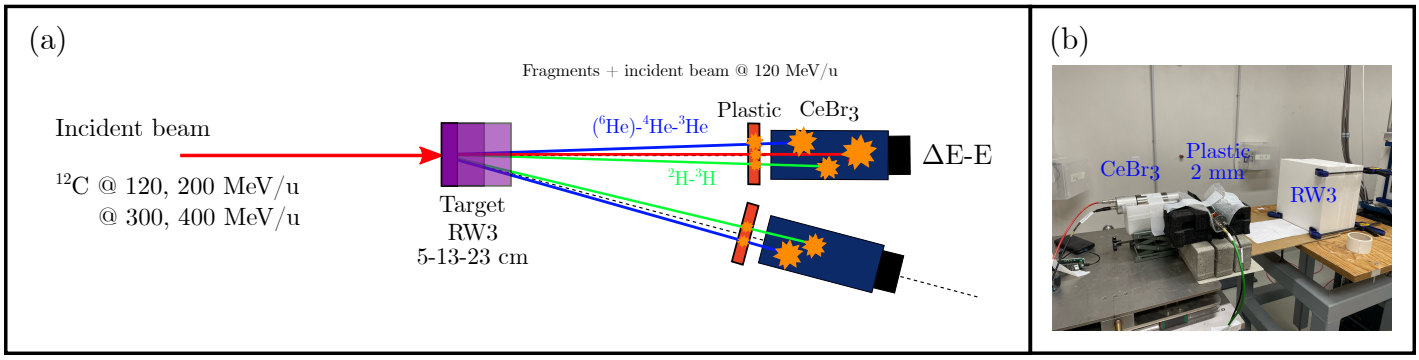


Figure 1: Scheme of the experimental setup used at CNAO.

was considerably reduced (down to +350V for ^{12}C ions) compared to the usual operating voltage (around +1200V), following recommendations from [4, 5].

The plastic scintillator, from Eljen Technology (EJ-228), is a square of 6×6×0.2 cm^3, put in coincidence with the $CeBr_3$, and coupled to XP3990 photomultiplier tube supplied with -1200V high-voltage.
Both detectors signals were aquired by a WaveCatcher digitizer module, developed by the LAL laboratory (Paris, France) [6]. The sampling frequency was set to be 3.2 GHz. This digitizer allows to reconstruct the full signal of the scintillators composing the telescope.
Obtained data were all compared with Monte Carlo simulations, carried out with Geant4 10.07, with the INCL physics list [7]. The INCL physics list implements the intranuclear Liège cascade model [8], in which nucleons are modeled as a free Fermi gas in a static potential well. A quasi-target composed of participating nucleons from the projectile entering the target as well as a quasi-projectile was built. The pre-defined QGSP_INCLXX_HP physics list was used for this final list.

3 Results

3.1 Secondary particles energy spectra

An example of ΔE-E bidimensional plot, obtained at 0° with 398.84 MeV/u ^{12}C interacting with a 23-cm thick RW3 target, is presented on Figure 2. The energy deposited in the plastic scintillator is presented as a function of the total energy released in the cerium bromide. Several populations can be distinguished, each of them corresponding to a different charge Z of secondary particle. It can be clearly seen that the helium isotopes (Z=2) are the main contribution to the secondary particles field. The incident carbon beam (Z=6) is also clearly visible on the figure.
Bidimensional ΔE-E maps were used to perform graphical cuts, in order to select the different charges that could be identified. Figures 3 and 4 present the comparison between measured and simulated energy distributions of the different charges of particles, at two different angles (0 and 5°) and

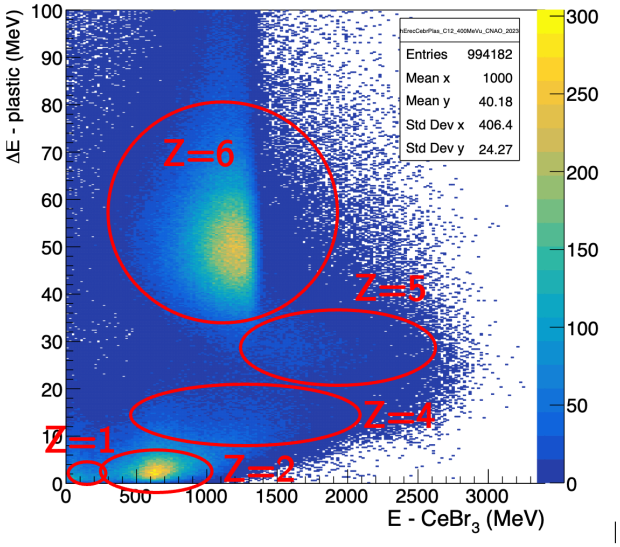


Figure 2: Energy deposited in the plastic scintillator (ΔE) as a function of total energy released in the cerium bromide (E), after the interaction of a 400 MeV/u ^{12}C beam with 23 cm of RW3 target.

two different target thicknesses (5 and 23 cm). Distributions were normalized by their integral in order to compare their shapes.
It can be observed that the shape and the average energy of measured secondary particles are in a relatively good agreement with the simulations for each Z, at 0° (Figure 3), except for Z=2 (Figure 3b), where an important underestimation of the simulated energy can be observed. At 5°, except for Z=1, the simulated energy distributions are all underestimated compared to the experimental data. In a similar fashion than what is observed at 0°, the underestimation of the energy by Geant4 is more important for Z= 2, with a factor 2 between the average measured energy and the simulated one.

4 Discussion

When the measured energy distribution of Z=2, corresponding to helium ions, is compared with the initial energy distribution of each simulated alpha-particle (Figure 5), it can be observed that both distributions are in a good agreement. It

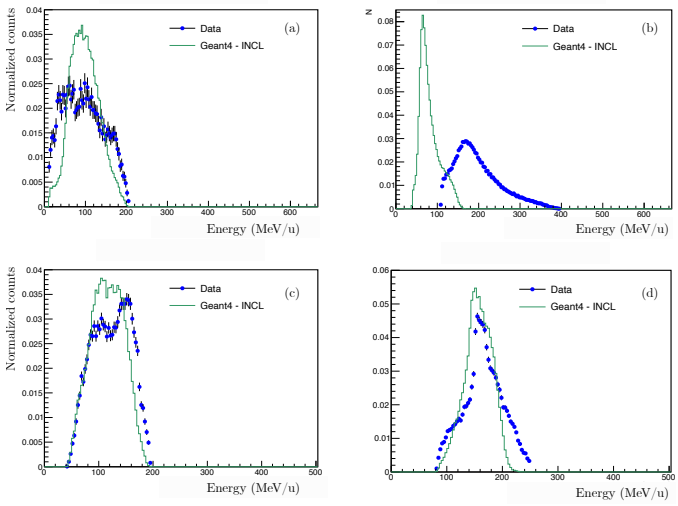


Figure 3: Comparison between simulated (green line) and measured (blue dots) energy distributions of Z=1 (a), Z=2 (b), Z=4 (c) and Z=5 (d) produced by a 400 MeV/u carbon ion beam interacting by 23-cm RW3 phantom.

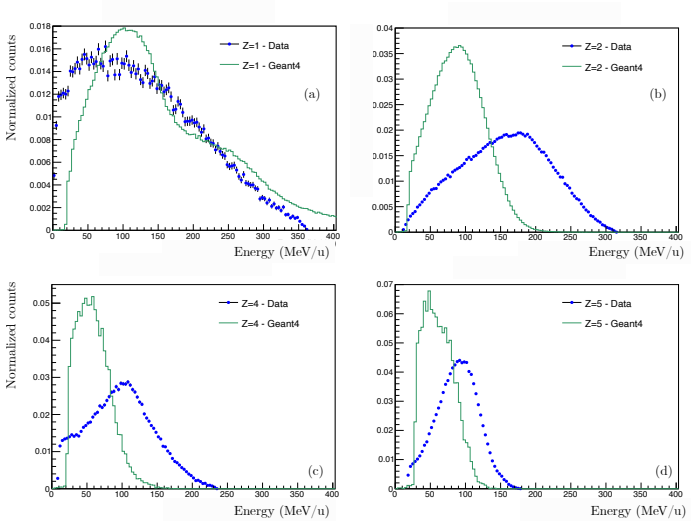


Figure 4: Comparison between simulated (green line) and measured (blue and red dots) energy distributions of Z=1 (a), Z=2 (b), Z=4 (c) and Z=5 (d) produced at 5° by a 200 MeV/u carbon ion beam interacting with 5-cm RW3 phantom.

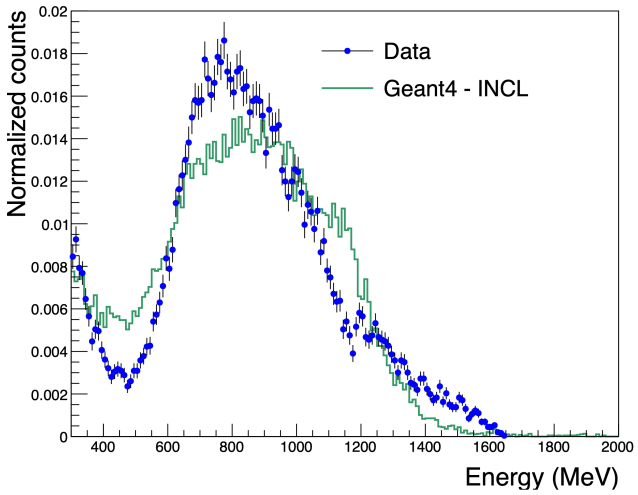


Figure 5: Comparison between the initial energy of each single helium ion simulated by Geant4 (green line) and the measured helium ion energy distribution (blue dots). Both distributions were normalized by their integral to allow better comparison.

should be emphasized that helium ions with initial energies predicted by Geant4, which are peaked around 800 MeV, do not stop into the cerium bromide scintillator of the ΔE-E telescope. Nevertheless, the measured energy in the telescope is peaked around 800 MeV (around 200 MeV/u), which indicates that the helium ions are stopped into the crystal in experimental conditions. This discrepancy between Geant4 and the measurement originates from the fact that the ΔE-E telescope measured two alpha-particles arriving on the detectors at the same time (therefore produced by the same initial carbon), while INCL is predicting that the energy corresponding to two helium-ions is actually carried by a single ion.

5 Conclusion

Preliminary results obtained at CNAO with ^{12}C ions interacting with RW3 targets have highlighted discrepancies between the INCL hadronic model predictions of Geant4 and the measured data, especially on the production of Z= 2 fragments. These results demonstrate the interest and the importance of performing measurements to benchmark Monte Carlo codes in the context of hadrontherapy. Further studies will be carry out with different types of ions and targets.

Acknowledgements

This work was possible thanks to the support of the HITRI+ project and the funding of the European Union's Horizon 2020 Research and innovation (Grant Agreement N°101008548). Authors also want to thank Petter Hofverberg from CAL facility, Angelica Facoetti from CNAO hadrontherapy center, and Michel Pellicioli, Jacky Schuler, Clément Haas and Nathalie Dick-Schuler from the Cyrcé facility. Fi-

nally, this work would also not have been possible without the work of Julie Gross and Carlotta Mozzi, and the valuable help of Cédric Mathieu.

References

[1] E. Haettner, H. Iwase, M. Krämer, et al. "Experimental study of nuclear fragmentation of 200 and 400 MeV/u12C ions in water for applications in particle therapy". *Physics in Medicine Biology* 58.23 (Nov. 2013), p. 8265. DOI: 10.1088/0031-9155/58/23/8265.

[2] J. Dudouet, D. Cussol, D. Durand, et al. "Benchmarking geant4 nuclear models for hadron therapy with 95 MeV/nucleon carbon ions". *Phys. Rev. C* 89 (5 May 2014), p. 054616. DOI: 10.1103/PhysRevC.89.054616.

[3] F. Horst, G. Aricò, K.-T. Brinkmann, et al. "Measurement of ^{4}He charge- and mass-changing cross sections on H, C, O, and Si targets in the energy range 70–220 MeV/u for radiation transport calculations in ion-beam therapy". *Phys. Rev. C* 99 (1 Jan. 2019), p. 014603. DOI: 10.1103/PhysRevC.99.014603.

[4] A. Miller, R. Machrafi, and A. Fariad. "Investigation of the LaBr3 scintillator response to heavy ions". en. *Radiation Measurements* 115 (Aug. 2018), pp. 43–48. DOI: 10.1016/j.radmeas.2018.06.009.

[5] F. Quarati, P. Dorenbos, J. van der Biezen, et al. "Scintillation and detection characteristics of high-sensitivity CeBr3 gamma-ray spectrometers". en. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 729 (Nov. 2013), pp. 596–604. DOI: 10.1016/j.nima.2013.08.005.

[6] D. Breton, E. Delagnes, J. Maalmi, et al. "The WaveCatcher family of SCA-based 12-bit 3.2-GS/s fast digitizers" (2014), pp. 1–8. DOI: 10.1109/RTC.2014.7097545.

[7] S. Agostinelli, J. Allison, K. Amako, et al. "Geant4-a simulation toolkit ". *Nucl. Instrum. Methods A* 506.3 (2003), pp. 250–303.

[8] A. Boudard, J. Cugnon, J.-C. David, et al. "New potentialities of the Liège intranuclear cascade model for reactions induced by nucleons and light charged particles". *Phys. Rev. C* 87.1 (2013), p. 014606.

Monte Carlo Modeling of Elekta VERSA HD for Out-of-Field Dose Characterization

Maxime Jacquet[1,2], Thomas Baudier[1,2], Eric Deutsch[3,4], Charlotte Robert[3,4], and David Sarrut[1,2]

[1] CREATIS, CNRS, Univ of Lyon, France
[2] Centre Léon Berard, 28 rue Laennec, 69373 Lyon
[3] UMR 1030 ImmunoRadAI, Université Paris-Saclay, Institut Gustave Roussy, Inserm, Villejuif, France
[4]Department of Radiation Oncology, Gustave Roussy, Villejuif, France

Abstract
Modern radiotherapy has significantly improved cancer patient survival rates, yet challenges remain in optimizing both tumor response and quality of life. Out-of-field doses play a role potentially impacting patient overall survival, but are not well-estimated by the conventional treatment planning systems. We propose here the development of a full Monte Carlo simulation model of the Elekta VERSA HD accelerator to assess the out-of-field deposited doses. We introduce an original Compton splitting method aiming to compensate for simulation low convergence in the out-of-field region. Results emphasize that achieving a 5% statistical precision on the far out-of-field deposited dose requires 200 days of simulations on a single thread. The proposed variance reduction technique led to a 1.4 efficiency gain. This model will be used to generate out-of-field dose maps on a large database of patients that will be used to train a deep learning model for quick dose calculation [1].

1 Introduction

Photon therapy modalities, such as Intensity-Modulated Radiation Therapy (IMRT) or Volumetric Modulated Arc Therapy (VMAT), use multiple beams with distinct configurations, enabling a highly conformal dose deposition within the targeted tumor volume. This involves an expansion of the irradiated volume compared to traditional 3D-Conformal Radiation Therapy (3D-CRT) [2]. However, the Out-Of-Field (OOF) deposited dose, often considered to be less than 5% of the prescribed dose, is notably lower than that within the irradiation field, but still is around 100 mGy at 15 cm far from the 5% isodose curve [3].
These very low doses can contribute to the development of radiation-induced malignancies or lymphopenia (lymphocyte deficiency), both of which reduce overall patient survival post-radiotherapy [4]. To quantify these effects and strive for the inclusion of lymphoid organs as potential organs at risk in Treatment Planning Systems (TPS), a key consideration is the precise estimation of the OOF deposited dose within the patient. However, conventional TPS are designed for in-field dose calculations and exhibit significant discrepancies when compared to experimental measurements, reaching up to 100% in the "near" OOF region [5, 6]. Consequently, several research initiatives focused on improving dose estimation in the OOF regions, employing analytical calculations, as reviewed in [7], or Monte Carlo (MC) simulations [8]. While the former method is relatively fast (taking several minutes), its precision can be outperformed by MC simulations. However, the latter approach is time-consuming, and the statistical precision to achieve is often defined from in-field simulations.
In the following, we present the statistical errors associated with the OOF deposited dose for the Elekta Versa HD, as a function of the simulation running time and introduce a Variance Reduction Method (VRM) named Compton Splitting to speed up the simulation.

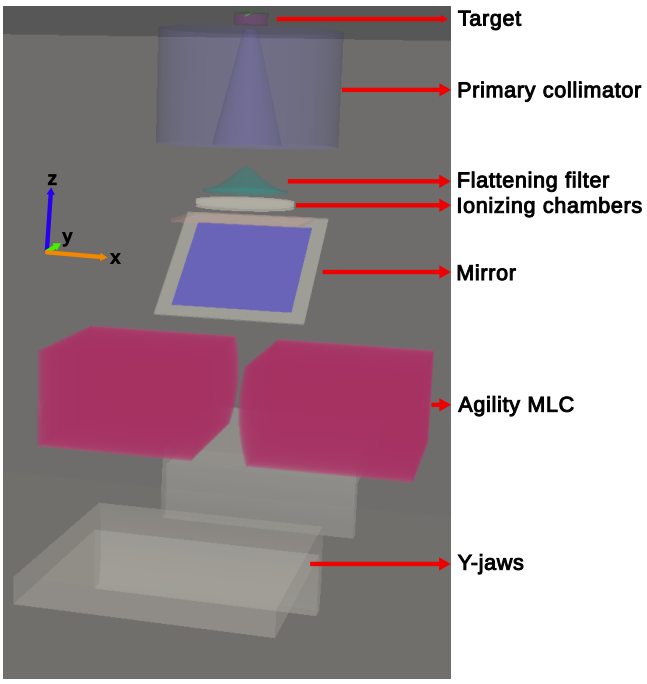


Figure 1: Simulation view of the different parts of the LINAC head. The jaws representation is here simplified. A phase space is inserted between the mirror and the MLC and acts as a particle source.

2 Materials and Methods

The MC simulation work is performed using GATE [9], an open-source software based on the GEANT4 toolkit [10], to model a LINAC head based on available information about Elekta Versa HD (Figure 1). The physics list used for the MC simulations was *G4EmStandardPhysics option3*.
The LINAC head simulation is divided into two parts. The

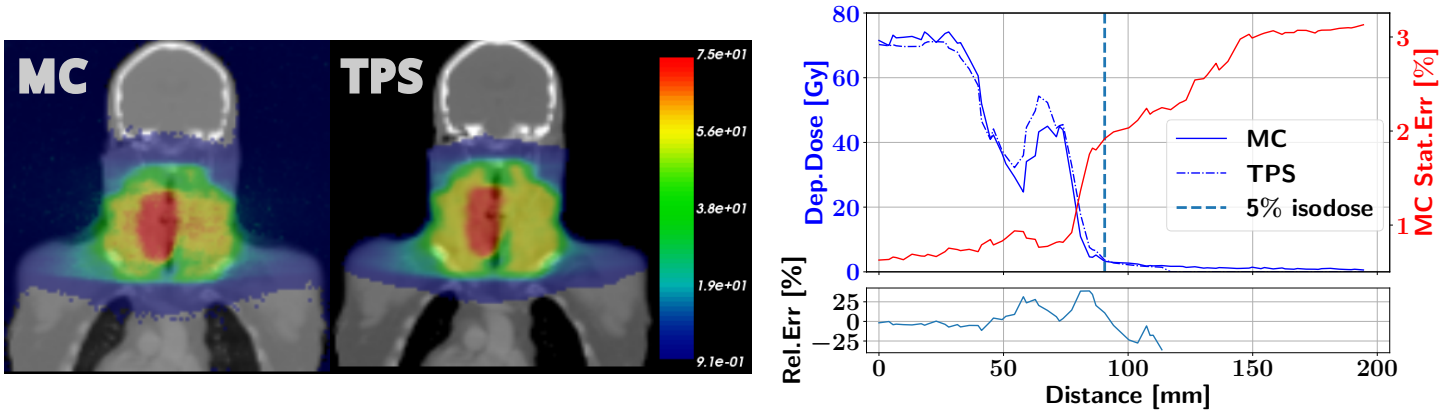


Figure 2: Left: Comparison, between the MC simulation (left) and Monaco TPS (right), of deposited doses, on a coronal plan, generated according to a DICOM RT plan. The dose is normalised at 70 Gy corresponding to the deposited dose at the LINAC isocenter position. The colormap scale is in Gray. Right: Comparison between the TPS and the MC simulation of deposited dose 1D profiles along a path starting from the isocenter to the minimum deposited dose value within the patient. The start of the OOF region (5% isodose) is depicted for visualisation purposes.

first part encompasses all patient-independent devices, including the electron target, primary collimator, flattening filter, ionizing chambers, and mirror. Particles exiting this part are stored in a phase space file and subsequently used as a particle source for the second simulation part. This part involves target-specific components, the Agility™ Multi-Leaf Collimator (MLC) and Y-jaws. The simulated LINAC head irradiates either a water tank or a voxelized CT-scan, on which a *dose actor* is attached. It enables to store of both the deposited dose and corresponding statistical uncertainties in $4\times4\times4$ mm^3 dose scoring voxels. The simulated accelerator parameters were fine-tuned to reproduce the dose deposition experimentally measured, i.e percentage depth-dose curves and dose profiles in X and Y-axis. The simulated dose well reproduces experimental data, and ongoing work involves refining simulation parameters.

The statistical assessment was conducted in a two-step process. The first step investigated, for MLC apertures from 2 to 11 cm in both axis, the impact of the three main mechanisms generating the OOF dose: (1) low-energy scattered photons within the target (patient-scattered photons, or PS photons), (2) scattered photons at the MLC/Y-jaws tips, or in the flattening filter (head-scattered photons or HS photons), and (3) photons crossing the MLC/Y-jaws (head leakage photons, or HL photons). The second step assessed deposited dose and its uncertainty as a function of the OOF distance, starting from the 5% isodose curve. The required simulation time was defined as the simulation duration, on a single thread (Intel® Core™ i9-13950HX), to achieve mean statistical uncertainties of 5% on a given dose scoring voxel. We also devised a method that reads any DICOM RT Plan to replicate a whole volumetric irradiation (e.g. few arcs, several hundreds of control points). It allows us to compute an entirely MC-based patient absorbed dose distribution, which is compared afterwards to the one obtained from Monaco TPS. Photons responsible for the OOF dose deposition (mainly HS and HL photons) are relatively rare compared to photons going in the field. Since they are mostly generated by photons undergoing Compton processes within the MLC/jaws before interacting within the patient, we decided to split Compton events occurring in the MLC/jaws regions. This splitting, coupled with Russian roulette for scattered photons with backward direction angles relative to the patient, is confined to the first two Compton processes of the particle (either primary or split). Each split particle is then conventionally tracked, with an associated weight bias. The simulation acceleration was assessed by computing the efficiency, denoted as ε, where $\varepsilon = \frac{1}{\sigma^2 t}$. σ^2 represents the total deposited dose variance, and t corresponds to the simulation running time. This efficiency was calculated from a box-shaped phase-space placed in the far OOF area. This efficiency is subsequently compared to the one obtained with a reference simulation without any applied bias.

3 Results

Figure 2 (left) illustrates the deposited dose within the patient according to the MC simulation (left) and the TPS (right). The MC simulation effectively reproduces a VMAT session, despite observable discrepancies in the deposited dose between the two images. A quantitative comparison is presented in Figure 2 (right), highlighting the deposited dose along a path starting from the isocenter to the coordinates of the minimum deposited dose, within the patient. Along this path, the difference in the in-field dose deposition is less than 5% near the isocenter but can reach up to 35% in the dose gradient. It has to be noted that here, below 1 Gy, the TPS no longer calculates the deposited dose.

The phase space emits $\sim 10^{11}$ γ, with $\sim$ 50000 particles per second simulated, requiring 23 days of simulation on a single thread. The resulting OOF deposited dose uncertainties, 10 cm beyond the 5% dose limit, remain low, around 3%.

Figure 3 provides a summary of the required simulation running time to achieve a 5% statistical uncertainty in $4\times4\times4$

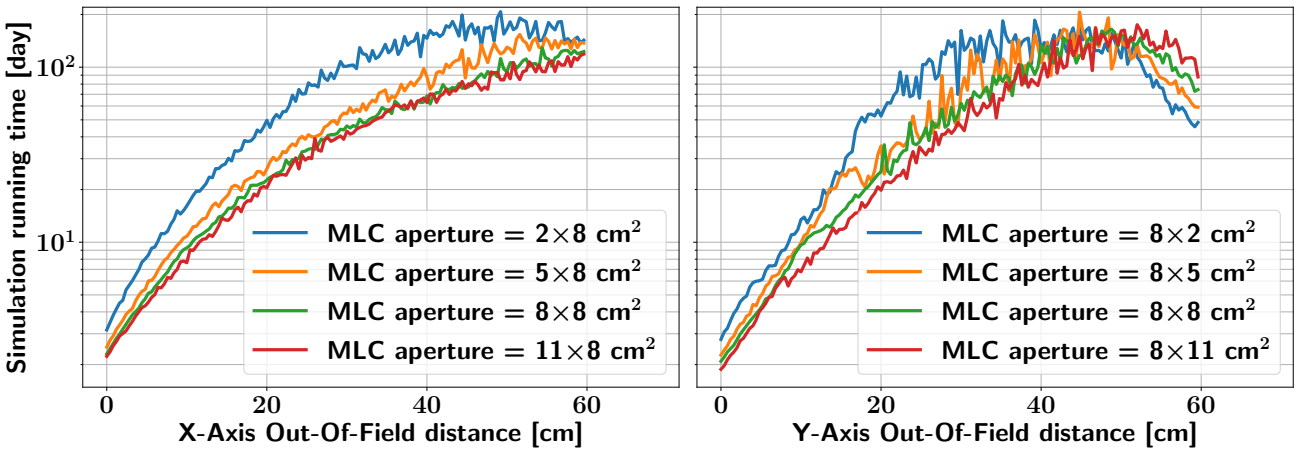


Figure 3: Required simulation running time to achieve 5% of statistical uncertainties within the OOF region, as a function of the MLC aperture and OOF distance from the 5% isodose curve, along the X-axis (left) and the Y-axis (right).

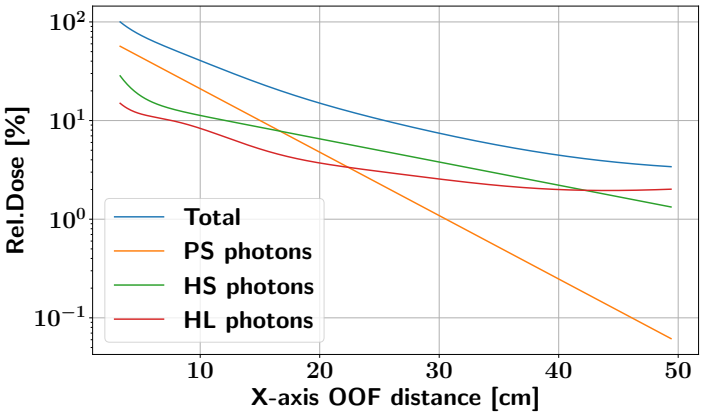


Figure 4: Individual component contributions to the OOF deposited dose plotted against the X-axis OOF distance for a MLC aperture of 8×8 cm^2. Each process is fitted based on its distinct behavior and the fit is presented solely to illustrate the trends within each process. The patient-scattered (PS) contribution is represented in orange, the head-scattered (HS) contribution in green, and the head-leakage (HL) contribution in red.

mm^3 voxels. The required time is plotted against the out-of-field (OOF) distance along the X-axis (left plot) and the Y-axis (right plot). The results are presented for the different MLC apertures simulated (from 2 to 11 cm). As expected, the simulation running time depends on the OOF distance (Figure 3).

In the near OOF region (approximately 0 - 15 cm), the dominant contribution comes from patient-scattered photons and the simulation duration peaks at a maximum of around 10 days (Figure 4). In the middle OOF region (approximately 15-40 cm), PS photons are primarily absorbed within the patient, and the dose is deposited through head-scattered photons and head-leakage photons. The deposited dose decreases and leads to an augmentation of the simulation running time to a maximum of 100 days. In the far OOF region (> 40 cm), the deposited dose increases because of photons passing beyond the collimation system. Since they are not interacting in the collimation system, they enhance the photon statistics, potentially reducing the simulation running time.

Splitting factor	q_D	$\frac{\varepsilon}{\varepsilon_{ref}}$
5	0.97 ± 0.08	1.33 ± 0.04
10	0.98 ± 0.07	1.39 ± 0.04
20	0.98 ± 0.06	1.44 ± 0.04
100	0.95 ± 0.11	1.43 ± 0.06

Table 1: Efficiency gain of the Compton splitting method for far OOF deposited dose. q_D is the quality in terms of deposited dose. q_D = 1 means that the deposited dose estimator is unbiased. ε_{ref} is the efficiency of the reference simulation.

Table 1 presents an unbiased deposited dose estimator (q_D close to one), with a modest simulation efficiency improvement, reaching up to 44%, in comparison to the reference simulation. The efficiency plateaus around a splitting factor of 10.

For higher values, all photons entering the collimation system consistently pass through it, ensuring no waste in tracking primary photons stopped by the collimators. As a result, all simulated events have the potential to attain the OOF region, contributing to the deposited dose. While a further increase in the splitting factor augment the number of split tracks passing through the collimation system, it concurrently extends the time spent in the collimation system per event. Hence, increasing the splitting factor does not enhance simulation efficiency further.

4 Discussion

We presented a model to evaluate the simulation running time required to achieve a specific statistical uncertainty (5%) in the OOF region. The trend of the curves aligns with the current state-of-the-art [11, 12].

The dose distribution generated by the MC simulation, which

replicates the radiotherapy session, exhibits differences when compared to the dose map calculated by the TPS, and suggesting a potential need for further refinement in the MC model. However, even though a simulation is fine-tuned according experimental data, previous research highlighted that large discrepancies peaking to 30% may remain with respect to the TPS [13].

In the far OOF regions, the required simulation running time peaks at 200 days for a single thread. This time can be significantly reduced, by employing multi-job simulations (about one day on our cluster). Despite this acceleration, training a deep-learning algorithm still demands the computation of around hundred images. Creating such a patient dataset, even with the support of a cluster, remains a time-consuming process taking several months. This highlights the need to strive for efficiency improvements in Monte Carlo simulation performance.

This statistical assessment will also provide a reference for calculation efficiency in OOF regions. This efficiency will serve as a ground truth for estimating potential gains in efficiency through the application of variance reduction methods [14]. A preliminary evaluation indicates that, even if allowing a 1.4 efficiency gain, Compton splitting alone may not be dedicated or sufficiently effective to enhance simulation efficiency. Alternative pseudo-deterministic VRMs, as SeTLE [15], or deep learning-based techniques to denoise low-statistic Monte Carlo dose distribution maps [16] could improve further simulations efficiency.

5 Conclusion

This study introduces a MC simulation model for the Elekta Versa HD LINAC using GATE. The model is used to generate a dose map according to a DICOM RT-plan, and is then employed to estimate simulation running times in the OOF region for various collimation system apertures. It was shown that achieving a precise simulation with a 5% statistical uncertainty demands around 200 days of simulation on a single thread. The proposed Compton splitting method allowed to 1.4 efficiency gain. These data can serve as a benchmark to improve MC simulation efficiency or for dataset creation to train deep-learning algorithms tailored to LINAC applications.

References

[1] N. Benzazon and C. Robert. “An innovative and versatile deep learning approach to estimate the out-of-field dose.” *ESTRO 2023* (2023), PD–0169.

[2] E. J. Hall and C.-S. Wuu. “Radiation-induced second cancers: the impact of 3D-CRT and IMRT”. eng. *International Journal of Radiation Oncology, Biology, Physics* 56.1 (May 2003), pp. 83–88. DOI: 10.1016/s0360-3016(03)00073-7.

[3] D. Saint-Hubert, F. Suesselbeck, F. Vasi, et al. “Experimental validation of an analytical program and a Monte Carlo simulation for the computation of the far out-of-field dose in external beam photon therapy applied to pediatric patients”. *Frontiers in Oncology* (2022). Publisher: Frontiers, p. 3189.

[4] N. Joseph, A. McWilliam, J. Kennedy, et al. “Post-treatment lymphocytopaenia, integral body dose and overall survival in lung cancer patients treated with radical radiotherapy”. en. *Radiotherapy and Oncology* 135 (June 2019), pp. 115–119. DOI: 10.1016/j.radonc.2019.03.008.

[5] N. S. Shine, R. Paramu, M. Gopinath, et al. “Out-of-Field Dose Calculation by a Commercial Treatment Planning System and Comparison by Monte Carlo Simulation for Varian TrueBeam®”. *Journal of Medical Physics* 44.3 (2019), pp. 156–175. DOI: 10.4103/jmp.JMP_82_18.

[6] R. M. Howell, S. B. Scarboro, S. F. Kry, et al. “Accuracy of out-of-field dose calculations by a commercial treatment planning system”. en. *Physics in Medicine & Biology* 55.23 (Nov. 2010), p. 6999. DOI: 10.1088/0031-9155/55/23/S03.

[7] N. Benzazon, J. Colnot, F. de Kermenguy, et al. “Analytical models for external photon beam radiotherapy out-of-field dose calculation: a scoping review”. *Frontiers in Oncology* 13 (May 2023), p. 1197079. DOI: 10.3389/fonc.2023.1197079.

[8] S. F. Kry, U. Titt, F. Pönisch, et al. “A Monte Carlo model for calculating out-of-field dose from a Varian beam”. en. *Medical Physics* 33.11 (2006). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1118/1.2360013, pp. 4405–4413. DOI: 10.1118/1.2360013.

[9] D. Sarrut, N. Arbor, T. Baudier, et al. “The OpenGATE ecosystem for Monte Carlo simulation in medical physics”. en. *Physics in Medicine & Biology* 67.18 (Sept. 2022). Publisher: IOP Publishing, p. 184001. DOI: 10.1088/1361-6560/ac8c83.

[10] S. Agostinelli, J. Allison, K. Amako, et al. “Geant4—a simulation toolkit”. en. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 506.3 (July 2003), pp. 250–303. DOI: 10.1016/S0168-9002(03)01368-8.

[11] C. W. Schneider, W. D. Newhauser, L. J. Wilson, et al. “A physics-based analytical model of absorbed dose from primary, leakage, and scattered photons from megavoltage radiotherapy with MLCs”. en. *Physics in Medicine & Biology* 64.18 (Sept. 2019). Publisher: IOP Publishing, p. 185017. DOI: 10.1088/1361-6560/ab303a.

[12] P. Hauri, R. A. Hälg, J. Besserer, et al. “A general model for stray dose calculation of static and intensity-modulated photon radiation”. en. *Medical Physics* 43.4 (2016), pp. 1955–1968. DOI: 10.1118/1.4944421.

[13] S. Gholampourkashi, J. E. Cygler, J. Belec, et al. “Monte Carlo and analytic modeling of an Elekta Infinity linac with Agility MLC: Investigating the significance of accurate model parameters for small radiation fields”. en. *Journal of Applied Clinical Medical Physics* 20.1 (2019), pp. 55–67. DOI: 10.1002/acm2.12485.

[14] V. Champciaux. “Calcul accéléré de la dose périphérique en radiothérapie”. fr. PhD thesis. Université Paris-Saclay, Jan. 2021.

[15] F. Smekens, J. M. Létang, C. Noblet, et al. “Split exponential track length estimator for Monte-Carlo simulations of small-animal radiation therapy”. eng. *Physics in Medicine and Biology* 59.24 (Dec. 2014), pp. 7703–7715. DOI: 10.1088/0031-9155/59/24/7703.

[16] S. Martinot, N. Bus, M. Vakalopoulou, et al. “Weakly supervised 3D ConvLSTMs for Monte-Carlo radiotherapy dose simulations”. en (2021).

Deformable Voxel Geometry for Photon and Electron Monte Carlo Dose Calculations

Björn Zobrist[1], Hannes A Loebner[1], Jenny Bertholet[1], Daniel Frei[1], Werner Volken[1], Florian Amstutz[1], Marco FM Stampanoni[2], Peter Manser[1], and Michael K Fix[1]

[1]Division of Medical Radiation Physics and Department of Radiation Oncology, Inselspital, Bern University Hospital, and University of Bern, Bern, Switzerland
[2]Institute for Biomedical Engineering, ETH Zürich and PSI, Villigen, Switzerland

Abstract

Purpose: To account for dynamically deforming anatomy during photon and electron dose calculation by implementing and validating a deformable voxel geometry for Monte Carlo dose calculations: DefVoxMC.
Materials and Methods: DefVoxMC is implemented in the Swiss Monte Carlo Plan framework and validated for non-rigid deformations for photon and electron beams with a compression test and a Fano test. DefVoxMC is used for a clinically motivated case with a breathing motion, and the dose differences between the reference (static) and deformed situations are assessed.
Results: The compression and Fano tests were successful. For the clinically motivated case, areas with dose differences of up to 1.4 Gy for electron beams and 6.9 Gy for photon beams are observed between the static and deformed situations with a prescribed dose of 50 Gy.
Conclusion: DefVoxMC is successfully implemented and validated for photons and electrons and can be used to consider deformations in clinically motivated cases.

1 Introduction

A crucial aspect of radiation therapy is dose calculation, aiming to accurately estimate the delivered dose. However, the dose calculation is usually performed using a static CT, while the dose is delivered to a patient with a dynamically deforming anatomy. Deformations can happen during the dose delivery, e.g. breathing motion, or between fractions, e.g. weight loss, tumor regression or progression.

A promising approach to consider deformations for dose calculations is to use a deformable voxel geometry for Monte Carlo (MC) simulations. MC simulations are considered the gold standard for dose calculation in radiation therapy [1]. The deformable voxel geometry consists of dodecahedron-shaped voxels (Figure 1) designed so that each voxel can be approximately deformed according to the deformation in the anatomy, quantified by a deformation vector field (DVF) [2]. The underlying assumption is that each voxel corresponds to the same mass element in every anatomical shape. Using this method, the geometry of the planning CT can be deformed to match the anatomical shape at other time points during the treatment as investigated by Heath et al. [3]

In this work, we implemented and validated a deformable voxel geometry for MC dose calculations to consider dynamically deforming anatomy during photon and electron dose calculations. The deformable voxel geometry extends the in-house Swiss Monte Carlo Plan (SMCP) [4,5], a framework that can be used for different particle transport algorithms for photon and electron beams.

2 Materials and Methods

2.1 Implementation and Validation

For the deformable voxel geometry, the six rectangular sides of regular voxels are divided into two triangles, resulting in a dodecahedron-shaped voxel as in Peterhans et al. [2] This shape allows independent shifting of every corner point, as depicted in Figure 1.

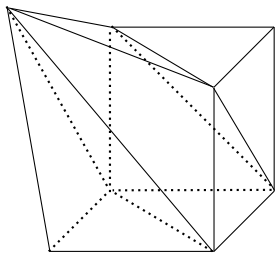


Figure 1: Dodecahedron-shaped voxel with one corner point shifted.

The mass and material of the voxels are set to be conserved with deformation. Therefore, as the volume of the voxels changes with deformations, the density is adapted. Within SMCP different particle transport algorithms can be used with any geometry. Available transport algorithms are VMC++ [6], EGSnrc [7], and other algorithms; in this work, the results are calculated with VMC++ except for the Fano test, which is calculated with EGSnrc. While the particle transport code simulates the interactions, it relies on the geometry code to calculate the path of the particles through the geometry. The geometry code is adapted for the deformable voxel geometry, considering the more complicated dodecahedron-shaped voxels.
The particle transport takes place in the deformed geometry, simulating the dose calculation in the deformed situation while scoring the dose in the geometry of the original situation. This MC dose calculation, using the deformable voxel geometry, is referred to as DefVoxMC in this work.

DefVoxMC is validated with a compression test, using non-rigid deformations on a cubic phantom with a density of 1.0 g/cm^3 for photons (0.2 g/cm^3 for electrons, as otherwise their range is too short). The compressed phantom has a density of 1.25 g/cm^3 for photons (0.25 g/cm^3 for electrons). Dose calculations are performed for a 6 MV photon field and a 6 MeV electron field using a voxel size of 0.125^3 cm^3. A first MC dose calculation is performed on both phantoms

using a regular voxel geometry. Next, the geometry of the original phantom is deformed to match the compressed phantom during the dose calculation using DefVoxMC. In both cases, the depth dose profile is calculated along the beam direction. The compression test is successful if the depth dose profile calculated with DefVoxMC on the deformed original phantom and calculated with a regular voxel geometry on the compressed phantom correspond.

Another test for the deformable voxel geometry is the Fano test [8]. The Fano Theorem states that if the atomic properties per unit mass are uniform and a uniform source (per unit mass) is used, then the fluence is also uniform. In this work, the Fano test is used with the specific conditions outlined by Tessier [8]. A homogeneous water phantom is used, and density changes are introduced by applying a deformation vector field (DVF). This DVF stretches voxels in the center of the x-direction while uniformly compressing the other voxels, conserving the overall size and ensuring a uniform atomic interaction cross-section. A uniform, isotropic, and density-scaled source with an energy of 15 MeV is used. Finally, an infinite simulation space is ensured by creating a new particle with the same direction and energy as the departing one on the opposite side of the geometry. The theoretical Fano value is calculated according to the following equation:

$$D_{Fano} = n\frac{E_0}{m}$$

Where n is the number of particle histories simulated, m is the mass of the phantom, and E_0 is the energy of the particles [8]. For the phantom with a mass of 1000 kg and 2.0×10^8 simulated particle histories, the Fano value is 4.81×10^{-7} Gy. The Fano test is successful if the resulting dose distribution is uniform and matches the theoretical Fano value.

2.2 Clinically Motivated Case

DefVoxMC is used for a clinically motivated case from the 4D Lung data set [9]. A breast case is simulated by contouring the left breast as the planning target volume (PTV). An arc mixed-beam radiotherapy (arc-MBRT) [10–12] plan is generated with two 6 MV photon arcs and two non-isocentric 9 and 12 MeV electron arcs. The dose distribution is calculated on a regular voxel grid and using DefVoxMC under breathing motion from the 4DCT.

The CT at the breathing phase 0% is chosen as the planning CT. Deformable image registration (DIR) is applied between the planning CT and the remaining nine breathing phases. This work uses the B-Splines image registration of plastimatch [13] for DIR. Additionally, DIR is applied between the neighboring phases, those DVFs are used for interpolation. The schematic of the resulting DVFs with one interpolation point between phases is shown in Figure 2, but more interpolation points between the phases are possible for a higher time resolution.

The vectors of the resulting DVFs are applied to shift the corner points of the deformable voxel geometry of the planning CT to match the deformed situation at every time point. In the particle simulation, the monitor unit (MU)-fraction of a particle is sampled, as the beam parameters are given as a function of MU-fraction. To match the breathing motion to the beam parameters, the number of MUs delivered per time is estimated. This is done using the limiting linac axes, such as gantry speed, maximum dose rate, etc. Using this information the timestamp of each sampled particle is determined. The closest neighboring DVF is used to deform the geometry, and the whole particle history is simulated. This process is repeated for each sampled particle and is depicted schematically in Figure 2. The resulting dose distribution is on the planning CT (CT at phase 0%).

The difference between the dose distributions with a breathing motion and for the static situation is evaluated. The calculation times and RAM usage are assessed for one electron and one photon arc.

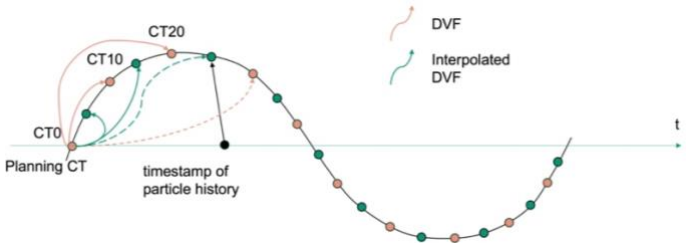


Figure 2: Schematic of DefVoxMC. Light red dots depict the original 4DCT, while green dots are interpolated CTs. The CT at phase 0% is chosen as the planning CT, and the DVFs point from this CT to the remaining CTs. The black dot represents the timestamp of a sampled particle, and the geometry is deformed with the nearest DVF. The y-axis represents the deformation due to the breathing motion. The arrows depict the DVFs.

3 Results

The results of the compression test are shown in Figure 3. The relative dose differences between the dose on the compressed phantom and the deformed original phantom are within 1% for electron and photon fields, indicating that DefVoxMC dose calculations in deformed phantoms are accurate for both particle types.

The results of the Fano test are shown in Figure 4. The dose is averaged over the y- and z-direction so that the difference in mean dose along the direction of the deformation can be evaluated. The relative dose difference to the theoretical Fano value is less than 1‰, indicating that the dose calculation with DefVoxMC is performed correctly.

The absolute dose difference between the dose distribution with a breathing motion and the static situation for the clinically motivated case is shown in Figure 5. The figure shows the electron (left) and photon (right) contributions of the arc-MBRT plan separately. The overall dose distribution is normalized to cover 50% of the PTV with the prescribed dose of 50 Gy. It is important to note that the

electron and photon contributions differ. The electrons contribute 28% to the target coverage, while the photons contribute 72%. There are areas with absolute dose difference of up to 1.4 Gy for the electrons and 6.9 Gy for the photons.

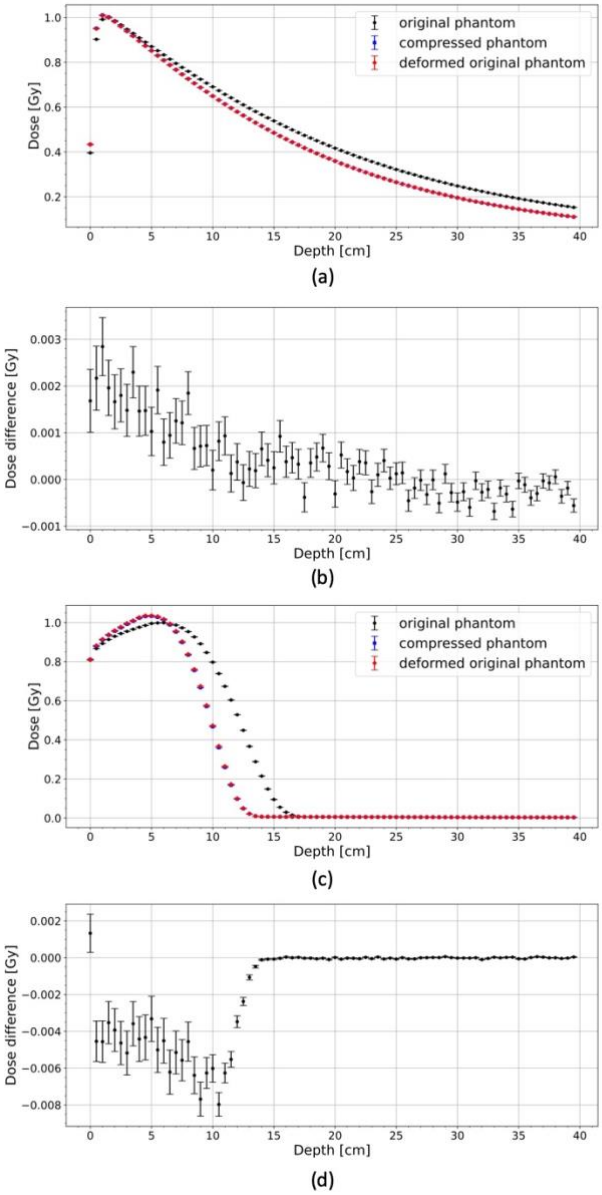


Figure 3: Results for dose calculations on phantoms with different densities. Only every fourth data point is depicted for readability. Error bars represent the standard deviation of the mean. (a) Depth dose profile of 6 MV photon field. (c) Depth dose profile of 6 MeV electron field. (b), (d) Difference between dose on compressed phantom and deformed original phantom.

The dose calculation is parallelized on the UBELIX HPC cluster, and 10^8 particle histories are simulated for the results above. The mean relative dose error for voxels with a relative dose bigger than 50% is 0.59%.

To compare calculation times only one CPU is used, and 10^5 particle histories are simulated. Table 1 shows the calculation times for one electron and one photon beam to compare the performance of DefVoxMC to a regular voxel geometry. The preparation includes the loading and interpolation of the DVFs. The particle simulation time for electron beams is 1.2 times longer for deformable compared to regular voxel geometry; for photon beams, it is 1.8 times longer. For the clinically motivated case, 6 GB of RAM was needed to store 50 DVFs for the 10 phases and 4 interpolation points between the phases.

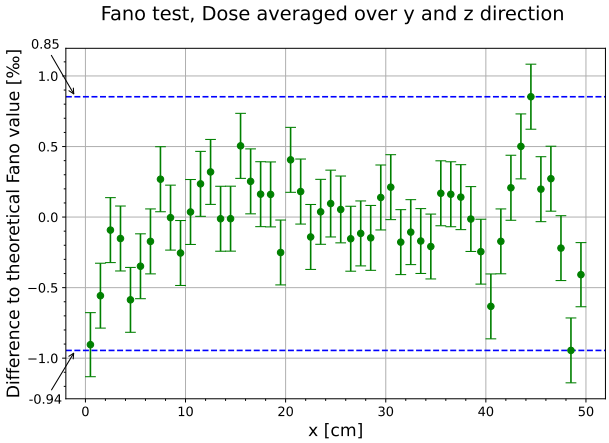


Figure 4: Dose difference along the x-direction averaged over y- and z-direction for the Fano test. The blue lines indicate the maximal dose difference to the theoretical Fano value. Error bars represent the standard deviation of the mean.

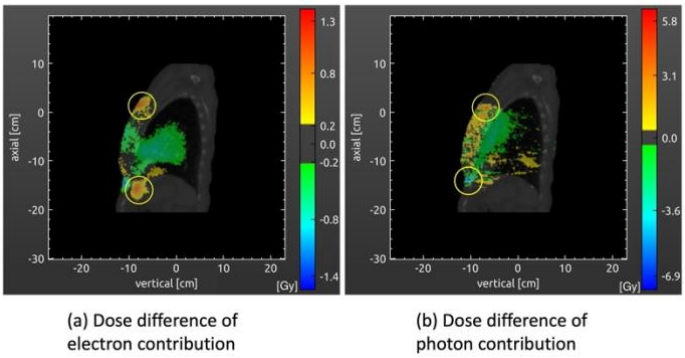


Figure 5: Absolute dose difference between the dose distribution with a breathing motion and the static situation in Gy in a sagittal slice. The mean relative dose error for voxels with a relative dose bigger than 50% is 0.76% for the photon contribution and 0.53% for the electron contribution. Yellow circles highlight areas with the largest absolute dose difference.

Table 1: Calculation times for photon and electron beams

	Preparation [min]	DefVoxMC [min]	Regular Voxel [min]
Electrons	3:10	2:45	2:15
Photons	3:10	1:55	1:05

4 Discussion

A deformable voxel geometry, DefVoxMC is successfully implemented and integrated within the SMCP framework. DefVoxMC is validated for non-rigid deformation with a compression test and a Fano test. The agreement of the compression test is within 1%, similar to the results of Heath et al. [3] In this work, the deformable voxel geometry approach is applied to electron beams for the first time. We further demonstrate the application of DefVoxMC to a

dynamically deforming situation for a clinically motivated treatment plan combining photon and electron arcs.
The calculation time compared to MC with a regular voxel geometry is increased due to the more complicated geometry methods of DefVoxMC. The geometry methods are called less often for electrons than for photons because electrons have a limited range. Therefore, the time increase between regular voxel geometry and DefVoxMC is smaller for electron than for photon beams.

The underlying assumption of the deformable voxel geometry is that each voxel corresponds to the same mass element in the original and the deformed situation. Other approaches like dose warping [14,15] and energy remapping [16] calculate the dose distribution on the deformed situation and apply the DVF to the dose or energy, mapping it back to the planning CT. However, the DVF holds the information of how the anatomy deforms and not the dose distribution. Unlike dose warping and energy remapping, for DefVoxMC, the DVF directly deforms the anatomy.
As a limitation of the study, the underlying assumption of each voxel corresponding to the same mass element in the original and deformed situation may not be fulfilled in certain situations. DefVoxMC could be used for adaptive radiotherapy to accumulate the daily dose on the planning CT by performing DIR on the daily CT. However, if the patient loses weight and some mass elements disappear, the assumption is not fulfilled anymore. Also, for the breathing motion, the mass of the inhaled air might influence the results. Another limitation is the DIR itself. DefVoxMC strongly relies on the quality of the DVF, but it is known that different DIR algorithms produce different DVFs [17], ultimately resulting in differences in the calculated dose distributions. To have accurate results on clinical cases, the quality of the DVF must be further investigated.

5 Conclusion

The SMCP framework is successfully extended with a deformable voxel geometry for dose calculations on deforming anatomy. DefVoxMC is validated with a compression test and a Fano test for photon and electron beams.
DefVoxMC can be used for clinical cases to consider dynamically deforming anatomies.

Acknowledgments

This work was supported by Varian, a Siemens Healthineers Company. Dose calculations and optimizations were performed on UBELIX (wwww.id.unibe.ch/hpc), the high-performance cluster at the University of Bern.

References

[1] Krieger T, Sauer OA. Monte Carlo- versus pencil-beam-/collapsed-cone-dose calculation in a heterogeneous multi-layer phantom. Phys Med Biol, vol. 50, 2005, p. 859–68. https://doi.org/10.1088/0031-9155/50/5/010.

[2] Peterhans M, Frei D, Manser P, Aguirre MR, Fix MK. Monte Carlo dose calculation on deforming anatomy. Z Med Phys 2011;21:113–23. https://doi.org/10.1016/j.zemedi.2010.11.002.

[3] Heath E, Seuntjens J. A direct voxel tracking method for four-dimensional Monte Carlo dose calculations in deforming anatomy. Med Phys 2006;33:434–45. https://doi.org/10.1118/1.2163252.

[4] Fix MK, Manser P, Frei D, Volken W, Mini R, Born EJ. An efficient framework for photon Monte Carlo treatment planning. Phys Med Biol 2007;52. https://doi.org/10.1088/0031-9155/52/19/N01.

[5] Fix MK, Manser P, Frei D, Volken W, Mini R, Born EJ. Efficient photon treatment planning by the use of Swiss Monte Carlo Plan. Journal of Physics: Conference Series OPEN ACCESS n.d. https://doi.org/10.1088/1742-6596/74/1/012004.

[6] Kawrakow Iwan and Fippel M. VMC++, a fast MC algorithm for Radiation Treatment planning. In: Schlegel Wolfgang and Bortfeld T, editor. The Use of Computers in Radiation Therapy, Berlin, Heidelberg: Springer Berlin Heidelberg; 2000, p. 126–8.

[7] Standards NRC of CanadaMRCentreIR, ionisant C national de recherches du CanadaC de recherche en métrologie. É de rayonnement. EGSnrc: software for Monte Carlo simulation of ionizing radiation 2021. https://doi.org/10.4224/40001303.

[8] Tessier F. EGSnrc update and Monte Carlo simulation verification n.d. https://people.na.infn.it/~mettivie/MCMA%20presentation/16%20Aula%20Magna%201%20sez/9_15_Tessier_ID_214.pdf (accessed January 4, 2024).

[9] Hugo GD, Weiss E, Sleeman WC, Balik S, Keall BJ, Lu J, et al. Data from 4D Lung Imaging of NSCLC Patients (Version 2) [Data set]. The Cancer Imaging Archive 2016.

[10] Guyer G, Mueller S, Mackeprang P-H, Frei D, Volken W, Aebersold DM, et al. Delivery time reduction for mixed photon-electron radiotherapy by using photon MLC collimated electron arcs. Phys Med Biol 2023;68:215009. https://doi.org/10.1088/1361-6560/ad021a.

[11] Mueller S, Fix MK, Joosten A, Henzen D, Frei D, Volken W, et al. Simultaneous optimization of photons and electrons for mixed beam radiotherapy. Phys Med Biol 2017;62:5840. https://doi.org/10.1088/1361-6560/aa70c5.

[12] Mueller S, Fix MK, Henzen D, Frei D, Frauchiger D, Loessl K, et al. Electron beam collimation with a photon MLC for standard electron treatments. Phys Med Biol 2018;63:25017. https://doi.org/10.1088/1361-6560/aa9fb6.

[13] Sharp G, LI R, Wolfgang J, Chen G, Peroni M, Spadea M, et al. Plastimatch – an open source software suite for radiotherapy image processing., 2010.

[14] Moriya S, Tachibana H, Kitamura N, Sawant A, Sato M. Dose warping performance in deformable image registration in lung. Physica Medica: European Journal of Medical Physics 2017;37:16–23. https://doi.org/10.1016/j.ejmp.2017.03.016.

[15] E Heath, J Seco, Z Wu, G C Sharp, H Paganetti, J Seuntjens. A comparison of dose warping methods for 4D Monte Carlo dose calculations in lung. J Phys Conf Ser 2008;102:012013. https://doi.org/10.1088/1742-6596/102/1/012013.

[16] Siebers J V, Zhong H. An energy transfer method for 4D Monte Carlo dose calculation. Med Phys 2008;35:4096–105. https://doi.org/https://doi.org/10.1118/1.2968215.

[17] Fatyga M, Dogan N, Weiss E, Sleeman WC, Zhang B, Lehman WJ, et al. A voxel-by-voxel comparison of deformable vector fields obtained by three deformable image registration algorithms applied to 4DCT lung studies. Front Oncol 2015;5. https://doi.org/10.3389/fonc.2015.00017.

A GPU-based patient-specific CT dosimetry tool combining Monte Carlo calculations and ML-based segmentations

Ronan Lefol[1], Yannick Lemaréchal[1], Jonathan Boivin[2], and Philippe Després[1,2]

[1]Département de physique, de génie physique et d'optique, Université Laval, Québec, Canada
[2]Service de radio-oncologie et Centre de recherche CHU de Québec-Université Laval, Québec, Canada

Abstract The continuous growth of Computed Tomography (CT) examinations underscores the pressing need for enhanced personalized dose estimation in medical imaging. Existing methodologies, reliant on either the CT Dose Index (CTDI) or anthropomorphic models, provide estimations with limited consideration of the patient's anatomy. While patient specific Monte Carlo simulations are recognized as the gold standard, they remain prohibitively time-consuming and remain unfit for massive, population-scale calculations. This study aims to extend the GPU Monte Carlo Dose (GPUMCD) software for clinical CT acquisitions and showcasing fast dose-to-organ extraction facilitated by automated segmentation tools. Doses simulated within anthropomorphic phantoms closely approximated measured values (within 7%) for various scenarios. Study of an open-access CT dataset of 58 patients showed a mean organ dose variation of up to 50% from the cohort average with identical scanning parameters. The personalized dose pipeline was executed on a single A100 40GB GPU for a runtime of 5 to 7 minutes. Notably, the segmentation task took an average of 140 seconds and the GPUMCD simulation, 71 seconds. The runtime was further reduced to an average of 4 minutes when employing 2 GPUs in parallel, use of more GPUs provided limited improvement to the pipeline as the CPU bound pre- and post-pocessing steps became the predominant computation cost. This works demonstrates the ability to generate a DICOM compliant, three-dimensional dose distributions, segmentations, and dose reports for CT acquisitions within minutes. This pipeline lays the groundwork for swift, personalized CT dosimetry and facilitates large-scale dose-to-organ studies.

1 Introduction

X-ray Computed Tomography (CT) is a diagnostic imaging method with relatively high radiation exposure to the patient. The clinical usage of CT machines has seen a rapid increase in use, causing a significant increase in medical radiation exposure per capita. Current means of dose estimations are performed through indices such as the Dose Length Product (DLP). Such indices were designed to be a standardized measure of the radiation output of a CT system and not a measure of individual patient dose. Studies have outlined that estimations of patient dose and cancer risks through such generic representations can carry an uncertainty of ±40% and may grow to a factor of ±5 when used to determine any one individual's effective dose or cancer risk [1, 2].
MC simulations are notoriously slow to reach an adequate level of accuracy due to the significant amount of required calculations (histories). Such tools have previously been validated for tasks such as CT imaging simulations. However, due to the long computation time, MC simulations have not been used in routine operations despite various improvements to MC calculations. Since the turn of the century, new adaptations of MC software have been developed for the Graphics Processing Unit (GPU) offering significant improvement in computation time. The GPU Monte Carlo Dose (GPUMCD) [3, 4] software platform, for example, was utilized for External Beam Radiation Therapy [3, 4], Low Dose Rate (LDR) and High Dose rate (HDR) brachytherapy [5–7], Internal Radiation Therapy (IRT) [8], and proton therapy [9].

At the core of any dose-to-organ examination is the need to have accurate and precise organ/region segmentation. This contouring task is typically done by experts, and remains a very time-consuming and highly repetitive task. Organ identification in CT modality is done through anatomical knowledge and recognition of image contrast values. Contrary to other types of images, the gray values on CT images carry a physical meaning. Due to the very similar atomic composition of many organs, the Hounsfield Units (HU) values often fall within the same range, that is, between 0 and 100 HU [10]. In scenarios where such organs are nearby, or where their boundaries are unclear, it becomes increasingly difficult to have a clean and consistent contour.

Despite this ever-increasing need for quality contouring, a proper standard of validation for organ delimitation has yet to be developed [11]. A previous study on inter-observer variation in segmentation of brain ROI showed inter-observer scoring of as low as 0.4 Dice score (DSC) for the chiasm and nerves, but above 0.8 DSC for the brainstem and eyes [12]. This broad variation in segmentation brings in turn significant variations in measurements or metrics which derive from the volume definition. Metrics of organ dose would be affected, as these are typically reported as a measure of the volume of the ROI, as in the case of the Dose Volume Histogram (DVH).

This study aims to present a new software pipeline, exploiting the power of automatic segmentation tools to provide consistent and accurate organ segmentations, and a GPU-enabled MC simulation through GPUMCD for patient-specific dose deposition originating from CT imaging. This pipeline allows patient-specific dose-to-organ measurements to be performed in a completely automated fashion, utilizing only the parameters present within the provided DICOM study.

2 Methods

A new CT extension was developed for GPUMCD, incorporating parameters such as direction, pitch, and current of the X-ray source, originating directly from the DICOM CT-examination files. Material and density assignment of the simulated patient volume were performed using standard calibration procedure for CT volumes.
Experimental validation of the new GPUMCD extension involved standard weighted CT Dose index ($CTDI_w$) and in-air ($CTDI_{100,air}$) measurements under known conditions, comparing results to simulation outcomes. Measurements were conducted with varied pitch factors and volume sizes. A series of complete CT acquisitions were additionally taken to perform dose calibration. Anthropomorphic head volume measurements were also performed using the Stereotactic End-to-End Verification Phantom (STEEV), (CIRS, Norfolk, USA) to assess dose deposition in more intricate geometries.
To assess GPUMCD's performance, 58 chest CT studies from the open dataset *Lung CT Segmentation Challenge 2017* (LCTSC) were simulated using three different NVIDIA GPUs: RTX 2070 super, RTX 3090, A100-SXM4-40GB. The recorded metrics included the average time required for 10^6 histories and a requirement of achieving a maximum uncertainty of 2% within the volume. Each study was run with 10^{10} primary histories, and a subsequent average runtime for 10^6 primary histories was determined. Additionally, the simulations were tested on an NVIDIA DGX platform equipped with multiple A100 (40GB) GPUs, leveraging GPUMCD's parallel GPU computation capabilities.
GPUMCD was containerized in a Docker [13] image, requiring as input a single DICOM CT series. Dose depositions were simulated with GPUMCD for each CT series in the LCTSC dataset [14] until the maximum uncertainty reached 2% (excluding in-air dose). An additional layer of data pre- and post-processing was added to the Docker images. The pre-processing steps extract simulation parameters from the DICOM study headers such as the kVp, tube current and modulation, collimation, pitch, and imaged volume size and resolution. The imaged volume is also converted into an egsphant compliant file format for ingestion by GPUMCD. The resulting dose and uncertainty volumes from the simulation were stored in associated DICOM RTDOSE files and a DICOM SR was produced with simulation parameters, GPUMCD version, and console logs as part of the post-processing steps.
Within the DICOM tags of the LCTSC study, the information for the tube current, tube potential, and spiral pitch factor were absent. In order to perform an initial analysis on a small cohort, the simulations for this dataset were all set to have a tube potential of 120 kVp and a constant tube current of 100 mAs with a pitch factor of 1. Hard-coding these values for these 58 patients made a *true* measure of deposited dose impossible, but enabled a cursory examination of dose deposition independent of the typical major dose drivers from scanning parameters and dependent on mainly patient anatomy and positioning.
The segmentation tools, TotalSegmentator V2 [15, 16] and Lungmask [17] were similarly containerized in individual Docker images. Each image was in turn used to segment LCTSC dataset. The resulting segmentations were stored in DICOM SEG or RTStruct format, and a DICOM Structured Report (SR) was generated for each segmentation task with detailed information on the provided parameters, segmentation tool and version, and application console logs.
For each CT study the associated segmentation and dose grids were overlaid in order to extract dose-to-organ metrics. The resulting metrics were stored in a DICOM SR file.

3 Results

3.1 Phantoms

The simulated and measured doses for a full abdomen protocol of a 32 cm PMMA CTDI volume are displayed in Table 1 along with the two measurement positions within the STEEV phantom. Simulated values varied by a maximum of 6.6% from measurement values. This falls within the experimental measurement uncertainties of 7.4%, following scenario 2 of IAEA-TRS 457 section 8.3.2 [18]. Similar comparisons were observed in the STEEV phantom with 80 kVp, 100 kVp, and 140 kVp tube potentials with a maximum of 7.2%, 6.4%, and 6.8% difference respectively.

Phantom	Measured (mGy)	Simulated (mGy)	Δ_{dose} (%)
PMMA 32cm center	5.22	5.26	0.6
PMMA 32cm mean periphery	14.75	15.58	5.5
STEEV isocenter	20.56	21.84	6.2
STEEV anterior	20.43	21.79	6.6

Table 1: Simulated and measured average dose values for a constant 100 mA tube current over a full volume scan for 120 kVp tube potential. A spiral pitch factor of 1 was used for the PMMA volumes, and 0.6 for STEEV.

3.2 Patients

Mean organ dose values are shown in Figure 1, illustrating results obtained using the LungMask and TotalSegmentator segmentation tools, alongside the ground truths from the LCTSC dataset. Only organs segmented by at least two of the methods are presented in the figures. Notably, a consistent mean organ dose is observed for either lung, with the maximum difference not exceeding 0.5 mGy.
Clear differences were observed when comparing the segmentations of the heart and spinal cord provided by the ground

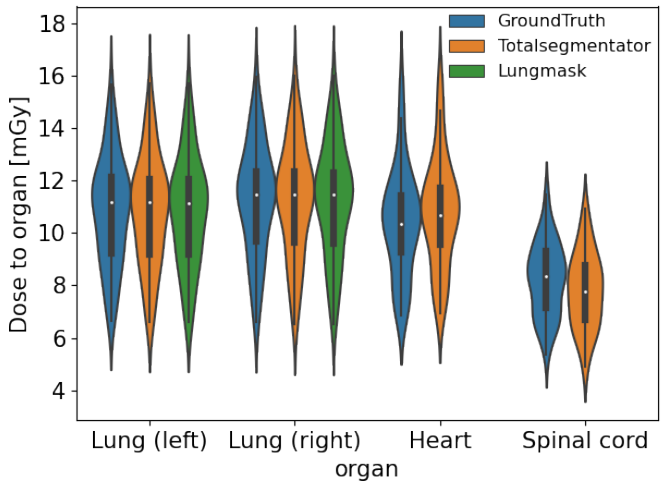


Figure 1: Mean dose to organ using GPUMCD and various segmentations.

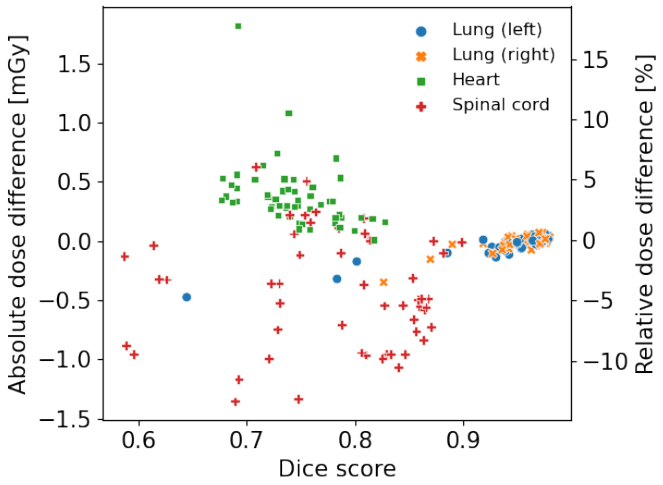


Figure 2: Mean organ dose differences yielded by Totalsegment versus LCTSC ground truth segmentations.

truth and TotalSegmentator. The spinal cord segmentations extended throughout the whole CT volume for TotalSegmentator, but were not present within the first and last few slices within the ground truth. The heart segmentation for the ground truth was a very simplified shape as compared to that provided through TotalSegmentator. These significant variations in segmentation volumes produced a maximum difference in mean dose for the heart and spinal cord of 14% and -17% respectively, with an average deviation of 3.5% and -5.5% mGy respectively.

The execution time for the pipeline's data pre-processing took on average 65 seconds; the main cause of this variation being the total amount of slices within the series. Utilizing a single A100 40GB GPU, the segmentation pipeline took an average of 140 seconds for the TotalSegmentator tool and 16 seconds for LungMask. The time to perform the simulation through GPUMCD for a maximum 2% uncertainty was on average 36 seconds. Reading parameter files and writing results for GPUMCD took an average of 34s. The extraction of organ dose values from the resulting DICOM files was performed in 8 seconds. The average runtime of the entire pipeline while operating in series ranged from 5 to 7 minutes. Utilizing 2 GPUs in parallel, the pipeline completed in an average of 4 minutes. Further GPU-parallelizing of the task will reduce the computation time of both the simulation and segmentation tasks to under a minute.

Operating the pipeline on RTX 3090 and RTX 2070S GPUs was on average 35% and 330% slower respectively, with a total computation time ranging from 7 to 10 minutes and 16 to 23 minutes respectively.

4 Discussion

In this study, an extension of the existing Monte Carlo tool, GPUMCD, was developed to accommodate modern CT machines. This extension was validated through a comprehensive comparison between simulated outcomes and measured results across diverse scenarios. Simulations were conducted for various homogeneous volume configurations, and an anthropomorphic volume was employed in the analyses.

The distribution of dose to various organs for the LCTSC patient cohort is presented in Figure 1. The figure demonstrates significant inter-patient dose variations. Organ volume and patient sex were not identified as primary sources of variation. As the pitch, tube current, and tube potential were kept constant across all patients of the LCTSC dataset, variations were predominantly attributed to factors such as patient positioning (e.g., lowered or raised arm(s)) and patient anatomy. Organ doses exhibited variations of up to $\pm 50\%$ from the cohort average. Comparisons with anthropomorphic organ dose calculations typically align with the cohort average, emphasizing the utility of such models in representing an average patient. However, these models fail to account for the unique characteristics not only of each patient but also their position during the CT examination.

The impact of the segmentation tool on the obtained mean organ dose is demonstrated in Figure 1 and Figure 2. The mean organ dose values for both lungs presented minimal dependence on the segmentation tool. However, the amount of CT slices segmented, and contour simplicity varied significantly between TotalSegmentator and the ground truth for the heart and spinal cord. This significant segmentation variation, presenting Dice scores between 0.6 and 0.9, showed greater dose variations, typically under 0.5 mGy (5%) but reaching 1.7 mGy (17%) in extreme cases.

A significant constraint in establishing a tool to consistently and accurately calculate patient dose in CT exposures has been computation time. The presented pipeline in this project addresses this challenge, making it feasible to perform hundreds of organ-dose calculations daily with reasonable hardware requirements. Each component of this pipeline is encapsulated in a self-contained Docker container, facilitating

straightforward and independent execution if needed.
This pipeline is presented with a runtime of under 7 minutes on an A100 40GB GPU with the ability to further reduce the simulation time of GPUMCD can be achieved by leveraging multiple GPUs with minimal associated overhead. With a total of 2 GPUs the runtime is reduced to an average of 4 minutes, however the use of more GPUs provides little reduced runtime. Within both the segmentation and simulation, along with the pre- and post-processing steps, are CPU bound processes which are not sped-up by additional GPUs. These CPU processes present a 1 to 2 minutes of computation time, leading to a bottleneck. While optimization and parallelization of these subtasks have the potential to enhance performance, such endeavors fall outside the scope of the present project.

5 Conclusion

The results presented in this study showcase a fast pipeline capable of leveraging diverse segmentation software and a new GPUMCD extension for CT simulation, enabling the calculation of personalized organ doses in CT examinations in under 7 minutes on a commodity GPU. The mean organ dose exhibits considerable variability between patients with identical scanning parameters, varying between 5 mGy and 18 mGy for the lungs and heart, and between 3 mGy and 12 mGy for the spinal cord.
Inter-segmentation dose variation for lungs was consistently within 0.3% with dice scores above 0.9. Segmentations for the heart and spinal cord had Dice scores varying between 0.6 and 0.9, with an organ dose variation of up to 17%.
The capability of this pipeline to rapidly generate quantitative 3D dose distributions for individual patients holds great potential for personalized protocol optimization and facilitates large-scale epidemiological studies.

References

[1] C. H. McCollough and B. A. Schueler. “Calculation of effective dose”. en. *Medical Physics* 27.5 (May 2000), pp. 828–837. DOI: 10.1118/1.598948.

[2] C. J. Martin. “Effective dose: how should it be applied to medical exposures?” en. *The British Journal of Radiology* 80.956 (Aug. 2007), pp. 639–647. DOI: 10.1259/bjr/25922439.

[3] S. Hissoiny, B. Ozell, H. Bouchard, et al. “GPUMCD : A new GPU-oriented Monte Carlo dose calculation platform”. en. *Medical Physics* 38.2 (Jan. 2011), pp. 754–764. DOI: 10.1118/1.3539725.

[4] S. Hissoiny, A. J. E. Raaijmakers, B. Ozell, et al. “Fast dose calculation in magnetic fields with GPUMCD”. en. *Physics in Medicine and Biology* 56.16 (Aug. 2011), pp. 5119–5129. DOI: 10.1088/0031-9155/56/16/003.

[5] S. Hissoiny, M. D’Amours, B. Ozell, et al. “Sub-second high dose rate brachytherapy Monte Carlo dose calculations with **bGPUMCD** : HDR dose calculations with bGPUMCD”. en. *Medical Physics* 39.7Part2 (July 2012), pp. 4559–4567. DOI: 10.1118/1.4730500.

[6] E. Bonenfant, V. Magnoux, S. Hissoiny, et al. “Fast GPU-based Monte Carlo simulations for LDR prostate brachytherapy”. en. *Physics in Medicine and Biology* 60.13 (July 2015), pp. 4973–4986. DOI: 10.1088/0031-9155/60/13/4973.

[7] C. Joachim-Paquet. “Extension et validation de bGPUMCD, un algorithme de calcul de dose Monte Carlo pour la curiethérapie implémenté sur processeur graphique”. MA thesis. Québec: Université Laval, 2022.

[8] A. Frezza, C. Joachim-Paquet, M. Chauvin, et al. “Validation of irtGPUMCD, a GPU-based Monte Carlo internal dosimetry framework for radionuclide therapy”. en. *Physica Medica* 73 (May 2020), pp. 95–104. DOI: 10.1016/j.ejmp.2020.04.010.

[9] D. Maneval. “Conception d’un formalisme de pouvoir d’arrêt équivalent et accélération graphique: des simulations Monte Carlo plus efficaces en protonthérapie”. PhD thesis. Université Laval, 2019.

[10] A. Kalra. “Developing FE Human Models From Medical Images”. en. *Basic Finite Element Method as Applied to Injury Biomechanics*. Elsevier, 2018, pp. 389–415. DOI: 10.1016/B978-0-12-809831-8.00009-X.

[11] R. Mir, S. M. Kelly, Y. Xiao, et al. “Organ at risk delineation for radiation therapy clinical trials: Global Harmonization Group consensus guidelines”. en. *Radiotherapy and Oncology* 150 (Sept. 2020), pp. 30–39. DOI: 10.1016/j.radonc.2020.05.038.

[12] M. A. Deeley, A. Chen, R. Datteri, et al. “Comparison of manual and automatic segmentation methods for brain structures in the presence of space-occupying lesions: a multi-expert study”. en. *Physics in Medicine and Biology* 56.14 (July 2011), pp. 4557–4577. DOI: 10.1088/0031-9155/56/14/021.

[13] D. Merkel. “Docker: Lightweight Linux Containers for Consistent Development and Deployment”. *Linux J.* 2014.239 (Mar. 2014). Place: Houston, TX Publisher: Belltown Media.

[14] J. Yang, G. Sharp, H. Veeraraghavan, et al. *Data from Lung CT Segmentation Challenge*. 2017. DOI: 10.7937/K9/TCIA.2017.3R3FVZ08.

[15] J. Wasserthal, H.-C. Breit, M. T. Meyer, et al. “TotalSegmentator: Robust Segmentation of 104 Anatomic Structures in CT Images”. *Radiology: Artificial Intelligence* 5.5 (Sept. 2023). Publisher: Radiological Society of North America, e230024. DOI: 10.1148/ryai.230024.

[16] F. Isensee, P. F. Jaeger, S. A. A. Kohl, et al. “nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation”. en. *Nature Methods* 18.2 (Feb. 2021), pp. 203–211. DOI: 10.1038/s41592-020-01008-z.

[17] J. Hofmanninger, F. Prayer, J. Pan, et al. “Automatic lung segmentation in routine imaging is a data diversity problem, not a methodology problem”. en. *arXiv:2001.11767 [physics, stat]* (Jan. 2020). arXiv: 2001.11767.

[18] I. Atomenergie-Organisation and I. Atomenergie-Organisation, eds. *Dosimetry in diagnostic radiology: an international code of practice*. en. Technical reports series 457. Vienna: International Atomic Energy Agency, 2007.

Accuracy comparison between GATE and TOPAS for 6MV Varian CLINAC IX

Zakaria Ait Elcadi[1] and Othmane Bouhali[1]

[1]Electrical and Computer Engineering, Texas A&M University at Qatar, Doha, 23874, Qatar.

Abstract Monte Carlo simulations in radiation therapy often face CPU timing challenges, prompting the development of various Geant4-based platforms for optimization. Our study compares GATE and TOPAS, both Geant4-rooted, aiming to guide users toward platforms offering optimal CPU tools and high precision.
The assessment involves CPU timing and electron beam configuration comparisons. Firstly, exploring platform-specific physics, phase space, and variance reduction techniques; secondly, scrutinizing electron beam configurations using stringent criteria.
Results show GATE's 7% efficiency edge over TOPAS in CPU timing, attributed to distinct techniques. However, electron beam assessments reveal discrepancies in mean energy and spot size. TOPAS excels in agreement with measurement percentage dose points due to its advanced physics configuration.
This analysis emphasizes GATE's computational efficiency and TOPAS's electron source modeling precision. It stresses the importance of platform-specific capabilities when selecting radiation therapy simulation tools.

1 Introduction

Clinical linear accelerators (LINACS) allow the production of electrons or photons with energy of several MeV, commonly used for the treatment of cancer. These LINACS are based on the principle of acceleration of electrons by high frequency electromagnetic waves. The goal of this therapeutic weapon is to deliver the highest possible dose of radiation to the tumor volume while limiting the irradiation of surrounding healthy tissues to reduce the risk of complications. MC methods are based on a microscopic modeling of particle transport resulting in a 1D, 2D and 3D simulation. The trajectories of the particles are simulated individually from a series of random numbers and cross sections responsible for reproducing the physical phenomenon. The solution is derived from the statistical mean of the quantities obtained from the simulation of a large number of primary particles. Despite this precision, they are difficult to apply in clinical routine because of excessive computation times. Several Monte Carlo simulation software [1–5] have been developed for Over time, several platforms have emerged, leveraging Geant4-based tools, each striving to optimize CPU performance. Yet, within this cohort, disparities in options and their consequential impact on CPU timing endure as a persistent issue. In this context, our study assumes a critical role, aiming to meticulously compare and contrast two prominent platforms, GATE and TOPAS, both deeply rooted in the Geant4 framework. By elucidating the nuances and performance benchmarks of GATE and TOPAS, our study endeavors to offer a guiding compass, ensuring practitioners can navigate the complexities of Monte Carlo simulations in radiation therapy with a judicious amalgamation of computational efficiency and unparalleled accuracy. [6–10].

2 Materials and Methods

GATE (Geant4 Application for Tomographic Emission) is an open-source Monte Carlo simulation based on Geant4. The GATE/Geant4 v9.1 using Geant4 v10.7, brings faster simulation times, better tracking algorithms, and enhanced visualization. TOPAS (TOol for PArticle Simulation) is a Monte Carlo platform for proton and carbon ion therapy. TOPAS v3.9 based on Geant4 v10.7 offers new features like simulating spot scanning proton therapy for more precise tumor targeting. The Varianc Clinac IX were modeled (figure 1) based on data provided by the manufacturer [11]. Our simulations were conducted on a $30 \times 30 \times 30\ cm^3$ water tank divided into small voxels of $5 \times\ 5 \times\ 5\ mm^3$ for the standard field size of $10 \times\ 10\ cm^2$. Because the profile measurements were all taken along an axis (the axis of movement of the jaws closer to the beam, which we call the Y axis), we only use one voxel in the orthogonal direction, X. The number of voxels along the Y axis will cover the central region and the penumbra.

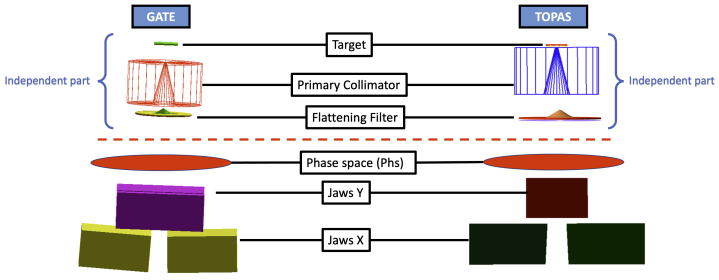


Figure 1: GATE and TOPAS modeling of Varian CLINAC IX head components for 6 MV photon beam

We used the high performance computing facility at Texas AM University in Qatar to reduce simulation computing time [12]. It is a Linux-based system outfitted with cutting-edge blade servers and a total of 4,128 traditional Intel Haswell CPU cores. All compute nodes are equipped with two 12-core processors and 128GB of RAM. It is linked to a 1PB storage system. The simulations were carried out initially with one node, and then we increased the number of nodes

from two to 94 to investigate the effect of nodes number on calculation time. TOPAS, in contrast to GATE, fully supports Geant4's multithreading simulation capability. We used the thread values from 2 to 94 in parallel to study the effect of varying the threads value on CPU timing.

2.1 Phase space

A phase space (PhS) encapsulates all potential states of a system as points in a space, defined by particles' attributes like energy, type, position, and direction. Collected by a surface in a specific geometry, this data is stored digitally for simulations [13]. Using a PhS streamlines initial simulation by needing only secondary collimation specifics for a fixed photon beam. It saves time by doing electron transport calculations just once.
PhSs are widely used in dosimetric applications. Modeling a photon beam with a PhS involves creating the PhS by collecting particles passing through a plane at the flattening filter exit and saving that data for reuse. While efficient, PhS files can become large and inflexible once made.
They come in formats like ROOT [14] or PHPS [15] and are created using software such as GATE and TOPAS. When optimizing multiple parameters, starting with less interdependent ones helps avoid biases. Adjustments in the optimization process may enhance the accuracy of final estimates.

2.2 Physics cuts

There are two kinds of physics cuts that can have a significant impact on CPU time. The first are production thresholds, which aim to avoid the generation of unnecessary low-energy secondary particles by ionization and bremsstrahlung processes. The minimum remaining kinetic energy for a track, that is, the energy below which particles are no longer transported, is the second type of physical cut that can help optimize the processus. The cutting energy prevents particle traces and thus reduces the time per history. The cut-off energy can be defined in two ways: either once for all the world or by a specfic modeled geometry. The user should define the production thresholds in GATE and TOPAS. This threshold should be defined as a distance or limit distance for individual materials, which is internally converted into energy. A region's production thresholds are established. Each volume is treated as a geometric region in GATE. TOPAS, unlike GATE, allows users to define the lower and upper edge production thresholds in terms of energy.

2.3 Variance reduction technics

Significant CPU time simulating an accelerator head involves tracking initial electrons to produce bremsstrahlung gammas. Bremsstrahlung splitting reduces time spent on electrons, favoring gamma tracking for better phantom reach. A secondary threshold is needed for particulate matter generation above the production threshold. Calculation precision affects time; "StepFunction" based on voxel size limits step size [16].
Choosing step size in Monte Carlo tracking is tough; TOPAS decides based on local factors, physics, and user input. Experimenting with different max sizes determines optimal precision. GATE Special Slices set limits on particle tracking. "KillActor" in GATE stops tracking particles outside a volume, saving computation. TOPAS filters particles upon creation for efficiency but caution is advised, as it's not a variance-reduction method.

3 Results

The need for long computation times in Monte Carlo (MC) simulations necessitates the use of various tools to reduce computation time while ensuring the accuracy of the final results. As described in section 2 , we utilized GATE and TOPAS to study different optimization techniques, prioritizing those that are less dependent on others. Unless otherwise specified, all results were obtained using 12×10^9 initial electrons with 6.0 MeV energy and a zero-width beam, 10×10 cm^2 field size and a division number of 100 for Bremsstrahlung splitting. The simulations used the Standard EM model opt3 for GATE, while opt4 was chosen for TOPAS with default parameters.
To initiate the optimization process, we analyzed the execution time at different locations of the accelerator head to gain insights into possible optimization factors. Initially, the entire accelerator head was simulated in a single step, enabling the simulation of particles interacting with various components until they reach the phantom located 100 cm away from the target. The entire Linac geometry was simulated with an initial electron energy of 6 MeV, a production threshold of 0.1 mm, and a reference field size of 10×10 cm^2. Subsequently, a phase file was added under each component of the Linac to calculate the computation time consumed by each component. The obtained results depicted in table 1 were as expected, indicating that most of the time was spent tracking electrons, especially in the target.

3.1 Phase space

The process of optimizing computation time began by utilizing the maximum number of nodes available to the user. To speed up simulations, a parallelization script was introduced to divide the simulation into 200 parallel tasks, each job containing 6×10^7 particles, and associated with the maximum number of nodes. This resulted in a 25% for TOPAS and 27% for GATE optimization when compared to the use of a single node.
Next, the PhS was used to divide the geometry into two parts: a dependent part and an independent part (figure 1). The PhS in this work was modeled as a cylinder with a diameter of 10 cm and a thickness of 1 nm in the z-direction,

Particle/Geometry region	Time/Time total (%)
e-/Target	81
e-/Primary collimator	1.1
e-/Flattening filter	1.2
e-/Jaws	1.1
e-/Outside	0.3
e-/Total	84.7
γ/Target	6.7
γ/Primary collimator	1.2
γ/Flattening filter	1.2
γ/Jaws	0.9
γ/Outside	5.1
γ/Total	15.1
e+/Total	0.2

Table 1: Fraction of time spent in each Varian CLINAC IX head components by particle type

located 7 cm above the secondary collimator or Y-jaws. The PhS format differs between GATE and TOPAS: in GATE, there is the phase space actor, which can be in PHPS, or ROOT format to avoid large format files; in TOPAS, the ROOT format cannot be used as a source for the second accelerator head module. Each PhS associated with a 6.10^7 particle job is three times larger than the one in root format. By considering only the dependent part and using the PhS as the beam source, the simulations were simplified. The calculation time was reduced by 66% for TOPAS and 68% for GATE when compared to the simulation of the entire accelerator head.

3.2 Physics cuts in the phantom

The use of a 0.1 mm cut-off in the phantom and throughout the entire geometry allows for the assumption that the deposited energy threshold is confined to the corresponding dose grid voxel. To verify this assumption, a PhS file was generated with a $10 \times 10\ cm^2$ field, which stopped at an XY plane located 1 cm before the phantom. Using various production threshold values, particles were eliminated, and the integral of the dose outside the creation voxel was calculated and compared to the CPU time gain. Table 2 summarizes the results, indicating that a production threshold value of 0.1 mm in the water phantom would not significantly affect the dose. The negligible impact on the dose is likely due to the fact that the energy threshold is confined to the corresponding dose grid voxel, as assumed, and because the phantom's cut-off is consistent with the world's cut-off .

3.3 Variance reduction techniques

Multiple simulations were conducted using variance reduction techniques. A division factor of 100 was utilized, and an energy slice of 1 mm, which corresponds to 1/5 of the

Cut in phantom (mm)	CPU time gain (%)
0.01	-24
0.05	-5
0.1	0
0.2	2
0.5	4
1	8

Table 2: CPU time gain when implementing a specific production threshold in the water phantom, utilizing the default GATE & TOPAS value of 0.1mm.

voxel dimension, was assigned. The step function was set to 0.1 mm. Furthermore, the software programs GATE and TOPAS have parameters called "setMaxStepSizeInRegion" and "MaxStepSize" for electrons, respectively, which allow for limiting the maximum step for a particle in a given region. To identify the optimal value for this parameter, various values ranging from 10 to 50 μm, with a pitch of 10 μm, were investigated. The chosen value was selected based on the recommended cut-off energy value and a reasonable calculation time, as presented in the table 3 [17].

Max Step (μm)	CPU Time gain (%)
10	10
20	-5
30	-15
40	-34
50	-71

Table 3: CPU time gain for different values of "setMaxStepSizeInRegion" in GATE and "MaxStepSize" in TOPAS.

The selection of the thread number is a variance reduction technique offered by TOPAS. To analyze the impact of this technique on execution time, several simulations were conducted using 2 to 64 threads. The results indicated that using more than 32 threads in a single simulation did not result in any significant improvement in the total CPU timing. One of the actors that GATE offer to contributes to variance reduction is the "KILL" actor, which is attached to a cylindrical geometry surrounding all of the other geometry and two boxes, one above the target and one below the water phantom. Table 4 summarizes the computational time savings for both TOPAS and GATE after implementing all processes (the first process is simulating the entire geometry), including PhS, cuts, and variance reduction techniques.

4 Discussion

The study explores efficient strategies in Monte Carlo simulations, emphasizing the need for optimization. Using GATE and TOPAS, it focuses on independence and reveals intriguing insights from simulations. Initial analysis showed electron tracking as a major time-consuming task, emphasizing

Process	CPU time gain (%) TOPAS	GATE
Jobs splitting	25.2	26.9
PhS (as a source)	65.82	67.82
Cuts	70	73.24
Variance reduction techniques	88.99	94.47

Table 4: CPU gain time for TOPAS and GATE after appliying all CPU optimization process

optimization importance. Fragmenting tasks led to faster CPU times. Parallelization scripts boosted optimization by 25-27%, and intelligent use of phase space reduced calculation time by 66-68%. Fine-tuning physics cuts and variance reduction techniques, like adjusting step sizes and using "KILL" in GATE, significantly cut computation time. GATE showed a 7% CPU optimization edge over TOPAS due to its adept use of Phase Space and Variance Reduction techniques.

5 Conclusion

This study underscores the vital need for efficient Monte Carlo simulations without compromising accuracy. By exploring GATE and TOPAS optimization tools, it unveiled crucial pathways for precision while minimizing computational expenses.

Identifying time-consuming aspects like electron tracking and utilizing fragmentation, parallelization, and phase space segmentation significantly reduced computation times.

Fine-tuning physics settings in the phantom highlighted the balance between computation time and accuracy, aiding in optimal parameter selection.

Notably, GATE's advantage over TOPAS by 7% in CPU optimization, thanks to its Phase Space and variance reduction techniques, offers users a choice. They can prioritize their platform based on specific needs while ensuring both accuracy and efficiency in their work.

References

[1] I. Kawrakow and B. R. B. Walters. "Efficient photon beam dose calculations using DOSXYZnrc with BEAMnrc". *Medical Physics* 33.8 (2006), pp. 3046–3056. DOI: 10.1118/1.2219778.

[2] G. Battistoni, J. Bauer, T. T. Boehlen, et al. "The FLUKA Code: An Accurate Simulation Tool for Particle Therapy". *Front Oncol* 6.116 (May 2016), pp. 1–24.

[3] R. Schwarz, L. L. Carter, and A. Schwarz. "Modification to the Monte Carlo N-Particle (MCNP) Visual Editor (MCNPVised) to Read in Computer Aided Design (CAD) Files". *doi:10.2172/843024* 38 (Aug. 2005). DOI: 10.2172/843024.

[4] J. Allison, K. Amako, J. Apostolakis, et al. "Geant4 developments and applications". *IEEE Transactions on Nuclear Science* 53.1 (Feb. 2006), pp. 270–278. DOI: 10.1109/TNS.2006.869826.

[5] R. Augusto, J. Bauer, O. Bouhali, et al. "An overview of recent developments in FLUKA PET tools". *Physica Medica* 54 (2018), pp. 189–199. DOI: https://doi.org/10.1016/j.ejmp.2018.06.636.

[6] P. Arce and J. I. Lagares. "CPU time optimization and precise adjustment of the Geant4 physics parameters for a VARIAN 2100 C/D gamma radiotherapy linear accelerator simulation using GAMOS". *Physics in Medicine 'I&' Biology* 63.3 (2018), p. 035007.

[7] K. J.-I. Park Hyojun Choi Hyun Joon and M. C. Hee. "Analysis of Dose Distribution According to the Initial Electron Beam of the Linear Accelerator: A Monte Carlo Study". *J Radiat Prot Res* 43.1 (2018), pp. 10–19. DOI: 10.14407/jrpr.2018.43.1.10.

[8] H. Jung, J. Shin, K. Chung, et al. "Feasibility of using Geant4 Monte Carlo simulation for IMRT dose calculations for the Novalis Tx with a HD-120 multi-leaf collimator". *Journal of the Korean Physical Society* 66.10 (May 2015), pp. 1489–1494. DOI: 10.3938/jkps.66.1489.

[9] S. Gholami, F. Longo, H. A. Nedaie, et al. "Application of Geant4 Monte Carlo simulation in dose calculations for small radiosurgical fields". *Medical Dosimetry* 43.3 (2018), pp. 214–223. DOI: https://doi.org/10.1016/j.meddos.2017.08.007.

[10] R. Onizuka, F. Araki, and T. Ohno. "Monte Carlo dose verification of VMAT treatment plans using Elekta Agility 160-leaf MLC". *Physica Medica* 51 (2018), pp. 22–31. DOI: https://doi.org/10.1016/j.ejmp.2018.06.003.

[11] Varian. *1999-2024 Varian Medical Systems,* URL: https://www.varian.com/.

[12] O. Bouhali, A. Sheharyar, and T. Mohamed. "Accelerating avalanche simulation in gas based charged particle detectors". *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 901 (2018), pp. 92–98. DOI: https://doi.org/10.1016/j.nima.2018.05.061.

[13] C. R., J. R., M. C.-M., et al. "Phase-space database for external beam radiotherpy." *Nuclear Data Section, International Atomic Energy Agency* (2006).

[14] I. Antcheva, M. Ballintijn, B. Bellenot, et al. "ROOT — A C++ framework for petabyte data storage, statistical analysis and visualization". *Computer Physics Communications* 180.12 (2009), pp. 2499–2512.

[15] R. Capote. "IAEA nuclear and atomic data for medical applications: Phase-space database for external beam radiotherapy nuclear data for heavy charged-particle radiotherapy". *Radiotherapy and Oncology* 84 (2007), S217–S217.

[16] opengatecollaboration. *http://www.opengatecollaboration.org/*. Tech. rep. opengatecollaboration, 2023.

[17] P. E and V. F. "Accuracy of the photon and electron physics in GEANT4 for radiotherapy applications." *Journal of Medical Physics.* 32 (2005), pp. 1671–1696.

Can Low-Cost Magnetic Resonance Distortion Phantoms be 3D Printed?

Haley Clark[1,2]

[1]Department of Radiation Oncology, BC Cancer, Surrey, Canada
[2]Department of Physics and Astronomy, University of British Columbia, Vancouver, Canada

Abstract Expanding use of diagnostic MR fusion, MR-simulation, and MR-linacs in radiotherapy has increased awareness of the quality assurance challenges associated with spatial distortions. While precise and effective commercial phantoms exist, they are costly and can present challenges when monitoring multiple MR scanners. In this work, we evaluate whether extremely low-cost 3D printed phantoms exhibit the spatial accuracy needed to monitor spatial distortions. Using several printed variations, we find that clinically sufficient accuracy is possible for up to four orders of magnitude cheaper cost than a commercial phantom.

1 Introduction

Spatial distortions in magnetic resonance (MR) images used for radiotherapy planning can negatively impact tumour and organ-at-risk localization [1]. Recently, quality assurance guidelines have emerged recommending routine monitoring of spatial distortions for MR scanners [2, 3].

Commercial phantoms made of high-grade materials can be used to monitor distortions, but they are generally expensive and bulky. BC Cancer radiotherapy makes use of a network of numerous external MR scanners, so use of commercial phantoms for each individual site would be not ideal. 3D printed phantoms, on the other hand, are lightweight and rigid, so could be more easily transported. Low production costs mean distributing phantoms to numerous MR scanners would be more cost effective with 3D printed phantoms, and the low-complexity construction means phantoms could, in principle, be made on-site as needed. However, spatial characteristics of such phantoms must be both accurate and consistent enough so that clinically relevant distortions remain detectable (i.e., ≤1-2 mm) [2].

In this work we investigate whether lightweight, low-cost phantoms can be 3D printed with sufficient accuracy for routine MR distortion monitoring.

2 Materials and Methods

A computer program was developed to generate $\mathbb{R}^3$ manifold surface meshes of rectilinear lattice phantoms with arbitrary cell dimensions and edge thickness. Four phantoms (one ‘small’ with 2x3x4 cells; two ‘medium’ with 4x4x8 cells; and one ‘large’ with 4x4x16 cells) were 3D printed using 1.75 mm diameter polylactic acid (PLA) filament (MakerBot MP05778). (See fig. 1.) Cells were isotropic with 15 mm extents. Edges were 3 mm in diameter.

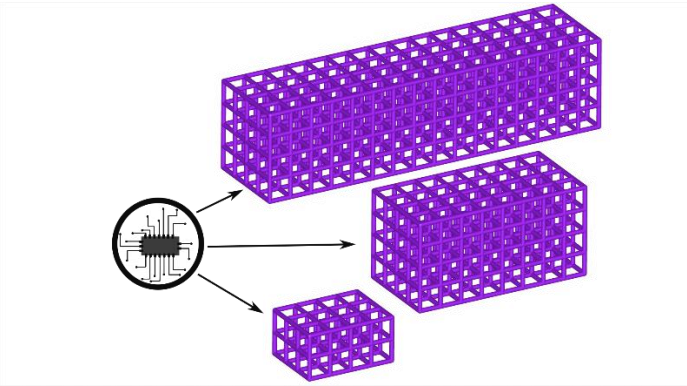


Figure 1: Small, medium, and large rectilinear lattice phantom models [4]. The regular structure can be used to monitor MR image distortions.

Two 3D printers were evaluated: a MakerBot Replicator Z18 and a Creality Ender 3. Print times spanned 3 - 12 hours, varying with phantom size. Medium and large phantoms analyzed in this study were printed with vertical supports and base-layer rafts to improve adhesion and print accuracy; the small phantom was printed with neither. Edge element thicknesses, cell inner and outer widths, and total phantom extents were each measured in 12 places with a caliper accurate to 0.02 mm. (See fig. 2.)

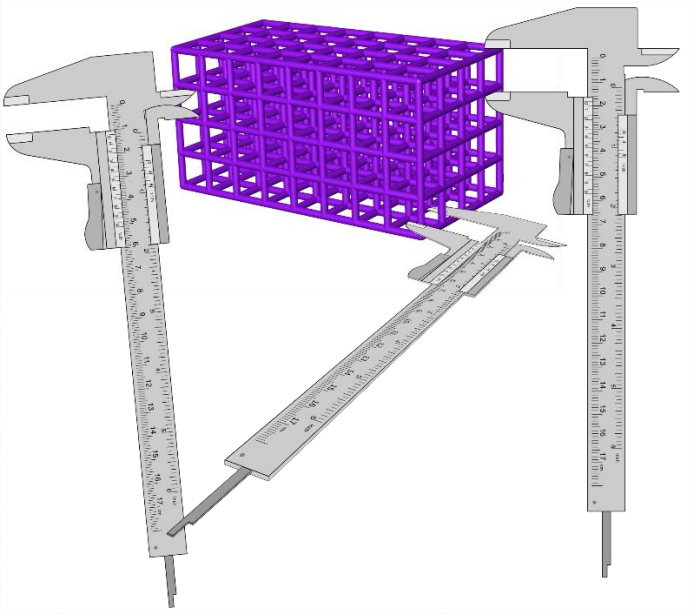


Figure 2: Schematic of caliper [5] measurements performed on all four phantoms. Lattice cell widths, edge diameters, and phantom extents were sampled in 12 places. Distortion (see fig. 4) was assessed with a square.

3 Results

All phantom variants printed without issues or failures. Printing supports were sufficiently thin (≤ 0.1 mm) that they could either be safely ignored or reliably eliminated during analysis of CT or MR images with a top-hat morphological transform [6]. Alternatively, even the largest phantom was able to be printed without the use of supports by bridging. (See fig. 3.)

Mean edge diameters for all phantoms were 3.04 mm (ideal: 3.0 mm, worst: 3.45 mm); mean cell inner widths were 11.91 (12.0 mm, 11.37 mm); mean cell outer widths were 17.88 mm (18.0 mm, 17.18 mm); mean total phantom extents near the extrema were 62.72 mm (63.0 mm, 62.69 mm); and mean total phantom extents near the middle were 62.04 mm (63.0 mm, 62.04 mm). Deflection and accuracy degraded worst along the printing axis. (See fig. 4.)

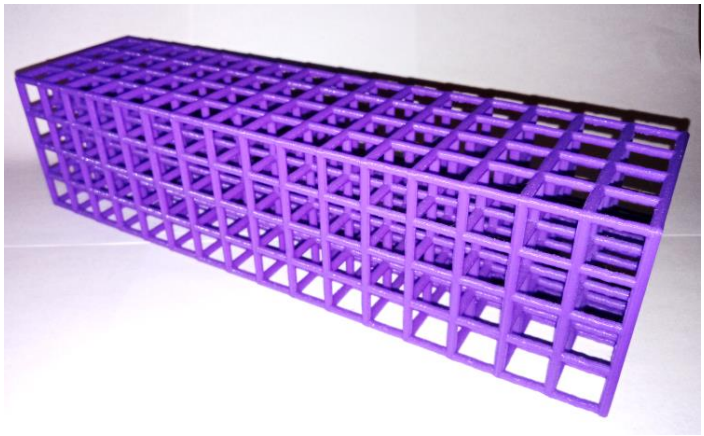


Figure 3: Example of a 3D printed 4x4x16 grid phantom, printed on a Creality Ender 3 printer.

The small phantom was the most accurate overall; the average discrepancy for all characteristics were ≤ 0.20 mm from ideal.

The medium (large) phantoms had average discrepancies ≤ 0.74 mm (≤ 0.76 mm) from ideal for all characteristics. The small phantom was printed without supports or a raft, yet still demonstrated excellent accuracy, minimal bridging artifacts, and good print adhesion.

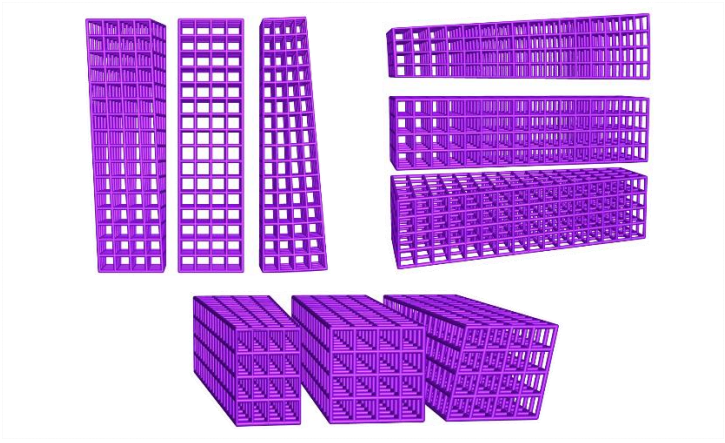


Figure 4: Exaggerated phantom warps that can result during printing, including skew, keystoning, and twists. Warps will confound distortion analysis. All can be ameliorated via use supports when printing, use of rafts if the printer bed exhibits adhesion issues, accurate bed leveling, and calibration of extrusion rates and printer bed temperatures.

4 Discussion

The 3D printed phantoms investigated in this work were all found to be sufficiently accurate for routine distortion monitoring purposes [3]. Commercial phantoms vary in cost but are typically $10,000 - $80,000 CAD. In contrast, the phantoms 3D printed for this study cost less than $300 CAD altogether, including the cost of the printer – a price difference of *two orders of magnitude*.

Material costs for printing a single phantom was approximately $5 CAD, or *four orders of magnitude cheaper* than a proprietary phantom. (See fig. 5.) While commercial phantoms are precision devices designed to accurately emulate clinical scenarios, use of low-cost 3D printed phantoms could conceivably assure greater quality through improving access to necessary equipment.

The limited spatial extent of the phantoms printed in this study presents a hindrance to whole-bore distortion checks. This can be overcome by chaining several phantoms together with interlocking connectors. Since the smallest phantom exhibited the most accurate dimensions, chaining small phantoms together might provide the best overall geometric accuracy. Additionally, this idea could be extended to create interlocking phantoms custom-fit for a specific MR scanner, patient immobilization device, coil, or support devices. However, further work is required to evaluate feasibility.

Note that while typical 3D printed materials will not directly produce signal in an MR scanner, the phantoms presented here can be imaged by submerging in liquid. Many MR-compatible 3D printed materials are not water soluble, so can withstand being submerged indefinitely, and air bubble

contamination can be countered with surfactants or ultrasonic vibration.

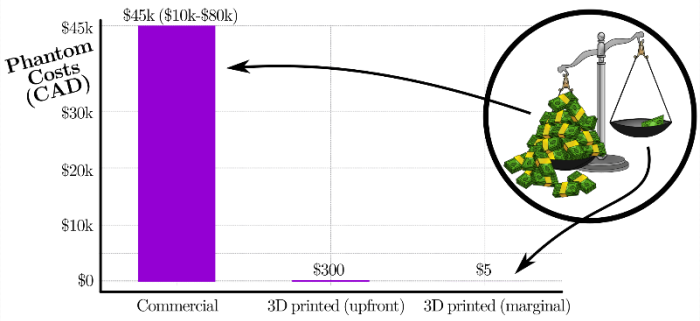


Figure 5: Quantitative cost comparison for commercial phantoms, 3D printed phantoms (+upfront costs), and 3D printed phantoms (marginal costs only). All are approximate. Inset: qualitative depiction of phantom costs; commercial on left, 3D printed (marginal costs only) on right.

5 Conclusion

3D printed MR distortion phantoms offer a compelling low-cost option for routine quality assurance when MR scanners will be used for radiotherapy. The phantoms presented in this work are straightforward to 3D print and can be done so quickly. Design files and surface meshes for these phantoms are open source.

References

[1] E Pappas et al., MRI-Related Geometric Distortions in SRT Planning, 2017, doi:10.1177/1533034617735454.

[2] Glide-Hurst CK, Paulson ES, McGee K, Tyagi N, Hu Y, Balter J, Bayouth J. Task group 284 report: magnetic resonance imaging simulation in radiotherapy: considerations for clinical implementation, optimization, and quality assurance. Medical physics. 2021 Jul;48(7):e636-70. doi:10.1002/mp.14695

[3] Canadian Partnership for Quality Radiotherapy: Technical quality control guidelines for magnetic resonance imaging for radiation treatment planning. 2020 Sept 8. Available from http://www.cpqr.ca/programs/technical-quality-control

[4] CPU ‘chip icon micro’ by Sinisa Maric, CC0, Pixabay: 1710300.

[5] Vernier caliper by J A Gaspar, modified by ed g2s, CC BY 2.5, Wikimedia Commons: 1365717.

[6] H Heijmans. Morphological image operators. Vol. 4. Boston: Academic Press, 1994.

Acknowledgements

The author would like to thank Steven Thomas and Scott Tyldesley for insightful discussions about MR distortions.

Automatic contour quality assurance (CAT-QA) to speed up online adaptive radiotherapy

L.S. Fankhauser[1,2], A. Smolders[1,2], R. Bellotti[1,2], D.C. Weber[1,3,4], A.J. Lomax[1,2], and F. Albertini[1]

[1]Center for Proton Therapy, Paul Scherrer Institute, Villigen, Switzerland
[2]Department of Physics, ETH Zurich, Zurich, Switzerland
[3]Department of Radiation Oncology, Inselspital, Bern University Hospital, University of Bern, Bern, Switzerland
[4]Department of Radiation Oncology, University Hospital Zurich, Zurich, Switzerland

Abstract One of the main bottlenecks in implementing online adaptive radiotherapy is the time required to contour and approve the daily contours. Quality assurance (QA) to identify those automatically generated contours that would benefit from manual revision would save a lot of time. This study aimed to develop a contour QA workflow for adaptive proton therapy for head and neck cancer. For the daily optimization, propagated organ-at-risk (OAR) contours are used. The proposed automatic contour QA (CAT-QA) uses geometrical and dosimetric tests to identify structures for manual re-contouring. It is split in two phases: an offline phase conducted before the treatment and an online phase conducted after acquiring the daily image. We applied this workflow to nine patients with multiple repeated CTs. Over all cases, maximum three structures were selected for manual revision. Using CAT-QA ensures to follow the initial prescription and avoids dose increase in OARs due to uncertainties of the propagated contours. In conclusion, although more validation is required, the defined workflow could enable safe use of auto contouring for adaptive radiotherapy whilst reducing time spent on manual contouring.

1 Introduction

Online adaptive radiotherapy aims to optimally conform the dose to the target by planning on daily anatomy ([1]). Using adaptation has shown to reduce the dose to healthy tissue for head and neck cancer (H&N) patients [2]. A daily adaptive workflow has to be fast to avoid anatomical changes between the acquisition of the daily image and the start of treatment. One of the main bottlenecks for clinical implementation of online adaptive is the time and human resources required to contour and approve the daily contours. These daily contours are essential to re-optimize the treatment plan on the daily anatomy, and, because of time limitations, need to be generated automatically by either automated segmentation and/or propagation.

Whereas it has been suggested that manual recontouring is often not necessary, especially for the organs-at-risk (OARs) [3], automatic algorithms can still yield unexpected results and therefore need quality assurance (QA). Including an automatic QA workflow that identifies those structures that need to be recontoured is an important step towards speeding up the daily recontouring task. Some work has been done in this direction [4–8], developing QA methods based on geometric differences. However, geometric differences do not necessarily result in clinically relevant dosimetric variations. Thus, correcting OAR-contours for which potential dose variations have little clinical impact only takes up time without improving treatment quality. In this study, we develop a contour QA workflow for adaptive proton therapy for head and neck cancer that identifies only the structures that benefit from manual revision, taking into account both geometric and dosimetric parameters.

2 Materials and Methods

Nine H&N patients, with on average four repeated CTs, have been used to investigate the proposed contour QA. Manual contours were delineated by radiation oncologists on all available CTs and are used as reference structures for the analysis outlined below. All cases were optimized on each of the available images with an IMPT plan with a SIB approach with two dose levels, mimicking a clinically plausible treatment plan. The automatic contour QA, CAT-QA, described below is tested retrospectively by simulating an online adapted workflow, using the available repeated CTs as potential daily images.

2.1 CAT-QA Workflow

The proposed contour-QA worklfow, CAT-QA, is split into two phases: an *offline* and an *online* phase (Fig. 1). During the offline phase, which is performed before the treatment, the sensitivity of the structures is assessed both geometrically and dosimetrically on the planning image. Non-sensitive structures identified during the offline phase are excluded from further evaluation in the online phase. In the online phase, the geometric and dosimetric sensitivity of only a subset of the initial structures is further evaluated on the daily anatomy. By reducing the number of structures to be evaluated online, the online QA workflow is more efficient and potentially suitable for an online adaptive workflow.

2.1.1 Offline phase

For the offline part, the planning CT, manual contours and the planned dose are used as input data (input data 1, fig 1). To assess the quality and sensitivity of the contours, we compare the manual contours to another structure set. Here, these alternative contours are generated with the deep-learning

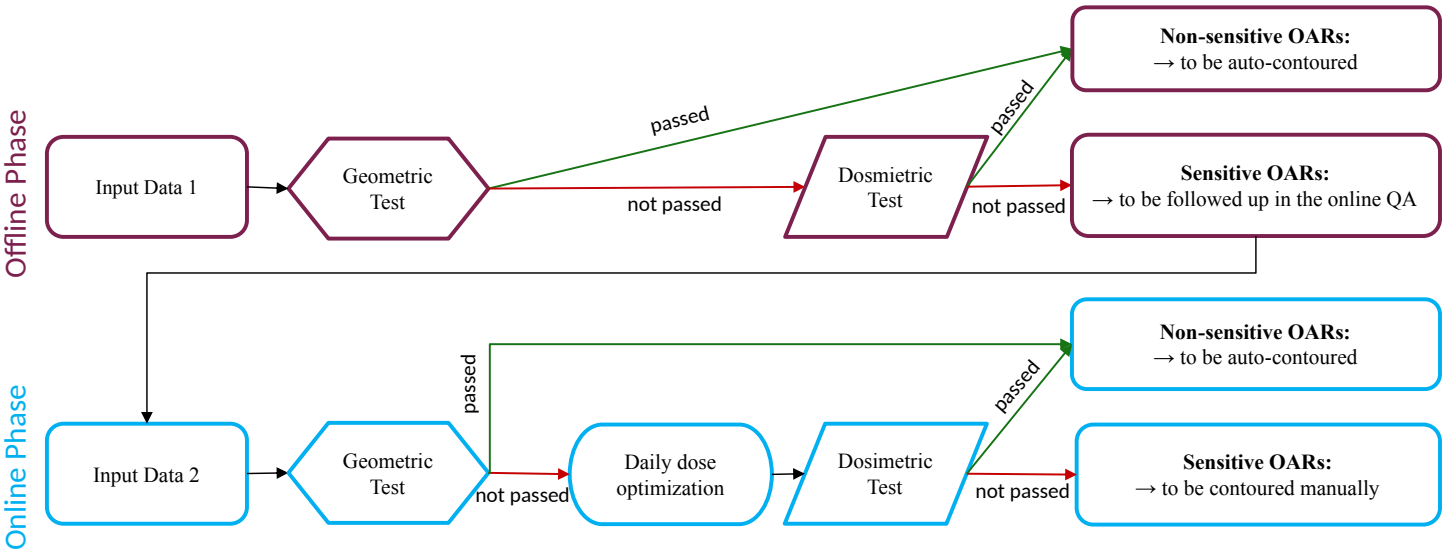


Figure 1: The offline phase is conducted after acquiring and generating the input data 1: the planning CT, the planned dose, the manual contours, and generating the auto-contours. The online phase is conducted after acquiring and generating the input data 2: the planning CT, manual contours of planning CT, planned dose on planning CT, the daily CT, the propagated contours. The daily dose is calculated by optimizing the dose using the propagated contours.

segmentation in Raystation (Raysearch AB, Stockholm, Sweden). First, the sensitivities of the structures are checked by carrying out 4 different geometric tests (described in detail in 2.1.3) by comparing the manual contours with the auto-contours. All OARs for which their auto-contour passed the geometric test are labeled as non-sensitive. These contours are not further considered in the offline or online workflow. The remaining OARs undergo a dosimetric sensitivity analysis. This is done by comparing the plan optimised with the auto-contours with the one optimised with the manual contours (described in detail in section 2.1.4). OARs that successfully pass the dosimetric test are deemed to have a low influence on the dose and are consequently excluded from further investigation in the online phase

2.1.2 Online Phase

For the online part, the planning CT, the manual contours of the planning CT, the planned dose and the daily CT are used as input data (input data 2, Fig. 1). Daily contours are propagated by registering the daily CT with the planning CT using plastimatch b-spline. Firstly, the sensitivity of the structures on the daily image is checked by performing four different geometric tests (see details in section 2.1.3). All OARs that pass the geometric test are marked as non-sensitive. This means that the automatically generated contours can be used to optimise the daily plan. The structures that did not pass the geometric tests are further evaluated to assess their dosimetric sensitivity (see 2.1.4). All OAR contours that passed this dosimetric test are also flagged as non-sensitive and all the others flagged as sensitive. For these finally flagged sensitive contours, manual revision is needed to proceed with the online adaptive workflow. For all the non-sensitive structures the b-spline contours are further used for the online adaptive workflow.

2.1.3 Geometric test

Both offline and online, four geometric tests are performed.
1. **Mean Hounsfield unit (HU):** Mean HU values are compared between the structure on the planning CT and the automatically generated one either on the planning CT for the offline part or on the daily CT for the online part. For the online part, the b-spline generated contours are used for the comparison of the HU values.Differences up to 100HU are accepted.
2. **HU range in contours:** The range of the HU values in the contours are tested. If the range changes are not more than $\pm$ 200 HU, the geometry of the structures is further evaluated by the third and fourth step.
3./4. **Dice similarity coefficient (DSC)/surface dice** are evaluated. For this, in the online workflow instead of the manual contours from the planning CT, a second set of auto-contours (this time generated using plastimatch demons) are used to take into account the anatomical deformations of the structures over the course of the treatment. If the DSC and the surface dice are higher than 0.8 (see AAPM report [9]), the contour has passed the geometric test.

2.1.4 Dosimetric test

The dosimetric test assesses the clinical endpoints outlined in the prescription for plans optimized with automatic contours in comparison to the reference plan optimized with manual contours. Offline, it compares the planned dose endpoint with the same endpoint in the dose reoptimized with the auto-contours, but with both evaluated on the manual contours. Online, these planned dose endpoints are compared to

the daily dose endpoints, evaluated in the b-spline contours themselves.Any dosimetric differences surpassing 20% (as suggested by [10]) or exceeding OAR tolerance categorize the structure as sensitive.

2.2 Evaluation of CAT-QA

The performance of CAT-QA is assessed by comparing the dosimetric differences between three plans: the plan optimised on the daily CT using manual contours (*daily reference plan*), the plan optimised using all automatic OAR contours (*fully automated plan*), and the plan optimised using a combination of the automatic OAR contours and the manual OAR contours selected as sensitive from the CAT-QA (*CAT-QA plan*). For each plan, the deviations of the prescription endpoints of the most relevant OARs and clinical target volumes (CTVs) from the reference plan are reported.

3 Results

For most OARs, the CAT-QA ensures that clinical endpoints defined in the prescription are within a 5% dose difference compared to the values achieved with the reference plan optimized on manual contours (Fig. 2). Seven out of 273 OAR structures have a dose difference larger than 5%. Our results show that by optimizing the plan with ALL automatic OAR structures, 21 structures would exceed the prescription value by more than 5% compared to the reference plan. Moreover increases in dose to healthy tissue of up to 26% of the prescribed dose would be detected in single fractions. As such, the workflow proposed here is an efficient tool to mitigate severe increases in the dose to the normal tissue. Nonetheless, the proposed workflow may not be sufficient to limit differences to less than 5% for all structures over all fractions. Indeed, for small organs, differences of up to 12% are detected for individual structures (e.g. right parotid in Fig. 2) with the defined tolerances used in the CAT-QA.
Similar considerations apply to target coverage, where, in

	total	offline sensitive	online sensitive
Case 1	37	16	1-3
Case 2	34	26	0-1
Case 3	42	32	0
Case 4	29	19	0
Case 5	15	2	0
Case 6	14	3	0-1
Case 7	15	0	0
Case 8	15	0	0
Case 9	15	2	0

Table 1: Number of structures selected by the workflow as sensitive in the offline and online phase. All OARs for which there is no automatic contour in the offline phase, are considered sensitive as well.

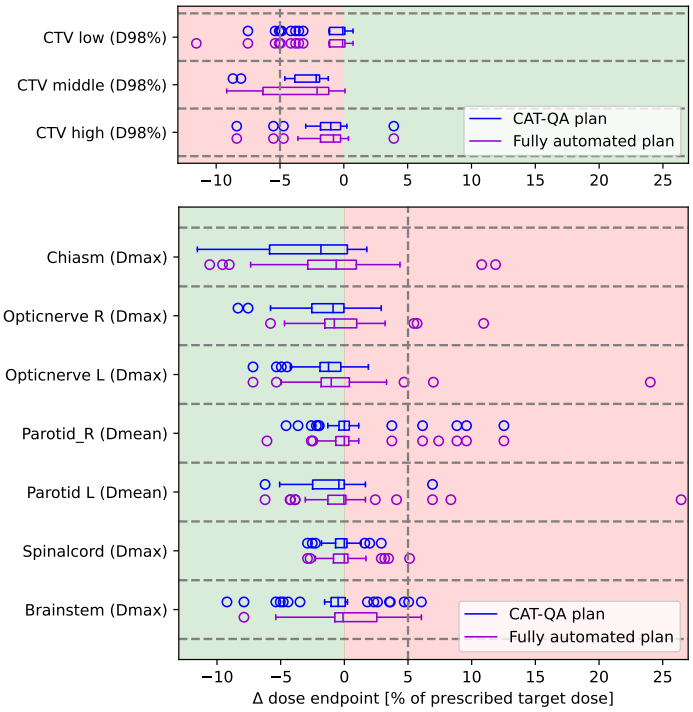


Figure 2: CAT-QA plan: Dose difference between the optimized dose using the manual contours and the optimized dose using the propagated contours combined with the manual contours for the structures selected by the workflow. Fully automated plan: Dose difference between the optimized dose using the manual contours and the optimized dose using the propagated contours.

most cases, the QA-Contour workflow is sufficient to limit the dose difference to below 5% for the daily plans However, higher differences are found for seven out of 89 target points evaluated (Fig. 2).
Nevertheless, the sensitivity analysis performed in the offline phase is an effective way to reduce the number of structures, to be further evaluated online, thus reducing the time needed for the execution of the QA tests in the online phase (Table 1). Indeed, on average, more than 60% (from 23% to 100%) of the initially defined structures are flagged as non-sensitive and are not analysed any further in the online phase.

4 Discussion

In this study, we present the framework of a automatic contour-QA workflow incorporating both geometric and dosimetric tests. The proposed workflow significantly reduces the number of structures requiring manual re-contouring, ensuring low doses to healthy tissue and high target coverage. Following the online phase, fewer than 3 structures, on average, are flagged as sensitive, necessitating manual contouring for the adaptive process. The workflow's effectiveness is highlighted by the substantial increase in doses (up to 26%) detected in OARs when not utilizing it.
While the CAT-QA workflow demonstrates positive results for most OARs, some outliers were not well detected, possibly due to uniform thresholds for geometric and dosimetric

deviations across all structures. Future improvements may involve OAR-specific tests to enhance results based on individual characteristics.
This study did not consider QA for propagation of target structures in the workflow, assuming they would be manually revised before daily optimization. A follow-up study will extend the workflow to include target contour QA.

5 Conclusion

The CAT-QA workflow, which considers both geometric and dosimetric parameters, shows promising utility for the safe integration of auto-contouring in online adaptive radiotherapy, while concurrently reducing the time dedicated to manual contouring. However, further refinement is required to minimize differences wih manually optimized plans below 2% before clinical implementation.

References

[1] H. Paganetti, P. Botas, G. C. Sharp, et al. *Adaptive proton therapy*. Nov. 2021. DOI: 10.1088/1361-6560/ac344f.

[2] C. B. Simone, D. Ly, T. D. Dan, et al. "Comparison of intensity-modulated radiotherapy, adaptive radiotherapy, proton radiotherapy, and adaptive proton radiotherapy for treatment of locally advanced head and neck cancer". *Radiotherapy and Oncology* 101.3 (Dec. 2011), pp. 376–382. DOI: 10.1016/j.radonc.2011.05.028.

[3] A. Smolders, E. Choulilitsa, K. Czerska, et al. "Dosimetric comparison of autocontouring techniques for online adaptive proton therapy". *Physics in Medicine and Biology* 68.17 (Sept. 2023). DOI: 10.1088/1361-6560/ace307.

[4] A. Kumarasiri, F. Siddiqui, C. Liu, et al. "Deformable image registration based automatic CT-to-CT contour propagation for head and neck adaptive radiotherapy in the routine clinical setting". *Medical Physics* 41.12 (Dec. 2014). DOI: 10.1118/1.4901409.

[5] W. J. Beasley, A. McWilliam, N. J. Slevin, et al. "An automated workflow for patient-specific quality control of contour propagation". *Physics in Medicine & Biology* 61.24 (Nov. 2016), p. 8577. DOI: 10.1088/1361-6560/61/24/8577.

[6] M. B. Altman, J. A. Kavanaugh, H. O. Wooten, et al. "A framework for automated contour quality assurance in radiation therapy including adaptive techniques". *Physics in Medicine & Biology* 60.13 (June 2015), p. 5199. DOI: 10.1088/0031-9155/60/13/5199.

[7] H. C. Chen, J. Tan, S. Dolly, et al. "Automated contouring error detection based on supervised geometric attribute distribution models for radiation therapy: A general strategy". *Medical Physics* 42.2 (Feb. 2015), pp. 1048–1059. DOI: 10.1118/1.4906197.

[8] S. Luan, X. Xue, C. Wei, et al. "Machine Learning-Based Quality Assurance for Automatic Segmentation of Head-and-Neck Organs-at-Risk in Radiotherapy". *Technology in Cancer Research and Treatment* 22 (Jan. 2023). DOI: 10.1177/15330338231157936.

[9] K. K. Brock, S. Mutic, T. R. McNutt, et al. "Use of image registration and fusion algorithms and techniques in radiotherapy: Report of the AAPM Radiation Therapy Committee Task Group No. 132: Report". *Medical Physics* 44.7 (July 2017), e43–e76. DOI: 10.1002/mp.12256.

[10] A. M. Barragán-Montero, G. Van Ooteghem, D. Dumont, et al. "Dosimetrically triggered adaptive radiotherapy for head and neck cancer: Considerations for the implementation of clinical protocols". *Journal of Applied Clinical Medical Physics* 24.11 (Nov. 2023). DOI: 10.1002/acm2.14095.

QATrack+: An open-source routine quality assurance platform

Randle Taylor[1] and James Kerns[1]

[1]Radformation, New York, NY, United States of America

Abstract QATrack+ is an open-source quality assurance platform able to record data, attach relevant files, and perform custom calculations. The software is extensible and allows users to write their own scripts for advanced data processing. Service events, reporting, and charting are all built-in and work seamlessly with data inputs. Advanced permissions also allow for specialized permissions and multi-role environments.

1 Introduction

Within the radiotherapy and clinical domains, handling routine quality assurance (QA) for all equipment within a clinic can be an arduous task. Such QA includes daily, weekly, monthly, quarterly, and annual requirements. [1, 2] Traditionally, this has been handled via spreadsheets or simple documents. The task of remembering the upcoming due dates, archiving binary data such as images or reference files alongside a spreadsheet, and performing any specialized tasks such as image analysis can have the clinician interfacing with multiple programs for even the most rote of tasks.

QATrack+ was developed at the the Ottawa Cancer Center as a customized solution for these problems, specifically for diagnostic [3] and regulatory requirements of QA [4]. It has since been open-sourced and used across numerous clinics worldwide.

2 Materials and Methods

The QATrack+ platform is written in the Python programming language [5] and built on the Django web framework library [6]. It is server-based and interfaced through a web browser; this means QATrack+ is deployable to the cloud or locally, depending on clinic needs.

The platform revolves around the concept of a test, which is an input or output item, within a "test list" which is a collection of tests. An example would be a daily high-dose-rate afterloader check-out with a 10-20 items to complete. The input-type tests can be entered by the user and the values can be compared with a reference value as well as a tolerance range. Each test can have independent references and tolerances. Output-type tests can also be created and linked to input tests, such as a dose calculation from ion chamber readings. These output tests can also have references and tolerances attached. Completing a test list creates an instance of that test list with all the values attached, hereafter called a session, including the performer, timestamp and other relevant details.

One of the most powerful features of the platform is being able to write Python scripts to calculate complicated output for a test. The special calculation test type can use inputs of other tests to perform basic math comparable to a spreadsheet, but also complicated calculations such as statistical variations, regression fitting or testing, retrieving data from another database, or performing image analysis such as with Pylinac [7]. The ability to write scripts within the context of other simple data records mean the user is not required to use multiple applications to perform all aspects of routine QA.

The advantages of a platform vs a spreadsheet or set of spreadsheets are several, including the ability to attach files to the session such as screenshots, log data, or reference materials. QA sessions can be reviewed by the appropriate authority later on, due to the multi-role paradigm. Data from the QA sessions can be charted over time to show intra-test trends and can also show inter-test correlations. Reports can be created within the browser and downloaded to a PDF, CSV, or Excel document so that a set of hard copies of all data is available, adding to redundancy. Email notifications can be configured to alert users about tests that had values out of tolerance, tests awaiting review, upcoming due dates, and more. A web application programming interface (API) also exists, allowing users to interact programmatically with the application, allowing for in-house applications to communicate with the server.

The platform is deployable to a number of different operating systems and compatible with a majority of the commonly-used databases including PostgreSQL [8], MySQL [9], and MSSQL. Being server-based, the platform only requires a browser to use, reducing the requirements of local resources.

3 Results

An example of entering data in QATrack+ is shown in Figure 1. Feedback from inputs is immediate and will render a red or amber warning if the value is outside the respective tolerance.

Reviewing data is a separate task in QATrack+ and allows the appropriate authority to approve, reject, and/or comment on the data session as well as attach further documentation or files, shown in 2

A chart of acquired data for several tests is shown in Figure 3. The timescale is configurable as well as the tests to display, allowing the user to compare the results of arbitrary

Perform Unit05 : Pentaguide Centred CCW

Last done: Day 2 (Pentaguide Offset CW)

Category	Name	Value	Skip	Status
MV Imaging	MV Before the Shift (Centred - CCW) - AP Horiz (mm)	-11		OK(-0.50)
MV Imaging	MV Before the Shift (Centred - CCW) - AP Vert (mm)	16.5		TOL(3.00)
MV Imaging	MV Before the Shift (Centred - CCW) - RT Horiz (mm)	11		OK(-0.50)
MV Imaging	MV Before the Shift (Centred - CCW) - RT Vert (mm)	15		OK(1.50)
kV Imaging	kV Interrupt Functional	No / Yes		OK
kV Imaging	kV Volume View Table Correction (Centred - CCW) - Lateral (cm)			Not Done
kV Imaging	kV-MV error (Centred - CCW) - Lat (mm)			Not Done
kV Imaging	kV-MV error (Centred - CCW) - Long (mm)			Not Done
kV Imaging	kV-MV error (Centred - CCW) - Vert (mm)			Not Done
MV Imaging	MV Inconsistency (Centred - CCW) - Sup-Inf Before (mm)	1.50000		OK(1.50)

Figure 1: Performing a Test List in QATrack+ demonstrating numerical & boolean input tests, as well as calculated output tests.

Review 5 Jun 2018 2a. Electron Beam Outputs

Test	Value	Reference	Comment	History 04Jun18	01Jun18	31May...	30May...	29May...	Status
									---
12 MeV 10x10 Cone 100MU	86.7	86.8		86.9	86.9	86.7	86.8	87.2	Approved
9 MeV 10x10 Cone 100MU	85.9	86		86.1	86.1	86	86	86.4	Approved
6 MeV 10x10 Cone 100MU	89.3	89.4		89.5	89.6	89.4	89.4	89.8	Approved
15 MeV 10x10 Cone 100MU	90.7	90.7		91	91	90.8	90.9	89.5	Approved

Attachments

Delete History Edit Update

Figure 2: Reviewing a QA session in QATrack+. Reference and historical values give contextual meaning to the current values.

tests across multiple machines.

The role-based permissions allow for composite and multi-role users including therapists, physicists, service engineers, and administrators to all use the platform with granular permission sets controlled by the clinic. Integration with Active Directory is also supported and allows users to log into QATrack+ via their network login, eliminating the need for local credentials.

4 Discussion

The intent of the platform is to provide a singular, comprehensive way to perform, manage, and track routine QA data. QATrack+ has been in use in numerous clinics around the world and has seen adoption particularly in budget-constrained clinics. The ability to enter data, perform calculations on that data, as well as record related data such as screenshots or reference materials all in one application reduces necessary cognitive load on the user.

Given that the platform is open-source, clinics can choose to adopt and adapt the platform to their own needs as desired and host locally in the clinic or on the cloud and invites validation from the community as well as enables budget-constrained clinics to leverage a powerful, state-of-the-field platform.

5 Disclosures

The authors are employed by Radformation, which uses QATrack+ as the base for a commercial product.

References

[1] E. E. Klein, J. Hanley, J. Bayouth, et al. *Task group 142 report: Quality assurance of medical acceleratorsa*. 2009. DOI: 10.1118/1.3190392.

[2] J. Hanley, S. Dresser, W. Simon, et al. *AAPM Task Group 198 Report: An implementation guide for TG 142 quality assurance of medical accelerators*. Oct. 2021. DOI: 10.1002/mp.14992.

[3] L. Buckley, G. Heddon, I. Byrne, et al. "Improved X-Ray Safety, Quality Control, and Resource Management in Medical Imaging Using QATrack+". *Journal of Medical Imaging and Radiation Sciences* 51 (1 Mar. 2020), pp. 22–28. DOI: 10.1016/j.jmir.2019.12.004.

[4] C. Angers, R. Bottema, L. Buckley, et al. "Streamlining Regulatory Activities Within Radiation Therapy Departments Using QATrack+". *Health Physics* 117 (3 Sept. 2019), pp. 306–312. DOI: 10.1097/HP.0000000000001119.

[5] P. Teams. *What is Python? Executive Summary*. 2023.

[6] D. S. Foundation. "Django: The Web framework for perfectionists with deadlines". *Djangoproject.Com* (2013).

[7] J. R. Kerns. "Pylinac: Image analysis for routine quality assurance in radiotherapy". *Journal of Open Source Software* 8 (92 Dec. 2023), p. 6001. DOI: 10.21105/joss.06001.

[8] T. P. G. D. Group. "What Is PostgreSQL?" *Postgresql.Org* (2021).

[9] MySQL. *MySQL Documentation*. 2018.

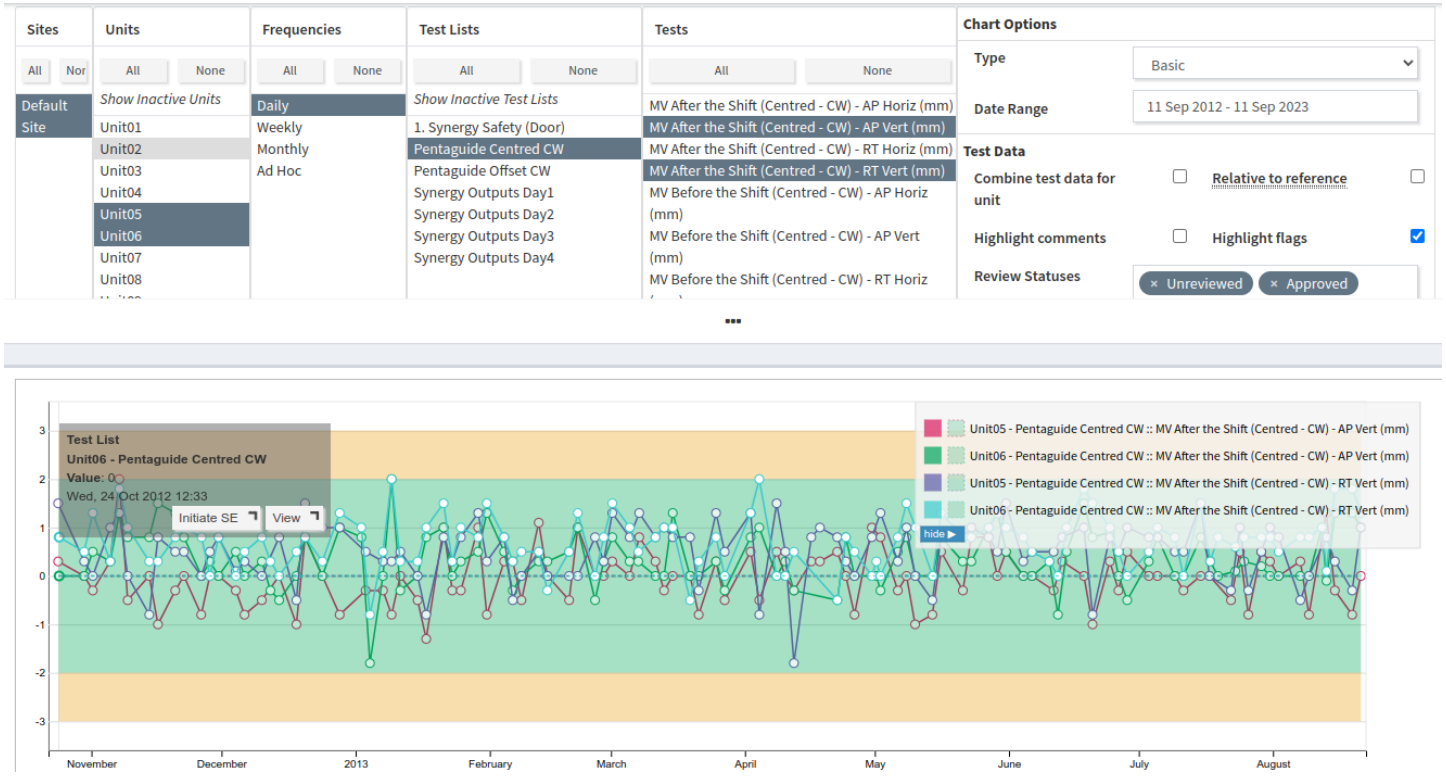


Figure 3: QATrack+'s charting feature allows you to plot data from multiple tests & machines concurrently. Reference values & tolerance bands can also be toggled.

Big data methods for CT doses analysis and monitoring

Pierre-Luc Asselin[1], Yannick Lemaréchal[2], Gabriel Couture[2], Samuel Ouellet[1], Jonathan Boivin[3], and Philippe Després[1,2,3]

[1]Département de physique, de génie physique et d'optique, Université Laval, Québec, Québec, Canada.
[2]Centre de recherche sur le cancer, Université Laval, Québec, Québec, Canada.
[3]Service de physique médicale et de radioprotection, CHU de Québec–Université Laval, Québec, Québec, Canada.

Abstract
Computed Tomography is the principal contributor to ionizing radiation exposure of medical origin for the Canadian population. The Canadian Computed Tomography Survey by Health Canada provides guidance in the safe use of radiation-emitting devices, notably by proposing Diagnostic Reference Levels (DRLs) for typical CT exams (e.g. head, chest, abdomen). Dose associated with typical studies can be compared to DRLs, but this remains a periodic one-off procedure that, although part of a quality assurance program, might fail to capture the reality of clinical activities. There is a need to monitor more closely and continuously the use of ionizing radiation in CT, as increasingly demanded by regulatory authorities. This project aims to implement big data methods to provide CT dose visualization and analytic tools to managers and technical personnel in a large institution. For instance, dashboards were deployed to provide an overview of radiation usage in CT, leading to opportunities to study clinical practice trends and facilitating the identification of outliers to improve heath care. The technology used was found to be well adapted to the analysis of massive datasets without interfering with clinical activities.

1 Introduction

Computed Tomography is currently the principal contributor to ionizing radiation exposure of medical origin for the Canadian population [1], a situation that typically prevails in developed and developing countries. CT examination requests are also on the rise, with an increase of 30% from 2010-2020 and an additional expected increase of 18% within the next twenty years [2]. The Canadian Computed Tomography Survey by Health Canada provides guidance in the safe use of radiation-emitting devices, notably by proposing Diagnostic Reference Levels (DRLs) for typical CT exams (e.g. head, chest, abdomen). Further dosimetric recommendations are also given by the Safety Code 35 to promote better radiation exposition management within health institutions [3]. Proper quality assurance programs respect such guidelines, but might still fail to capture the reality of clinical activities.

In 2016, authorities in Quebec (Canada) surveyed current dosimetric practices within the province and demanded both to better conform to Health Canada guidelines, and to monitor more closely the use of ionizing radiation in CT. Currently, no clear guidelines or best practices are proposed to health institutions in Québec for radiation exposure monitoring. Therefore, this project aims at developing a dashboard to provide visualization and analytic tools to managers and technical personnel to monitor radiation usage in CT.

2 Methods

2.1 DICOM

All developments were aligned with the Digital Imaging and Communications in Medicine (DICOM) standard, to maximize semantic and technical interoperability across workflow components. The DICOM standard is well established for the communication and the storage of medical imaging, clinical data and associated metadata. DICOM's standardized format ensures a robust and predictable structure, unlike more generic formats such as XML and JSON. DICOM is a strong foundation to build on, especially with FAIR data management principles in mind.

The DICOM standard provides more than 3600 data fields to store data and contextual information. For instance, a DICOM object could contain information such as the age of the patient, along with weight, date of birth and sex. DICOM data files can be seen as collections of standardized data fields related to a shared subject, with unique identifiers for sound data management. These files can then reference each other to better represent the complex relations between them, as expected in a medical imaging context. For example, a DICOM data file containing the actual image could reference other files pertaining to the patient or study such as segmentations, dose maps or dose reports.

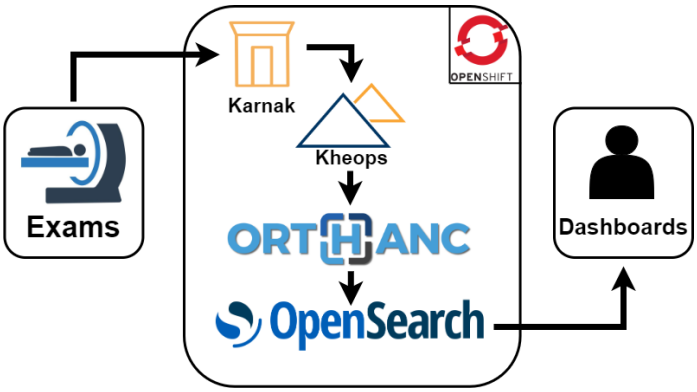


Figure 1: General flow within the PARADIM platform to expose data to OpenSearch.

2.2 Data flow

CT dose monitoring dashboards are fed by a data flow generated by routine clinical activities. Software components supporting this data flow are hosted at Université Laval within our custom-made PARADIM platform (*Plateforme d'annotation, de réutilisation et d'analyse d'images médicales* - Medical image annotation, reuse and analysis platform)[1]. PARADIM includes these components:

- Clinical data extraction and de-identification (`Karnak`)
- Data governance (auth/autz) (`Kheops`)
- Data storage (`Orthanc`)
- Data viz environment (`OpenSearch`)

`Karnak` - `Karnak` is an open-source DICOM gateway for data de-identification and DICOM attribute normalization. This software acts as a DICOM listener connected to the PACS (*Picture Archiving and Communication System*) and feeds de-identified data to Kheops through DICOMWeb. Data de-identification is supported field by field and includes: replacement by a dummy value, time-shift for dates or outright field removal. Various de-identification profiles can be managed within Karnak, including the DICOM basic privacy profile[4].
`Kheops` - `Kheops` is a DICOM-compliant image archive under MIT license offering data governance functionalities, including granular authentification/authorization (auth/autz) for data sharing for human an machine as well (through revokable API tokens). Images within `Kheops` are managed as data collections (called albums) accessible only with proper rights. Read and write permissions, as well as download and sharing rights are granted separately. A collection within Kheops was constituted for this particular project.
`Orthanc` - `Orthanc` is an open-source DICOM server used as the main DICOM database in PARADIM, with S3-compatible Ceph (RedHat) as an object storage backend. `Orthanc` provides a graphical web interface and RESTful API to request DICOM data. Moreover, `Orthanc` includes a Python (or lua) code interpreter to handle/transform DICOM data and ensure further data normalization through personalized scripts. For instance, one such script is used to collect DICOM date-time fields to guarantee conformance to DICOM standard's expected format. Another script extracts and normalizes dosimetric data from imaging exam data. Each data normalization script collects data stored within DICOM fields and reshapes it to ensure conformance to expected data format from the DICOM standard (commonly by data type conversion with the python standard library). Additionally, these scripts are automatically executed by `Orthanc` on every new DICOM study stored within its database, through a trigger mechanism.

`OpenSearch` - Following data ingestion, curated medical data stored within `Orthanc` are routed to `OpenSearch` in JSON format and in conformance with HTTP protocol (with the HTTP library for python called requests[2]). `OpenSearch`, a fork of Elasticsearch, is an open-source software under Apache 2.0 license offering data indexing, visualisation and exploration functionalities. Curated data elements are stored as documents (lists of data fields containing the relevant DICOM data), which are then clustered in indexes enabling `OpenSearch`'s powerful search engine. For the particular needs of this project, new indexes are generated daily to support expected continuous data ingestion. This index generation is supervised by a specific index pattern, ensuring structural cohesion between indexes.

2.3 Dashboards

Following data ingestion and indexation, dashboards are built within `OpenSearch` to offer flexible and user-friendly interfaces for data exploration, as shown for example in Figure 2. Dashboard offers metrics, graphs and monitoring tools tailored for specific needs. The dashboards can be accessed from any computer, assuming appropriate authentication. It is worth mentioning that dashboards do not report medical data, only aggregated metrics. Several data visualization options are available [3], as well as search and filter tools to drill down in data. Multiple dashboards can be hosted on a single `OpenSearch` instance, each with their distinct dataset. Dashboards are built according to Big Data visualization recommendations [5, 6]

2.4 Data Management

Data management best practices are integrated within PARADIM and throughout the workflow. Chief among them are the FAIR principles[4], as well as a general DataOps philosophy focusing on machine-actionability. These principles and guidelines offer guidance to data scientists and professionals on how data should be managed and integrated. For instance, in our data flows we enforce the preservation of all metadata and we rely on persistent unique identifiers, which is facilitated by the DICOM standard (which also handles the interoperability component of FAIR).

[1] https://paradim.science/

[2] https://requests.readthedocs.io/en/latest/

[3] https://datavizproject.com/

[4] https://www.go-fair.org/fair-principles/

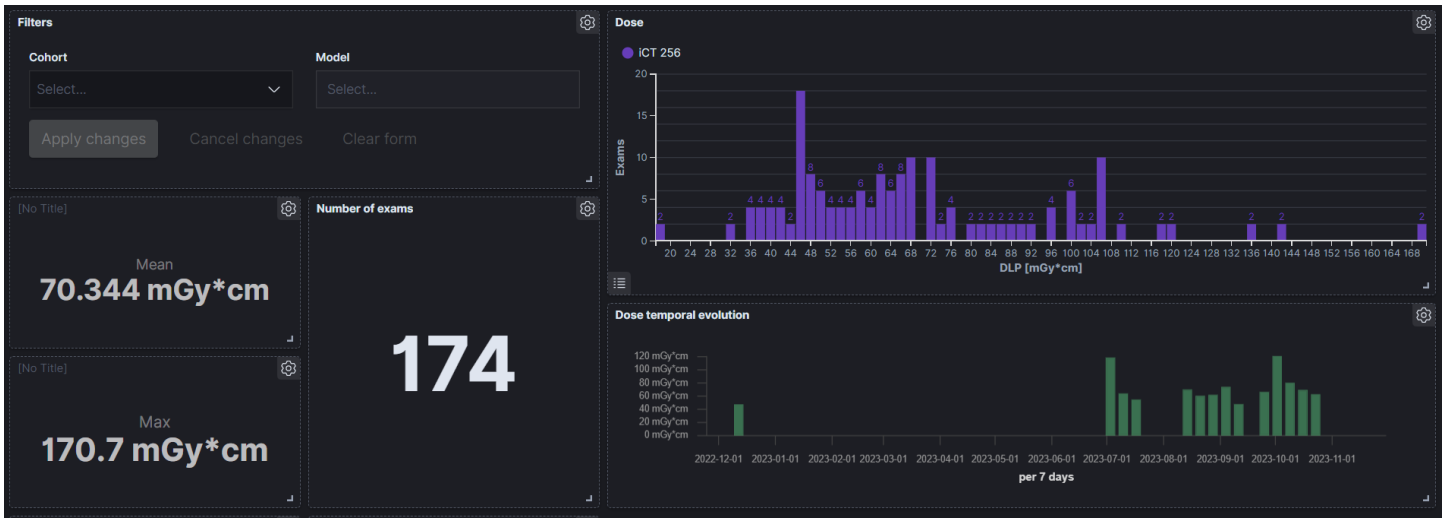


Figure 2: Dashboard example fed by the data pipeline.

3 Results

The robustness of the data pipeline was tested with the open access National Lung Screening Trial database[5]. The platform was able to ingest data at a rate of (5440 ± 145) image series per hour (from the `Karnak` gateway to visibility on dashboards), suggesting that real-time operations are possible for a large institution or even a regional/national setting.
The dashboard example presented in Figure 2 shows some available metrics, including mean and maximum dose, and stacked bar charts (for dose distribution and temporal evolution). Line graphs, pie charts, treemaps, gauge charts and tables can also be used. Additionally, dynamic filters can be integrated into the dashboard, such as the one at the upper left of the dashboard in Figure 2, which allows filtering on device model or clinical cohort. Multiple metrics, such as bar charts, can also be used to drill down to specific values with a single click.
Dashboards provide an overview of radiation usage in CT, leading to opportunities to study clinical practice trends and ease the identification of outliers to improve heath care.

4 Discussion

The proposed infrastructure is well suited to massive datasets and is designed with robustness in mind. PARADIM is deployed in a `kubernetes` environment, where applications are continuously monitored and can be restarted automatically upon failure [7].
The strict adherence to the DICOM standard enforces data normalization and preserves contextual information through rich metadata. Additionally, dosimetric data stored within DICOM radiation dose structured reports (RDSR), such as DLPs, was successfully extracted and ingested by dashboards. However, data kept within private tags is not normalized since their content is not controlled by the DICOM standard.
We have observed that access to rich dosimetric information often depends on the configuration of CT devices. By default, some do not send dosimetric data to the PACS or generate results that cannot be exploited easily (for example dosimetric reports added as images, requiring OCR to extract data, or dosimetry data in non-standard DICOM field). Furhtermore, manufacturers unfortunately offer distinct ways to store dosimetric data despite the availability of the Patient Radiation Dose Structured Report[8, 9]. This heterogeneity is a barrier to large scale dosimetric analyses.

5 Conclusion

A dashboard fed by secure data pipelines was developed to aggregate and visualize dosimetric data from clinical CT studies. The dashboard provides an overview of radiation usage in CT, give rise to opportunities to study clinical practice trends and ease the identification of outliers to improve heath care. Data ingestion is continuous and does not interfere with clinical workflow. Likewise, good data management practices are respected and integrated within the workflow. Moreover, the tool is compliant end-to-end to DICOM standard and uses only open-source, non-proprietary software. Next steps include the development of alerts for events (e.g. high dose, repeated exams) and the integration of data from other centers to generate a personalized report on cumulative dose from CT studies.
Additional work to promote sound data management within health institutions should be considered to maximize the tool's potential. Proper data management include data normalization, compliance to standards (such as DICOM) and richer metadata. Better compliance to the standard by vendors regarding dosimetry appears highly desirable.

[5] https://wiki.cancerimagingarchive.net/display/NLST

References

[1] UNSCEAR. *2008 Report to the General Assembly with Scientific Annexes, Volume I: Sources, Annex A: Medical radiation exposures.* Tech. rep. United Nations, 2008.

[2] A. M. Yi-Sheng Chao Alison Sinclair. "The Canadian Medical Imaging Inventory 2019 - 2020". *Canadian Journal of Health Technologies* 1.1 (2021), p. 215.

[3] H. Canada. "Safety Code 35: Safety Procedures for the Installation, Use and Control of X-ray Equipment in Large Medical Radiological Facilities" (2008), p. 88.

[4] NEMA. *DICOM Attribute Confidentiality Profiles*. URL: https://dicom.nema.org/dicom/2013/output/chtml/part15/chapter_E.html#table_E.1-1.

[5] A. Vellido. "The importance of interpretability and visualization in machine learning for applications in medicine and health care". *Neural Comput & Applic* 32.24 (2020), pp. 18069–18083. DOI: 10.1007/s00521-019-04051-w.

[6] bibinitperiod C. A. A. L. Wang G. Wang. "Big Data and Visualization: Methods, Challenges and Technology Progress". *Digital Technologies* 1.1 (2015). DOI: 10.12691/dt-1-1-7.

[7] A. B. et al. "A scalable, secure, and interoperable platform for deep data-driven health management". *Nat Commun* 12.1 (2021). DOI: 10.1038/s41467-021-26040-1.

[8] G. M. Systems. *Computed Radiography DICOM Conformance Statements*. URL: https://www.gehealthcare.com/products/interoperability/dicom/computed-radiography-dicom-conformance-statements.

[9] S. Healthinners. *DICOM Conformance Statements - SOMATOM Scope*. URL: https://www.siemens-healthineers.com/services/it-standards/dicom-conformance-statements-computed-tomography/somatom-scope.

AutoMRISimQA: an automated system for daily quality control of a 3T MRI simulator

Aitang Xing[1,2,3], Gary Goozee[1,2,3], Gary Liney[1,2], Sankar Arumugam[1,2,3], Shrikant Deshpande[1,2,3], Anthony Espinoza[1,2,3], Alison Gray[1,2,3], Vasilis Kondilis[1], Doaa Elwadia[1,2], Robba Rai [1,2], Phillip Chlap[1,2,3], and Lois Holloway[1,2,3]

[1]Liverpool & Macarthur Cancer Therapy Centre, Liverpool, NSW 1871, Australia
[2]Ingham Institute for Applied Medical Research, Liverpool, NSW 1871, Australia
[3]South West Sydney Clinical Campuses, University of New South Wales, Sydney, Australia

Abstract A software system named AutoMRISimQA was developed to monitor the daily performance of a wide-bore 3T scanner (MRI) which was designed and dedicated to radiotherapy simulation. AutoMRISimQA can monitor the performance of the MRI simulator not only by using image quality indices such as signal-to-noise ratio (SNR), uniformity, ghosting, and contrast but also performing a quick check of geometric accuracy as well as the external lasers quantitatively. It was implemented into the daily clinical workflow in 2013 and has been used for more than 10 years. It was also seamlessly integrated with QATrack+, allowing continuous monitoring of the consistency of the MRI simulator's performance.

1 Introduction

MRI simulator (MRISim) is a magnetic resonant imaging (MRI) scanner designed and dedicated for radiotherapy planning, dose calculation and treatment response [1, 2]. The use of MRI in radiotherapy requires the scanner to be more stable and reproducible than the one used for diagnostic purposes. A recently published TG 284 report recommended a list of tests to be done daily, but how to perform these checks depends on the individual centre [3].
MRISim was commissioned and integrated into the routine workflow in 2013 in our centre [4, 5]. A fully automated daily quality control (QC) program was developed and has been running for more than 10 years. The purpose of this paper is to report how the system was designed, implemented and integrated into a radiation oncology workflow.

2 Materials and Methods

2.1 MRI Simulator and phantom

MRI simulator is a wide-bore 3T MRI scanner (Siemens 3T Skyra, Erlangen, Germany) equipped with an external simulation laser and a flat couch [5]. The phantom is the AQUARIUS phantom (MRI-0005, S/N 0010260016, LAP of America, Laser application, USA), which was originally designed for checking the alignment of simulation lasers [4]. The phantom is cylindrical, and inside the phantom, there are four long tubes filled with CuSO4 solution and the rest is filled with water. As shown in Figure 1, the phantom was set on the flat table aligned with external lasers and then scanned using a T1 weighted spin echo sequence (echo time(TE) =30ms and repetition time(TR)=3410ms) daily. The MRI images were transferred to a shared network location on a server. An in-house system, named AutoMRISimQA was developed to monitor the MRI simulator's daily performance using phantom images.



Figure 1: Siemens Skyra 3T MRI Simulator with a flat couch top. The AQUARIUS phantom was set up on the couch and aligned with external lasers.

2.2 Design of AutoMRISimQA system

The system was designed and divided into six modules. Figure 2 showes the overall structure of the system. The first module is used to detect if a new image series arrives. If the new images were found, the second module will be triggered to sort image series, according to the acquisition date and time. The image pixel data were also extracted. The second module uses these data to calculate image quality indices such as signal-to-noise ratio(SNR), which characterize the performance of the MRI scanner. The calculated QC parameters are passed to the third module, whose role is to make a logical judgement. The parameters are compared with their corresponding reference values and tolerance. If any of the QC parameters are out of tolerance, an alert e-mail will be automatically sent to the responsible physicists. A PDF report was also generated at the same time to assist physicists

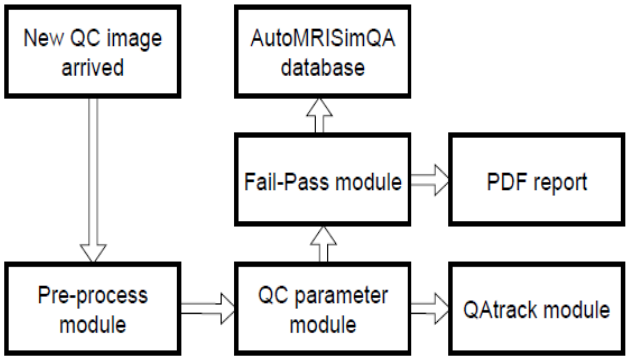


Figure 2: The modular structure of AutoMRISimQA

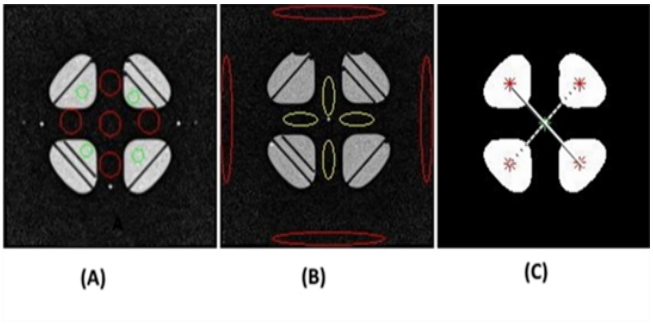


Figure 3: The centre slice of phantom images and ROIs used to calculate QC parameters.

in quickly diagnosing where the problem is. At our centre, all QC data for radiotherapy equipment were managed using QATrack+ [6], an open-source web-based system designed for radiotherapy quality control. The QATrack module is used to automatically update the database of QATrack+.

2.3 Implementation of the system

Matlab, Java and Python were used for system implementation. As QATrack+ was implemented in pure Python, the module for updating the QATrack+ database was written using Python. The access to the database was through the Python representational state transfer(REST)framework [7]. In this way, the risk of accidental damage to the database was minimized. Except for the report module, the rest of the system was implemented using Matlab. A rich of imaging process functions provided by Matlab's image processing toolbox is the major consideration of using Matlab as the main programming language. Matlab also can seamlessly integrate Java applications into the Matlab program. The report module was written in Java using the iText package [8]and integrated into Matlab.

The core component of the system is the one used for calculating the QC parameters from the phantom images. The image slice shown in Figure 3 (A) was used to calculate the uniformity and SNR using the four $1cm^2$ red circle regions. The percentage intensity uniformity (PIU) is defined as

$$PIU = 100 \times (1 - \frac{S_{max} - S_{min}}{S_{max} + S_{min}}) \quad (1)$$

Where S_{max} and S_{min} represent the maximum and minimum signal averaged over four red circles.

The SNR is calculated as the ratio of the mean signal averaged over four areas defined by the red circles to the standard deviation over the same ROI. Another parameter, termed output (MRI signal), is defined as the ratio of the standard deviation to the mean of the pixel values over the red circle ROI at the centre of the images.

The ghosting effect was calculated using the image slice and eight elliptical regions as indicated in Figure 3(B). The ghosting ratio (GR) was calculated as:

$$GR = \frac{|((S_{top} + S_{btm}) - (S_{top} + S_{btm}))|}{2S_{mean}} \quad (2)$$

Where S_{top} , S_{btm} , S_{left} and S_{right} are the signals averaged large eclipses at the top, bottom, right and left of the images, respectively. S_{mean} is the mean signal averaged over four small ellipsoids at the center of the image.

The contrast was calculated as the ratio of the mean signal value over areas defined by the four green circles in high contrast regions to the mean signal value determined over the four red circle regions. The image was segmented to isolate the four high signal regions and then the centroids of these regions were found. Two diagonal distances between these centers were calculated as shown in Figure 3(C). The forward and backward diagonal distance was recorded as d45 and d135 for reporting purposes. The laser coordinates Y is the slice position, where laser coordinates X and Z are the center of the central image slice, which were calculated using centroid coordinates of four high contrast regions.

3 Results

The system has been implemented into the MRI-based radiotherapy workflow since 2013. As a start of a daily routine, a radiographer or radiation therapist (RT) scans the laser phantom. Then the images are sent to a location sitting on a shared network drive. The system automatically processes the images and updates the QATrack+ database. The daily QC results can be viewed on any computer connected to the hospital's private network using the web-based interface of QATrack+ or the PDF report.

As an example, Figure 4 showed one QC result generated by autoMRISimQA results. The traffic lights were used to indicate the acceptability of the results along with the tolerance values.

The main setting for the system can also be configured via a text file as shown in Figure 5. The use of configuration files makes the system more flexible and maintainable.

The statistical analysis based on the data acquired over a certain period provided useful information regarding the MRI's performance and a tool for continuous monitoring of MRI scanner performance. As an example, Figure 6 shows the trend of SNR and MRI signals over a few months, which were generated by QATrack+.

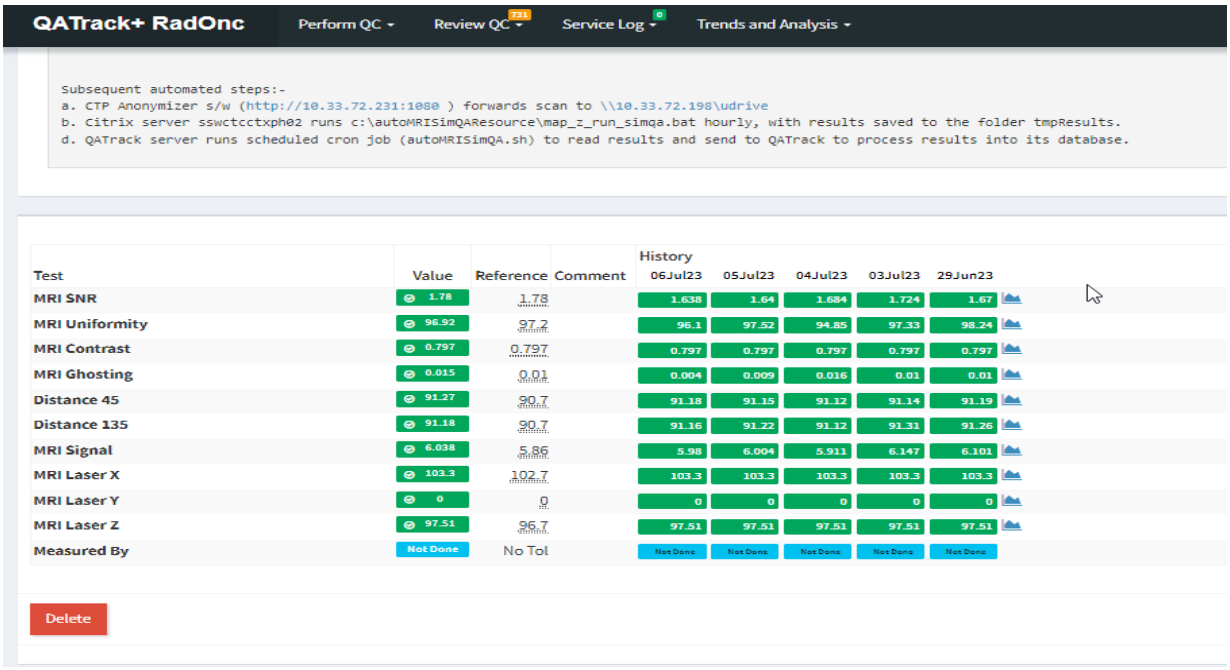


Figure 4: Updated daily QC results displayed through the QATrack+ web interface.

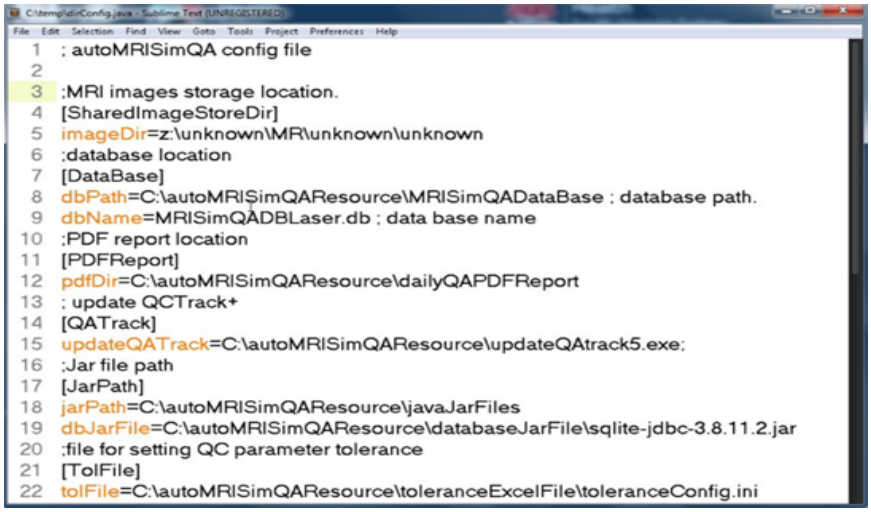


Figure 5: Configurability of the system using a simple text file.

4 Discussion

AutoMRISimQA achieved an automated constancy check of MRISim's performance by monitoring MRI imaging system in one go using multiple image quality indices, which are directly related to the use of MRI in radiotherapy planning and recommended by TG 284 [3], but also doing a quick check of geometric accuracy. In addition to the visual check of simulation laser alignment, the laser positions were also quantitatively monitored.

The work presented here demonstrates a feasible solution for developing an automated daily QC program for an MRI simulator, which can be used and implemented by other centres. The source code of the system is available on the request and will be on GitHub soon. Although the system implemented is based on the laser phantom, other commercial phantoms or custom-made phantoms can be used.

Several authors developed software tools for quality assurance of a diagnostic MRI scanner using ACR MRI phantom [9, 10]. An open-source image processing tool, IQWorks, was also used for MRI quality assurance purposes [11]. In comparison with these tools, autoMRISimQA has several unique features: (1) the system was fully automated once the phantom images were transferred to the network location from the MRI control console; (2) It was seamlessly integrated with an open-source QC data management system. The daily QA results and the trend of the system's performance can be viewed on any computer connected to the hospital network; (3) Automatic email notification allows physicists monitor the MRISim and take action only when it is required.

The clinical benefit of implementing the AutoMRISimQA system into the clinical workflow is efficiency, especially for a busy clinic with a large number of patients going through

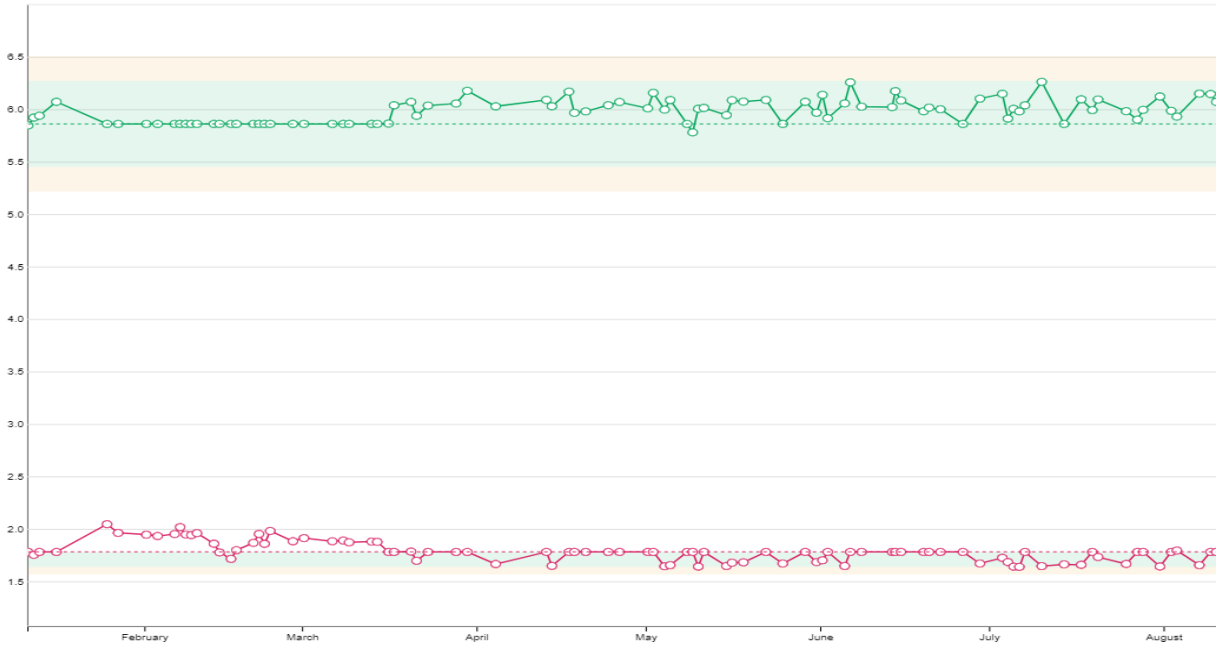


Figure 6: Continuous monitoring of MRI simulator using autoMRISimQA and QATrack+ web interface. SNR (green curve) and MRI signal (red curve) were generated by QATrack+.

the scanner. Before the use of this system, the RTs needed to do two QC tests daily: a customer QC using a uniform phantom and the check of the alignment of simulation lasers using the laser phantom. The average time spent on morning daily QC was half an hour. Now it was reduced to four minutes, the scanning time of laser phantom. Furthermore, the continuous monitoring of MRISim's performance ensures the short, medium and, especially, long-term follow-up of technical characteristics and their possible drift with time.

5 Conclusion

A software system was designed, and implemented for daily quality controls of an MRI simulator. As an open-source tool, the system provides a prototype solution for monitoring the performance of a dedicated MRI scanner used in a radiation oncology department.

References

[1] R. Christiansen, H. Jensen, and C. Brink. "Magnetic Resonance–Only Workflow and Validation of Magnetic Resonance–Based Dose Calculations for Radiation Therapy of Prostate Cancer". *International Journal of Radiation Oncology, Biology, Physics* 96.2 (2016), E645.

[2] A. D. King and H. C. Thoeny. "Functional MRI for the prediction of treatment response in head and neck squamous cell carcinoma: potential and limitations". *Cancer imaging* 16 (2016), pp. 1–8.

[3] C. K. Glide-Hurst, E. S. Paulson, K. McGee, et al. "Task group 284 report: magnetic resonance imaging simulation in radiotherapy: considerations for clinical implementation, optimization, and quality assurance". *Medical physics* 48.7 (2021), e636–e670.

[4] A. Xing, L. Holloway, S. Arumugam, et al. "Commissioning and quality control of a dedicated wide bore 3T MRI simulator for radiotherapy planning". *Int J Cancer Ther Oncol* 4.2 (2016), pp. 1–10.

[5] A. Xing, G. P. Liney, L. Holloway, et al. "Quality assurance of the radiotherapy workflow integrating a dedicated wide-bore 3T MRI simulator". *World Congress on Medical Physics and Biomedical Engineering, June 7-12, 2015, Toronto, Canada*. Springer. 2015, pp. 539–542.

[6] R. Studinski, R. Taylor, C. Angers, et al. "SU-E-T-103: Development and Implementation of Web Based Quality Control Software". *Medical Physics* 41.6Part12 (2014), pp. 246–246.

[7] *Python REST API Framework*. http://python-eve.org/. Accessed: 2024-1-1.

[8] *Java iText package*. https://itextpdf.com/. Accessed: 2024-1-1.

[9] J. Sun, M. Barnes, J. Dowling, et al. "An open source automatic quality assurance (OSAQA) tool for the ACR MRI phantom". *Australasian physical & engineering sciences in medicine* 38 (2015), pp. 39–46.

[10] J. Yung, W. Stefan, D. Reeve, et al. "TU-F-CAMPUS-I-05: Semi-Automated, Open Source MRI Quality Assurance and Quality Control Program for Multi-Unit Institution". *Medical Physics* 42.6Part36 (2015), pp. 3647–3647.

[11] J. Fazakerley, J. Brunt, J. Graham, et al. "The application of IQWorks image analysis software to quality assurance in MR imaging". *Biomedical Physics & Engineering Express* 2.2 (2016), p. 025011.

NoPAUSE: Feasibility of a general-purpose, physician-initiated automatic VMAT treatment planning system for curative treatments.

Nick Chng[1], Stay Miller[1,4], Quinn Matthews[1], Fred Cao[2], Maryam Golshan[3], and Robert Olson[1,4]

[1] BC Cancer – Prince George, Canada
[2] BC Cancer – Surrey, Canada
[3] BC Cancer – Kelowna, Canada
[4] Department of Medicine, University of British Columbia, Vancouver, Canada

Abstract

The Northern Plan Automation Services (NoPAUSE) project is a collection of in-house applications and scripts designed to improve efficiency and quality in radiotherapy treatment planning. The Treatment Planning Automation System (TPAS) is one of these services, and enables fully automatic VMAT plans to be produced within our Eclipse/ARIA environment. It supports a physician-initiated workflow that allows contouring, planning, and approval to occur in one encounter. This workflow is deployed at three centres within our institution for rapid-palliative VMAT. The purpose of this work is to investigate extending our system to curative protocols, beginning with prostate, rectum, and H&N. Retrospective clinical cases were used to tune the system. After tuning, a further five validation cases for each site were used to compare auto-plans to the manual clinical ones. For each site, TPAS produced clinically-acceptable plans that compared favorably to the matched retrospective control, and demonstrated comparable quality from a DVH perspective.

1 Introduction

Inverse treatment planning for IMRT or VMAT techniques is time-consuming, and significant contributor to the total delay between simulation and treatment. This is particularly the case when including delays associated with hand-offs between the physician who completes the contours and reviews the final plan, and the planning departments that are generally responsible for optimization. In busy clinics, many physicians opt to defer non patient-facing tasks like contouring and plan review until after hours, implying that at least a full business day may pass between those activities.

The Northern Planning Automation Services project is a collection of in-house applications and scripts designed to improve efficiency and quality in radiotherapy treatment planning. The core platform is the Treatment Planning Automation System (TPAS), which fully automates VMAT planning in Eclipse. There are many domains in which planning automation can be implemented[1]. In the present context the automation is inclusive of field-placement, planning structure creation, optimization, and dose constraint reporting. Physicians launch TPAS, submit a request, and receive an optimized plan for review and approval in a single encounter.

TPAS was originally deployed as an automation tool specific to rapid-palliative protocols, including two clinical trials. In these settings, it is considered sufficient for the plans to meet the constraints of those protocols for TPAS plans to be clinically acceptable. However, it is often possible to do much better than the hard constraint, so before expanding the system for curative treatments, it is desirable to ensure that the plans produced by the system are of comparable quality to those optimized manually. The purpose of this study was to assess the feasibility of extending TPAS using an RO-initiated workflow for curative protocols.

2 Materials and Methods

TPAS is implemented in C# for the Varian's Eclipse™ treatment planning system, using their ESAPI library. The system can be broadly divided into three components: The plan request submission interface, the distributor, and the planning servants. The specific approach to optimization for a given site or prescription is configured through a combination of user-editable (xml) protocol file and an associated Eclipse Clinical Protocol. TPAS also supports the use of Varian's RapidPlan™ product to determine appropriate optimization objectives, where such models exists. Protocol files can also specify logic to guide the automatic rescaling of priorities and re-optimization to address scenarios where clinical goals are not met.

Submission

The submission interface (Fig. 1) is a light-weight, single page form that begins the process. The user selects a protocol from a pull-down menu, which brings up the targets and organs-at-risk (OARs) relevant to the protocol. An aliasing system automatically maps detected patients contours to the protocol definitions when naming conventions (e.g. AAPM TG-263) are followed, but the interface allows this to be overridden if necessary. If the protocol allows variation in the prescription or fractionation, these adjustments can also be applied. Once this is complete, the request can be submitted and the interface closed. In uncomplicated cases, this typically takes a physician between 15-30s, after which they can close the form and continue other tasks.

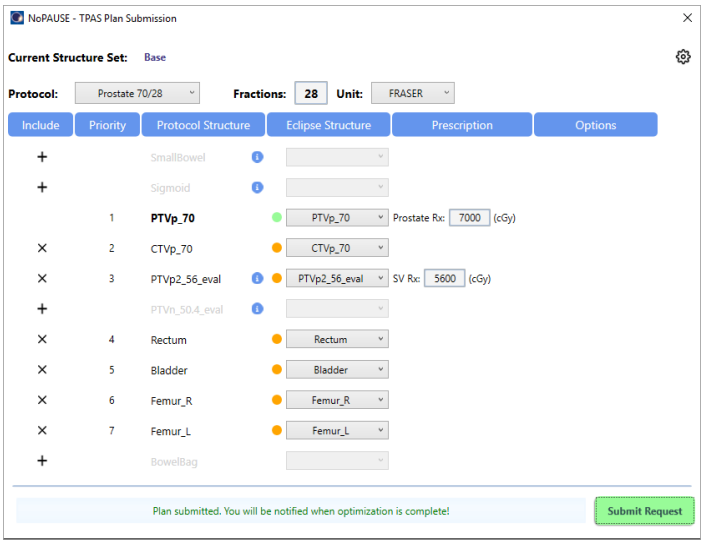


Figure 1: The submission interface, showing a request for a plan to be created according to the Prostate 70 Gy / 28# protocol. The only responsibility for the user-physician is to map the patient's contours to the protocol definitions, and adjust the prescription dose if necessary.

Distribution

The request is picked up by a central distributor, which sends an email to the user confirming the request, including the prescription and structure mapping. When the plan is completed by the planning servants, the distributor is also responsible for notifying the physician and staff subscribed to the protocol. The email includes a summary of the dosimetric clinical goals that were achieved or violated, as defined in the clinical protocol. The total planning time for a 2-arc TPAS plan is approximately 10-15 minutes from submission to completion.

Planning Servants

The planning servants are implemented as background ESAPI services. Each servant is capable of performing one plan per patient at a time. The distributor/servant model allows TPAS easily scales to larger departments with the installation of more planning servants, although this will eventually be limited by competition for optimization and dose calculation computation resources.

Plan Quality Validation

To assess plan quality against our clinical baseline, seventy-five plans across three sites (H&N, Prostate, Rectum) were retrospectively re-planned using TPAS. For the H&N site, our locally trained RapidPlan model was used. For each site, twenty plans were used to refine the configuration of the protocols, and the remaining five plans were used for validation. Plan quality was compared using the DVH quality metrics specific to each protocol.

3 Results

All plans in the validation-sets were considered clinically acceptable based on review of the 3D dosimetry and field/plan parameters by the principal author. There were minor variations in coverage between clinical and automatically planned cases. The average target coverage (V95%) for automated [manual] plans for each site was HN: 98.7% [99.1%], Prostate: 98.8% [98.9%], Rectum: 99.6% [99.1%]. High priority OAR doses were also comparable, although they trended lower for second-tier constraint doses where the TPAS protocol pushed harder than may have been deemed necessary in the clinical setting. For example, the average rectal D0.035% for prostate was 99.3% [99.5%], whereas the rectal V83% was 9% [13.4%]. In the H&N, minimizing salivary gland dose is usually a focus, and this was reflected by comparable results: Mean doses for combined parotids were 28.1 Gy [29.6 Gy]. Cord PRV D0.035cc doses were 41.5 Gy [47.1 Gy], and true cord D0.035cc was 35.7 Gy [42.0 Gy].

Figures 2-5 illustrate the relative dosimetric quality between TPAS and the retrospective clinical cases for the validation cases. Each column of the graph represents performance in one DVH metric, with the clinically acceptable constraint indicated by a horizontal bar. Upper constraints are indicated with a downward red arrow, and lower (target) constraints are indicated with an upward green arrow. Error bars on each column represent +/- one standard deviation.

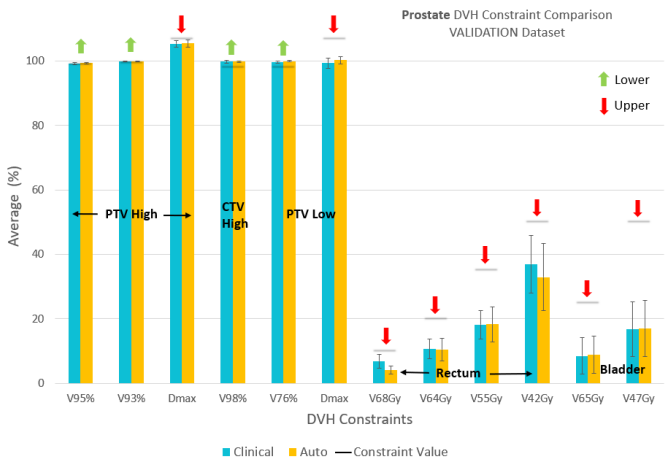


Figure 2 : Comparison of Prostate (70 Gy / 28 #) DVH quality metrics between automated and clinically optimized plans.



Figure 3 : Comparison of Rectum (50 Gy / 25 #) DVH quality metrics between automated and clinically optimized plans.

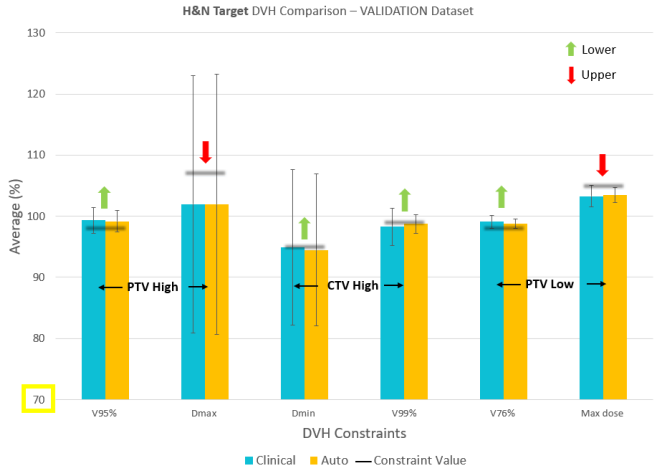


Figure 4 : Comparison of H&N (70 Gy / 35 #) DVH quality metrics for targets between automated and clinically optimized plans.

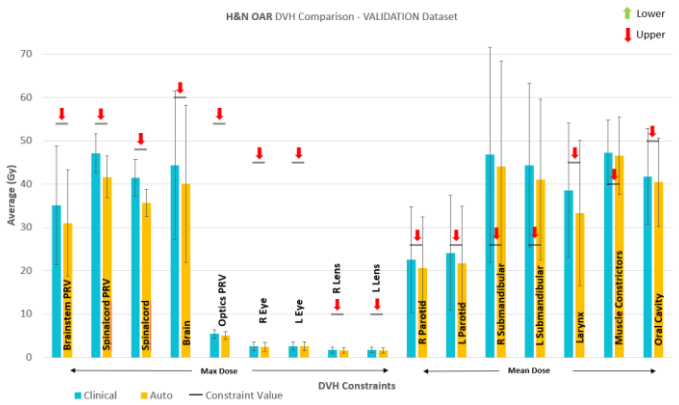


Figure 5 : Comparison of H&N (70 Gy / 35 #) DVH quality metrics for OARs between automated and clinically optimized plans.

4 Discussion

The quality of TPAS plans was broadly comparable to the clinical cases. For the prostate and rectum sites in particular, target coverage was very similar, with a slight trend toward higher (but clinically acceptable) maximum doses in the automated plans. This was attributed to greater sparing of adjacent OARs. TPAS protocols are highly configurable - in this study optimization objectives were used that initially prioritized OAR sparing, with rescaling of target priorities upward in subsequent iterations if coverage was not met. The patterns of heterogeneity in the targets were consistent with 'pushing' coverage later in the optimization process.

There was greater separation in average OAR sparing for the H&N protocols than for prostate and rectum. This is likely due to the way that TPAS segregates OARs into near-, mid- and distant (out-field-regions) and optimizes each region individually. Sparing in the mid- and distant- lower dose regions accounted for much of the average reduction. It is possible that manual planners did not focus on these regions once the overall OARs met clinical constraints.

This study was not powered to test for statistically significant differences between TPAS and manual planning. However, our results do suggest that such studies are worth pursuing with a larger sample size. The concept of clinical acceptability[2] is poorly defined and often institutionally specific and dose-volume metrics are incomplete characterization of plan quality. It is not clear whether the automated plans being produced using the protocols we developed would ultimately be preferrable to physicians, and DVH analysis can obscure features of the 3D distribution which may be important. It is likely that this will need to be further investigated in a prospective setting with physician preference as the endpoint.

5 Conclusion

This work suggests the TPAS system is a viable approach to RO-initiated automatic planning in a multi-site curative setting. Work is presently underway to assess the acceptability of TPAS plans for a variety of sites in a prospective setting.

References

[1] P. Meyer, M-C Biston, C Khampan et al. "Automation in radiotherapy treatment planning: Examples of use in clinical practice and future trends for a complete automated workflow", Cancer Radiother. (2021), DOI: 10.1016/j.canrad.2021.06.006

[2] H. Baroudi, K. Brock, W. Cao et al. "Automated Contouring and Planning in Radiation Therapy: What is 'Clinically Acceptable'?", Diagnostics (2023), DOI: 10.3390/diagnostics13040667

A complete portfolio for automated treatment planning for whole breast radiotherapy

Hana Baroudi[1], Leonard Che Fru[2], Deborah Schofield[2], Dominique L. Roniger[2], Donald Hancock[2], Kent A. Gifford[2], Tucker Netherton[2], Joshua S. Niedzielski[2], Adam Melancon[2], Manickam Muruganandham[2], Simona F. Shaitelman[3], Sanjay Shete[1], Melissa P. Mitchell[3], Laurence E. Court[1-2]

[1]The University of Texas MD Anderson Cancer Center UTHealth Houston Graduate School of Biomedical Sciences, Houston, Texas, USA

[2]Department of Radiation Physics, [3]Department of Radiation Oncology, [4]Department of Breast Radiation Oncology, Division of Radiation Oncology, The University of Texas MD Anderson Cancer Center, Houston, Texas, USA

Abstract This study aims to develop automated 3D conformal radiation therapy planning for the treatment of intact breast, considering individual patient factors and available resources.

In total, three automated treatment approaches were created. These comprise conventional tangents for whole breast treatment, and two options for additional regional lymph nodes treatment: wide tangents photon fields with supraclavicular nodes (SCLV) field, and a regular photon tangents field with a matched field to treat the internal mammary nodes, and a SCLV field. Each approach includes the choice of single or two isocenter setup (with couch rotation) to account for a full range of patient sizes.

All algorithms begin by generating automated contours for breast CTV, regional lymph nodes, and organs at risk using in-house nn-Unet deep learning models. Treatment fields are then automatically generated. Then gantry angles and field shapes are optimized, ensuring coverage of target contours while limiting dose to nearby organs. Algorithms were integrated into the Raystation treatment planning system, and then tested on 10 whole breast patients for clinical acceptability. Evaluation criteria included ensuring adequate coverage to targets and adherence to dose constraints for normal structures.

For the treatment of breast, the algorithms successfully replicated clinical plans, meeting dosage goals for all 10 cases. For breast and regional lymph nodes treatment, 5 out of the 10 plans met all preferred objectives. For the rest of the plans, axillary nodes coverage or lung or heart dose constraints were not met indicating that a different choice of treatment choice should be considered.

This study showcases the feasibility of a comprehensive automated treatment planning model for whole breast radiotherapy, accommodating diverse treatment paradigms.

1 Introduction

Radiation therapy is an integral part of breast cancer treatment, improving patients' outcomes and survival[1]. The incidence of breast cancer diagnoses is steadily increasing, with an anticipated 40% rise from the current rate by 2040. Despite this surge in demand, the radiation therapy workforce is projected to grow by only 25%, falling short of meeting the escalating requirements[2]. This imbalance is expected to be more pronounced in low- and middle-income countries, where the patient-to-oncologist ratio is already disproportionately high (1,000 patients for every 1 oncologist, compared to 150 patients for every 1 oncologist in developing countries)[3]. Treatment planning for radiotherapy is time-consuming and requires specific expertise. With the projected workforce shortage, planning will fall behind, and patients will have to wait longer to start their treatment. Even a four-week delay in adjuvant radiotherapy treatment for breast cancer has been associated with an increase in mortality rate[4]. Therefore, it is important to mitigate the consequences of the workforce shortage in treatment planning for breast radiotherapy. We propose the development of automated tools to facilitate the selection of the most suitable method for individual patient treatment, taking into account both clinical needs and available resources.

Various approaches exist for radiation therapy treatment for intact breast, the most common yet complex being 3D conformal radiation therapy (3DCRT). One comprehensive treatment with this approach consists of two wide opposing fields tangential to the breast, covering the internal mammary lymph nodes and an additional anterior field to cover the supraclavicular nodes. The extensive anatomical variation from patient to patient adds to the complexity of the individualized treatment and to the generalization of an automated approach.

One of the initial strides in automated planning was taken with a heuristic approach described by Purdie et al. [5]. This method necessitates the placement of BB markers at specific locations around the breast for functioning. Addressing this limitation, Sheng et al. introduced a more recent model utilizing a machine learning approach [6] . It is noteworthy, however, that their model is tailored to a standard breast tangent scenario and does not encompass considerations for lymph nodes. On the commercial front, Radformation's EZfluence provides a viable solution for automating breast treatment plans through fluence optimization. Nevertheless, it's important to acknowledge that, while effective, this solution may lean towards the higher end in terms of cost, which could pose a challenge for clinics with limited resources.

In this work, we propose an automated planning tool that provides a complete portfolio of treatment approaches, thus allowing the clinician to choose the approach based on the patient, their available resources and preferred clinical approach. This work is part of the Radiation Planning Assistant project, a web-based autocontouring and planning tool that is to be offered to clinics in low- and middle-income countries for minimal (likely zero) cost.

2 Materials and Methods

I-Algorithm Development:

All algorithms begin with automated contouring. The automated contours are generated using in house nn-Unet deep learning contouring models[7]. The generated target contours comprise of breast critical target volume (CTV),

supraclavicular lymph nodes CTV (SCLV), internal mammary nodes CTV (IMN) and axillary lymph nodes level I, II and III CTVs (AX1, AX2, AX3). The organs at risk (OARs) automated contours include heart, left ventricle, ipsilateral lung, trachea, esophagus, humerus head with acromioclavicular joint, cricoid cartilage, contralateral breast, and ipsilateral brachial plexus. All algorithms were implemented in the Raystation treatment planning system using scripting.

Three distinct automated treatment approaches were developed following clinical practice: Approach I; regular intact breast tangents, Approach II; wide tangents photon fields (breast + IMN) with SCLV field, Approach III: regular tangents with IMN electron matched field and SCLV field. For each approach, single and two isocenter solutions (with couch rotation) were developed to meet different clinical preferences, machine capabilities, and patient anatomy. An overview of the algorithm workflow is summarized in figure 1.

Isocenter Placement:

Single isocenter approach for breast and SCLV fields: The isocenter is placed on the caudal end of the SCLV auto-contour with a 7mm margin and centered between the breast structure's medial/lateral and anterior/posterior extent. If the distance between the isocenter and the inferior edge of the breast field is more than the jaw size limit of 20cm then the isocenter is moved inferiorly to meet this limit.

Two-isocenter approach for breast and SCLV fields: The isocenter of the SCLV field is placed as described above. The isocenter for the breast field is placed at the midpoint of the breast auto-contour. In this case, a couch rotation is applied to the tangents fields to achieve superior match with the SCLV field.

IMN field: The IMN isocenter is placed at the center of the IMN auto-contour then moved to 5cm away from the skin, in the beam's direction, for a treatment at an extended 105 cm SSD.

Aperture Generation:

For the SCLV field, the superior border is set to the inferior border of the cricoid cartilage. The inferior border is set on the isocenter. The medial border is closed on the SCLV contour with MLCs blocking the trachea. The lateral border is user defined with a default of 1.5cm from the AX1 contour. A humeral head and acromioclavicular joint block is created to protect the structure. An example of a SCLV aperture is shown in figure 2.

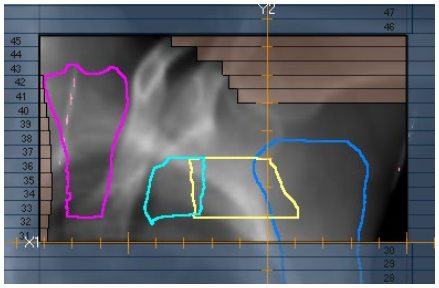


Figure 2. Beam's Eye View of a SCLV field aperture with a trachea and humeral head block. Automated contours are shown from left to right, SCLV, AX3, AX2, and AX.

For the breast field, the superior border is matched at the SCLV isocenter. The inferior border incudes a 2cm margin from the breast contour. A 2cm flash is created anteriorly to the breast contour. The posterior border is set to the target borders with a 7mm margin.

For the IMN field, the superior and inferior borders match the breast field. The medial border is placed 2 cm from the IMN contour. The lateral border is matched at the skin with the breast field.

Gantry Angle Selection:

For the SCLV field, the gantry angle is chosen within the range of 10 to 20 degrees (340-350 degrees for the left side). For each angle, the aperture is defined, and the minimum dose to the SCLV auto-contour, as well as the maximum dose to the esophagus, are calculated. The angle that optimizes coverage for the SCLV target while

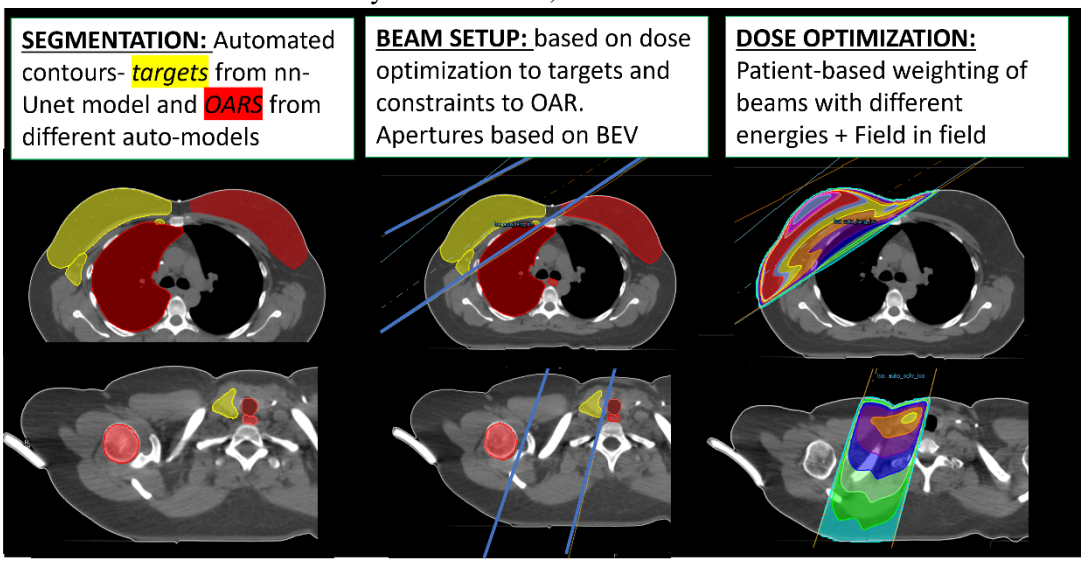


Figure 1. Design of the automated planning workflow for whole breast cancer radiotherapy treatment

ensuring that the maximum dose to the esophagus remains below 38 Gy is then selected.

For the breast field, the gantry angles for the medial and lateral fields are chosen from a range of 25 to 75 degrees (285-335 degrees for the left side). For each medial angle, a set of lateral angles is iterated over, and the volume of the ipsilateral lung receiving a 20 Gy dose, along with the dose to 5% of the volume of the contralateral breast, is calculated. The angles that minimize the volume of the ipsilateral lung receiving 20 Gy are selected. If the dose to 5% of the volume of the contralateral breast exceeds 5 Gy, angles that deviate from the direction of this structure are chosen. This process is repeated until the value of the dose to 5% of the volume is reduced to less than 5 Gy.

For the IMN field, the gantry angle is chosen within the range of 20-30 degrees. For each angle, the aperture is defined, and a contour based on the 20% isodose line is computed. The volume intersecting the isodose contour with the lung is then determined to reflect the shape of the dose at depth. The angle associated with the largest volume is selected, indicating the entrance of the beam through minimal skin curvature (assuming with curvature the isodose will be skewed, which decreases its volume). In this scenario, the medial gantry angle is designated as the IMN gantry angle plus 10 degrees (minus 10 degrees for the left side), and the corresponding lateral beam angle is determined to match the medial angle posteriorly.

Energy selection & Dose Optimization:

For the SCLV field: the energy and respective field weights are determined through iterative optimization, aiming to achieve the clinical objective of covering 90% of the SCLV volume with 90% of the prescribed dose.

For the breast field: field-in-field optimization is utilized to mitigate hot spots and enhance the coverage of the 100% isodose line. Concerning the IMN field, the energy selection is contingent on the depth of the IMN contour. If the depth is variable across the targeted area, an additional beam set is generated with a different energy to accommodate these variations.

II- Algorithm Testing:

Ten post-lumpectomy breast patients were selected retrospectively from our institutional database. Seven patients had right sided treatment while 3 were left sided. All the patients except for one had one of their arms above their head for their treatment setup.

In total, 50 plans were generated, encompassing three approaches for each patient. Each approach consisted of a version employing a single isocenter methodology, as well as another version utilizing a two isocenter methodology. Therefore, every patient had a total of 5 plans, each presenting a distinct perspective on the chosen methodology.

All automated plans were evaluated for clinical acceptability based on the 95% of the targets' volumes receiving more than 95% of the prescribed dose, in addition to meeting OAR constraints presented in table 1. Hot spots were evaluated for the SCLV and breast fields independently. The preferred criteria for the maximum dose was set at <115% of the prescription. The two isocenter approach was verified qualitatively by matching the superior border of the tangents field to a contour created from the external body contour at the SCLV isocenter.

OAR	*Constraint*
Heart	At most 5 Gy average dose
Left Ventricle	At most 25% volume at 5 Gy
Ipsilateral Lung	At most 35% at 20 Gy dose
Esophagus	At most 38 Gy at 0% volume
Ipsilateral Brachial Plexus	At most 53.5 Gy at 0% volume
Contralateral Breast	At most 5Gy average dose

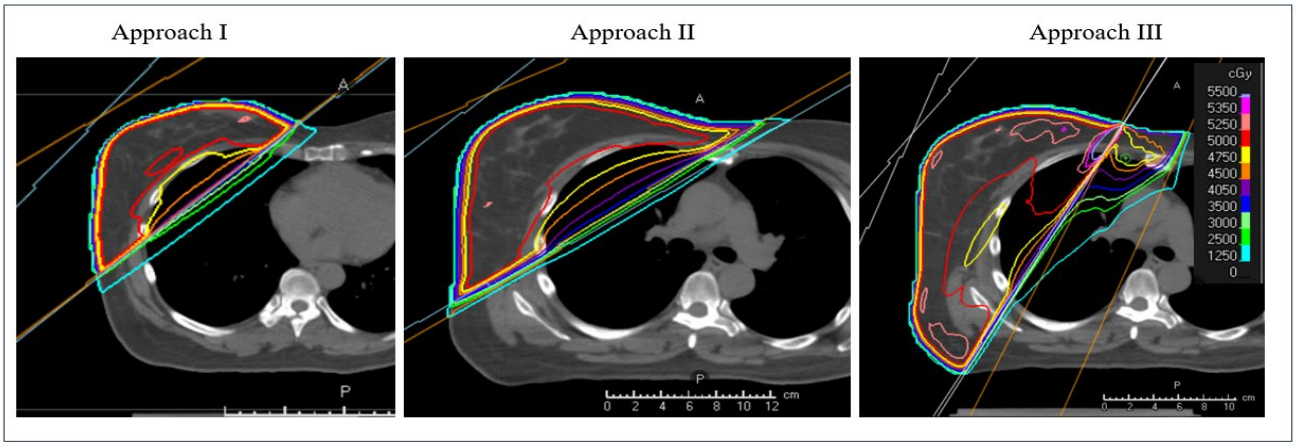


Figure 3. Example of axial slices with dose distribution of the 3 different approaches, approach I; regular tangent, approach III; wide tangents, approach III; tangents with electron IMN. Prescription was set to 50Gy for the breast and 45Gy for the lymph nodes

3 Results

For the regular tangents approach, all 10 automatically generated plans met the preferred objectives for targets coverage and OAR constraints as well as hotspot maximum. For the partial wide tangents approach, all targets and OARs have met their objectives for 50% of the plans. Three plans did not meet ipsilateral lung constraint (V20Gy= 40%, 43% and 42%). One plan with left sided treatment did not meet heart and left ventricle dose limit (mean heart dose= 5.8Gy and V5 left ventricle=28.3%). And one plan did not meet target coverage for AX1(D95%=38.4Gy) indicating the necessity for a posterior-anterior beam to encompass the deeper axillary nodes. Note, not meeting the lung or heart dose constraints may indicate that this approach is not appropriate for that patient.

For the third approach, with regular photon tangents field and a matched IMN electron field, 50% of the plans met all objectives and constraints. Four plans did not meet lung dose constraint, three of which were also flagged for the second approach. The plan that did not meet the heart dose limit in the second approach met the constraints following this method. The plan with low AX1 coverage was also flagged in this approach.

The two isocenters approach was successfully verified in all patients, correctly identifying the couch and collimator rotation angles to match the tangents and SCLV fields.
An example of all three approaches for a plan with a prescription of 50Gy to the breast and 45Gy to the lymph nodes is shown in figure 3.

4 Discussion

In this work, we have developed a portfolio of automated treatment plans for whole breast radiotherapy. This portfolio includes approaches for regular breast treatment without lymph nodes involvement and a comprehensive treatment option. Although in some cases, the automated plans did not meet the preferred objectives, this may be because of difficult patient anatomy. In all cases, it is likely that significant time savings could be realized through automation, followed by editing of the plans instead of starting anew. In addition, the clinical implementation of these automated models has the potential to mitigate the negative downfalls of radiation workforce shortage by improving workflows. Studies have shown that the integration of automated tools in clinical settings can effectively provide high-quality breast treatment plans, accompanied by potential time-saving benefits[8–10].

Effective screening and early diagnosis programs, combined with advanced radiation treatment planning tools, have the potential to contribute to shifting the standard of care for breast cancer in low-middle-income countries. This shift may involve a transition from mastectomy to lumpectomy care, emphasizing less invasive and more targeted treatment approaches. Our tools will hence have greater significance in helping resource constraint clinics where workforce expertise is lacking in transitioning their standard of care.

To our knowledge, this is the first work to automate electron treatment for IMN employing accurate energy selection and skin matching with the adjacent photon field. For clinics lacking machines with electron capabilities or having different clinical practice, a wide tangent approach was also developed for treatment of IMNs with photon beams. Along with breast and SCLV treatment, this therefore represents a first of its kind automation of all 3DCRT possible scenarios for comprehensive treatment of breast cancer.

Although no official physician review was performed for this study, informal feedback from MD Anderson's radiation oncologists sub-specialized in breast cancer, has guided the development of these tools. The next step in testing these tools will include a fully evaluation by radiation oncologists from multiple institutions.

5 Conclusion

The feasibility of an automated treatment planning model, encompassing different treatment paradigms and clinical preference for whole breast radiation therapy was proven. This work has the potential to improve the efficiency and outcome of whole breast radiotherapy globally.

References

1. EBCTCG (Early Breast Cancer Trialists' Collaborative Group). Effect of radiotherapy after mastectomy and axillary surgery on 10-year recurrence and 20-year breast cancer mortality: meta-analysis of individual patient data for 8135 women in 22 randomised trials. *Lancet Lond Engl*. 2014;383(9935):2127-2135. doi:10.1016/S0140-6736(14)60488-8
2. Yang W, Williams JH, Hogan PF, et al. Projected supply of and demand for oncologists and radiation oncologists through 2025: an aging, better-insured population will result in shortage. *J Oncol Pract*. 2014;10(1):39-45. doi:10.1200/JOP.2013.001319
3. Mathew A. Global Survey of Clinical Oncology Workforce. *J Glob Oncol*. 2018;(4):1-12. doi:10.1200/JGO.17.00188
4. Hanna TP, King WD, Thibodeau S, et al. Mortality due to cancer treatment delay: systematic review and meta-analysis. *BMJ*. 2020;371:m4087. doi:10.1136/bmj.m4087
5. Purdie TG, Dinniwell RE, Letourneau D, Hill C, Sharpe MB. Automated Planning of Tangential Breast Intensity-Modulated Radiotherapy Using Heuristic Optimization. *Int J Radiat Oncol*. 2011;81(2):575-583. doi:10.1016/j.ijrobp.2010.11.016
6. Sheng Y, Li T, Yoo S, et al. Automatic planning of whole breast radiation therapy using machine learning models. *Front Oncol*. 2019;9(AUG). doi:10.3389/fonc.2019.00750
7. Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat Methods*. 2021;18(2):203-211. doi:10.1038/s41592-020-01008-z
8. Dragojević I, Hoisak JDP, Mansy GJ, Rahn DA, Manger RP. Assessing the performance of an automated breast treatment planning software. *J Appl Clin Med Phys*. 2021;22(4):115-120. doi:10.1002/acm2.13228
9. Purdie TG, Dinniwell RE, Fyles A, Sharpe MB. Automation and intensity modulated radiation therapy for individualized high-quality tangent breast treatment plans. *Int J Radiat Oncol Biol Phys*. 2014;90(3):688-695. doi:10.1016/j.ijrobp.2014.06.056
10. Amit G, Purdie TG. Automated planning of breast radiotherapy using cone beam CT imaging. *Med Phys*. 2015;42(2):770-779. doi:10.1118/1.4905111

An open source multi-criteria optimization framework for radiotherapy

Tobias Becher[1,2,3], Amit Ben Antony Bennan[1,2], Remo Cristoforetti[1,2,3], and Niklas Wahl[1,2]

[1]Department of Medical Physics in Radiation Oncology, German Cancer Research Center – DKFZ, Heidelberg, Germany
[2]Heidelberg Institute for Radiation Oncology (HIRO) and National Center for Radiation Research in Oncology (NCRO), Heidelberg, Germany
[3]Department of Physics and Astronomy, Heidelberg University, Heidelberg, Germany

Abstract We present an implementation of a Pareto plan navigation framework in the open source treatment planning toolkit `matRad`. It expands on the existing weighted sum approach and allows for calculation and navigation of Pareto surfaces. Furthermore we introduce a lexicographic approach as an alternative to the weighted sum method.

1 Introduction

A central concern of the treatment planning problem in radiotherapy is the handling of different trade-offs like target coverage against organ at risk (OAR) sparing. Since the different objectives often contradict each other there is no single optimal plan, but rather an entire set of *Pareto optimal* plans that cannot improve on one objective without worsening at least another.

Methods to solve this multi-criteria optimization (MCO) problem evolved from an iterative tuning of weights in a weighted sum objective function to automated approaches using Pareto front approximation or lexicographic optimization. Pareto front approximation techniques automatically compute a set of Pareto optimal plans and allow navigation of the resulting Pareto front in an interactive user interface. Lexicographic approaches, on the other hand, require definition of a *wish list*, e. g., based on a clinical protocol prioritizing treatment goals, to automatically reach an acceptable plan. These approaches are now commonly implemented and used in commercial treatment planning systems. The open-source community is, however, lacking comprehensive implementation of such methods for research and education.

Within the open-source radiotherapy treatment planning toolkit `matRad`, which up to its latest release relies on the weighted sum approach for non-linear treatment plan optimization, we developed a MCO module to allow for Pareto front approximation and navigation as well as lexicographic optimization. The implementations are based on well-established methods published in literature.

For Pareto front approximation and navigation, we employ a so-called *sandwich* algorithm [1] that iteratively chooses weights automatically to generate a plan database approximating the Pareto front. Afterwards, a navigation algorithm finds convex combinations of the pre-calculated plans to allow near continuous and interactive exploration of the solutions.

For lexicographic optimization, `matRad` was extended with a priority and goal system to enable definition of a wish list. The $2p\varepsilon c$-method [2] was chosen to consecutively optimize individual objectives constrained by previous solutions and goals. In the end, this method results in a plan optimally fulfilling all the defined goals while prioritizing minimization of the objective functions according to their priority in the wish list.

In this work, we describe the implementation of said MCO framework in `matRad` and demonstrate both methods by application to a prostate case.[1]

2 Materials and Methods

2.1 Software framework

The basis for our MCO framework is `matRad` (version 2.10.1 *Blaise*) [3, 4]. It is written in `MATLAB` and allows for IMRT and IMPT dose calculation and plan optimization using a weighted sum approach with user chosen penalties. Photons, protons and carbon ions are available for planning including biological optimization for the latter. Optimization relies on `IPOPT` [5] or Matlab's `fmincon` using an L-BFGS approximation for Hessian matrices.

To implement Pareto front approximation, linear programs and convex hull calculations are required. These are solved using the `Matlab` functions *linprog* and *convhulln*, respectively.

2.2 The multi-criteria optimization problem in radiotherapy

The general multi-criteria RT treatment planning problem can be expressed through a set of functions $\mathbf{f}$ and hard constraints $\mathbf{c}$:

$$\begin{aligned} \underset{\mathbf{w}}{\text{minimize}} \quad & \mathbf{f}(\mathbf{d}) = [f_1(\mathbf{d}), f_2(\mathbf{d}), ..., f_n(\mathbf{d})] \\ \text{subject to} \quad & c_k^l \leq c_k(\mathbf{d}) \leq c_k^u \qquad \forall k, \\ & d_i = \sum_j D_{ij} w_j \qquad \forall i, \\ & w_i \geq 0 \qquad \forall i \end{aligned} \tag{1}$$

$\mathbf{d}$ is a discretized vector of dose values with elements d_i related to the beamlet weights $\mathbf{w}$ via the precomputed dose influence matrix $\mathbf{D}$ such that $\mathbf{d} = \mathbf{D}\mathbf{w}$. One approach minimizes the weighted sum, $\sum_i p_i f_i$, where p_i are weights/penalties applied to each objective.Varying the penalties allows to obtain different plan options.

[1]The implementation is available in the `matRad` repository on `Github` (Commit ID at submission: f19635f).

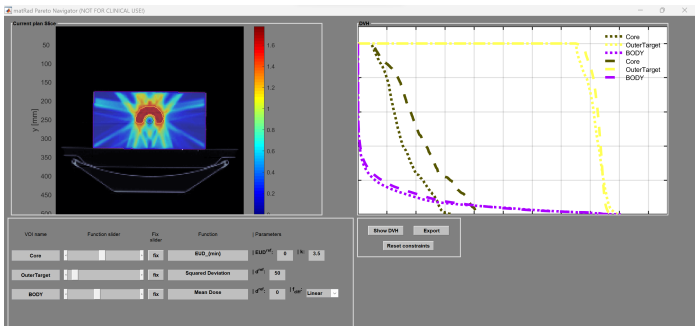


Figure 1: GUI for interactive plan navigation. Objective values can be adjusted by moving the sliders and limited through upper bounds. Additional buttons allow to show the current plan's DVH and to export it to the general `matRad` GUI.

2.3 Pareto front approximation

For Pareto front approximation, we rely on so-called *sandwich* algorithms that approximate the Pareto surface with polyhedral inner and outer approximations. Given an existing set of plans with corresponding objective function values $\mathbf{y}$ the inner approximation is formed by the corresponding convex hull while the outer approximation is bounded by the intersection of supporting hyperplanes. With the assistance of the maximum distance of inner and outer approximation a new weight vector is chosen to improve on the current approximation in the most improvable region.

The algorithm for our implementation was introduced by Rennen, Dam, and Hertog [1]. By optimizing each objective individually we calculate the anchor points of the Pareto front. The anchor point solutions are used to scale the objectives f_i to a range of 0 to 1, to better account for different dimensions of objective function values, using their highest values u_i and lowest value l_i. The scaled objectives f' are then given by $f'_i = \frac{f_i - l_i}{u_i - l_i}$. Afterwards the algorithm iterates over the bounding facets of the inner approximations and calculates new penalty vectors based on the normal vector of the facet with largest distance.

For optimization of individual weight computations found by the sandwich algorithm, we use `matRad`'s existing weighted sum optimization framework, allowing us to use any optimizer interfaced by `matRad`. The sandwich algorithm continues until the maximal distance between inner and outer approximation falls below a predetermined threshold or a certain number of iterations is passed.

Finally, all solutions are stored in form of the resulting beamlet intensity vectors for direct dose calculation or plan navigation.

2.4 Plan navigation on the Pareto front

For visualization of the different plans after database calculation, a dedicated widget is available for `matRad`'s graphical user interface (GUI). The widget is shown in Figure 1 with an example Pareto front approximation loaded.

Starting from an initial plan with objective function values $\hat{\mathbf{y}}$, a new plan can be investigated by moving across the Pareto surface using interactive sliders. Upon moving a slider, the corresponding objective function value is fixed and a new plan is found as a convex combination of pre-calculated plans using the algorithm introduced by Monz, Küfer, Bortfeld, et al. [6]. The core idea of the algorithm is to start from the point $\mathbf{y}^R$ obtained through the slider movement and then search along the line $\mathbf{y}^R + t\mathbf{1}$ for an intersection $\mathbf{y}'$ with the approximation of the Pareto surface, expressed as $\mathbf{y}' = \sum_k v_k \mathbf{y}^k$. $\mathbf{y}^k$ are the objective function values of the pre-calculated plans and v_k the corresponding convex coefficients, that are then used to calculate new weights $\mathbf{w}' = \sum_k v_k \mathbf{w}^k$ from the pre-calculated weights $\mathbf{w}^k$. While the new plans are not necessarily Pareto optimal,[2] they are still clinically relevant if the Pareto surface is sufficiently approximated.

To limit navigation to regions of interest, upper bounds may be added on specific objective functions. The range of the sliders is then adjusted to mimic the upper bound and since the change also limits the options for other objectives the respective slider ranges are changed as well.

2.5 Lexicographic optimization

Lexicographic optimization relies on a prioritized wish list formulating goals for each objective f_i. The chosen method for our implementation – the $2p\varepsilon c$-method [2, 7] – consists of two fully automated plan optimization phases of iterating to the wish list.

The first phase optimizes objectives individually one after the other (under the general constraints of Equation 1). In each step, moving to optimization of lower prioritized objectives adds additional constraints $f_j \leq \varepsilon_j$ to still respect higher priority objectives. ε_j is set to the corresponding aspired value g_j if it can be met and else to the achieved value f_j^* multiplied with a slack variable c (in our implementation chosen as $c = 1.03$).

The second phase then turns objectives that could not meet their goal value into hard constraints and they are not optimized again. For the remaining objectives it applies the ε-constraint method which can be formulated as

$$\begin{aligned} \underset{\mathbf{w}}{\text{minimize}} \quad & f_i(\mathbf{d}) \\ \text{subject to} \quad & c_k^l \leq c_k(\mathbf{d}) \leq c_k^u && \forall k, \\ & f_j(\mathbf{d}) \leq \varepsilon_j && \forall j \in \{1,..,n\} \backslash i, \\ & d_i = \sum_j D_{ij} w_j && \forall i, \\ & w_i \geq 0 && \forall i \end{aligned} \tag{2}$$

Initially the ε values are set as achieved in the first phase and after optimization are changed to cf_j^* for all following objectives.

This approach has the advantage that objectives that easily meet their goal value do not dominate the optimization in the second step, while ensuring that objectives which are not able to meet their goal value are still respected.

[2] for convex objectives it holds that $\mathbf{f}\left(\mathbf{D}\sum_k v_k \mathbf{w}^k\right) \leq \sum_k v_k \mathbf{y}^k$

	Objectives		Constraints	
Volume	Name	Parameters	Min[Gy]	Max[Gy]
Target	$f_{sq\ deviation}$	$\bar{d} = 50$	45	57
Core	f_{EUD}	$a = 3.5$	0	40
Body	f_{mean}	-	0	45

Table 1: Objectives and constraints used for the TG119 phantom.

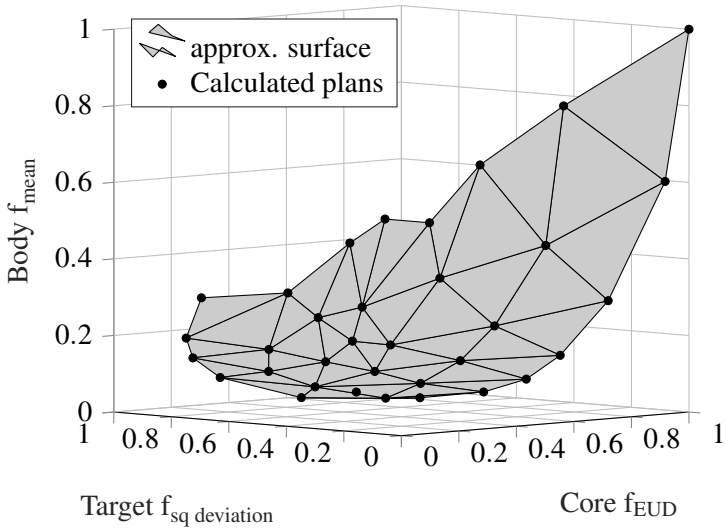


Figure 2: Example Pareto surface for the TG119 phantom. The (normalized) objective function values define the three axes. The plans obtained through optimizing the weight combinations found by the sandwich algorithm are shown as dots connected by the approximated Pareto surface.

3 Results

3.1 Pareto front approximation

The implementation of the Pareto front approximation algorithm was verified with the non-radiotherapy test case 1 from the original publication [1] in `MATLAB`.
It was then applied to two radiotherapy scenarios using the CORT dataset available within `matRad`. The first example uses the TG119 phantom shown in Figure 1. The three objectives set (see Table 1) allow to plot the Pareto surface approximation (see Figure 2).
The second example uses a prostate patient (shown in Table 2). In total, 105 plans were calculated for the database. In Figure 3, we show four plans: (1) anchor point plan focusing on the 68 Gy volume, (2) anchor point plan focusing on rectum, (3) a balanced plan from the precalculated database and (4) plan found with help of the navigation interface.

	Objectives		Constraints	
Volume	Name	Parameters	Min[Gy]	Max[Gy]
PTV 68	$f_{sq\ deviation}$	$\bar{d} = 68$	61.2	78.2
PTV 56	$f_{sq\ deviation}$	$\bar{d} = 56$	50.4	64.4
Bladder	f_{EUD}	$a = 3$	0	58
Rectum	f_{EUD}	$a = 8$	0	58
Body	f_{mean}	-	0	58

Table 2: Objectives and constraints used in the sandwich algorithm implementation for the prostate case.

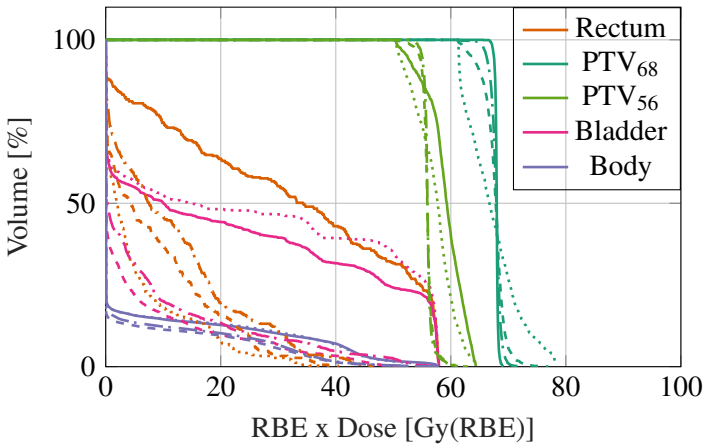


Figure 3: DVH for the four prostate plans. Anchor point plans (1) and (2) are shown as solid and dotted lines, respectively. The balanced database plans (3) and (4) are plotted dashed and dash-dotted.

3.2 Lexicographic optimization

To test the $2p\varepsilon c$-method a wish list on the prostate phantom presented in Table 3 was optimized.

Constraints				
Number	Volume	Name	Min[Gy]	Max[Gy]
1	Rectum	Min/Max Dose	0	58
2	Bladder	Min/Max Dose	0	58
3	Body	Min/Max Dose	0	58
Objectives				
Priority	Volume	Name	Param.	Goal
1	PTV 68	$f_{sq\ deviation}$	$\bar{d} = 68$	2
2	PTV 56	$f_{sq\ deviation}$	$\bar{d} = 56$	2
3	Rectum	EUD	$a = 8$	30
4	Bladder	EUD	$a = 3$	30
5	Body	Mean dose	-	10

Table 3: wish list for the prostate patient. Constraints for OAR are set as in the Pareto front approximation, while target constraints are omitted. The PTVs have highest priority and are are assigned a squared deviation objective each.

Table 4 shows achievable objective function values and chosen constraints in both phases of the $2p\varepsilon c$ approach. All objectives except for Rectum meet their aspired goal value in the first phase and are therefore optimized again. A DVH corresponding to the final plan can be found in Figure 4. The target coverage was achieved and while the Rectum goal is missed, the resulting plan was evaluated to be acceptable.

First step				
Priority	Volume	Goal	Achievable	New constraint
1	PTV 68	2	$5.9 \cdot 10^{-2}$	2
2	PTV 56	2	$2.6 \cdot 10^{-6}$	2
3	Rectum	20	22.6	23.3
4	Bladder	30	18.9	20
5	Body	10	5.1	10
Second step				
1	PTV 68	2	1.67	1.72
2	PTV 56	2	0.70	0.72
4	Bladder	30	23.8	24.5
5	Body	10	4.3	-

Table 4: Result table of the prostate patient. The goal on the rectum cannot be met in the first iteration and is therefore skipped in the second iteration.

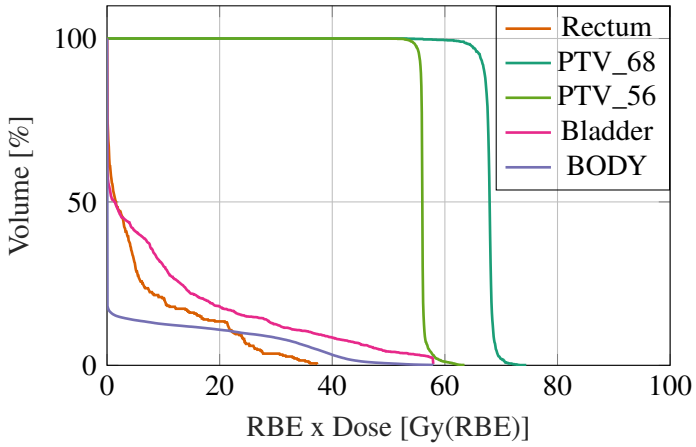


Figure 4: DVH for the prostate plan calculated with the $2p\varepsilon c$-method.

4 Discussion

Two established MCO treatment planning approaches were successfully implemented in the open-source treatment planning system `matRad`.

4.1 Pareto front approximation and navigation

The implemented sandwich algorithm was benchmarked against published results and the Pareto surface could be adequately approximated for two RT examples.
While navigation of the treatment plan is possible nearly in realtime, once the plan database is calculated using the sandwich algorithm, the computation time is limited by the constraint optimization method in `matRad`. Since `matRad` favors flexibility of computational performance optimization, the performance of the non-linear optimization approach is suboptimal for Pareto surface approximation. However, computational performance can be optimized in the future by users themselves or public community contributions.

Future improvements to the algorithm itself might include automatic clustering of correlated objectives to reduce the complexity of the calculation for high dimensional problems as well as the implementation of parallelizable sandwich algorithms.

4.2 Lexicographic optimization

Implementation of the $2p\varepsilon c$-method was demonstrated with a prostate plan and was shown to work as expected. It allows for a more automated and clinically oriented steering of the treatment plan than the weighted sum approach, however design of a suitable wish list depends on the planner's experience. Similar computational limits as in the Pareto front approximation approach apply, as the approach is limited in performance by `matRad's` underlying optimization algorithm.

5 Conclusion

We present our MCO framework in the open source treatment planning software `matRad` which includes: (1) A sandwich algorithm implementation to calculate Pareto optimal plans and a navigation interface to allow their exploration afterwards. (2) also introduced a lexicographic method for automated treatment plan generation. This package facilitates future research into advanced treatment plan optimization strategies.

References

[1] G. Rennen, E. van Dam, and D. den Hertog. *Enhancement of Sandwich Algorithms for Approximating Higher Dimensional Convex Pareto Sets*. Rochester, NY, June 30, 2009. DOI: 10.2139/ssrn.1427721.

[2] S. Breedveld, P. R. M. Storchi, and B. J. M. Heijmen. "The equivalence of multi-criteria methods for radiotherapy plan optimization". *Physics in Medicine & Biology* 54.23 (Nov. 2009), p. 7199. DOI: 10.1088/0031-9155/54/23/011.

[3] H.-P. Wieser, E. Cisternas, N. Wahl, et al. "Development of the open-source dose calculation and optimization toolkit matRad". *Medical Physics* 44.6 (2017). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/mp.12251, pp. 2556–2568. DOI: 10.1002/mp.12251.

[4] B. Ackermann, M. Bangert, A. B. A. Bennan, et al. *e0404/matRad: Blaise v2.10.1*. Version 2.10.1. Nov. 20, 2020. DOI: 10.5281/zenodo.7107719.

[5] A. Wächter and L. T. Biegler. "On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming". *Mathematical Programming* 106.1 (Mar. 1, 2006), pp. 25–57. DOI: 10.1007/s10107-004-0559-y.

[6] M. Monz, K. H. Küfer, T. R. Bortfeld, et al. "Pareto navigation—algorithmic foundation of interactive multi-criteria IMRT planning". *Physics in Medicine and Biology* 53.4 (Feb. 21, 2008), pp. 985–998. DOI: 10.1088/0031-9155/53/4/011.

[7] S. Breedveld, P. R. M. Storchi, M. Keijzer, et al. "A novel approach to multi-criteria inverse planning for IMRT". *Physics in Medicine & Biology* 52.20 (Oct. 2007), p. 6339. DOI: 10.1088/0031-9155/52/20/016.

Zero click treatment preparation: from CT to plan

Geert Wortel[1], Rita Simões[1], Johan Trinks[1], Peter Remeijer[1], Eugène Damen[1], Mark Gooding[2], Tomas Janssen[1]

[1]Department of Radiation Oncology, The Netherlands Cancer Institute, Amsterdam, The Netherlands
[2]Inpictura Ltd, Oxford, United Kingdom

Abstract Automation of segmentation and treatment planning is an important development in radiotherapy. While this automation can save time, the auxiliary workflow steps are still time consuming, and typically involve multiple hand-overs between personnel, hampering an efficient process. To create a fully automated workflow, segmentation and planning need to be embedded within an automation framework. In this work we describe a fully automated workflow of the complete treatment preparation chain from CT to RTPlan export. All key components of this workflow are currently in clinical use. Starting from the CT, auto-segmentation is performed by Mirada WorkflowBox. The CT and resulting structure set are sent to Pinnacle, which in combination with our automation framework FAST creates and exports RTPlan and Dose. We discuss the implications of full automation of all auxiliary steps on the architecture of the workflow.

1 Introduction

Automation of treatment preparation for radiotherapy is an ongoing and crucial development. Main steps in the process include image segmentation and treatment planning. A wide range of solutions, both commercially available and home-built, is used to perform these tasks[1-4]. However, for a full automation of the workflow, data flow, protocol selection and auxiliary steps also need to be performed automatically[5].

In clinical practice automatically generated structures and plans often still require manual verification. The auxiliary workflow steps (which include data transfer, data import, creation of help-structures, choice of isocenter, selection of CT reference point beam setup, choice of prescription, selecting isodose color scheme, etc.) are therefore often not automated as they can simply be performed by the user that is already present for the verification. This, however, does limit the potential time saving and hampers the development of full automation of the preparation workflow.

Automated treatment preparation executes a substantial part of the radiotherapy treatment chain. It saves time, boosts staff productivity and can result in consistent and high plan quality. The only requirements are a planning scan and a clinical protocol as input and an evaluation of the final output by a clinician. Due to its speed and efficiency, automated treatment preparation can also be highly beneficial for online adaptive radiotherapy[6] or the development of a one-stop-shop.

In this work we describe a fully automated workflow of the complete treatment preparation chain. This workflow is based on our clinically used automated segmentation and planning methods and designed such that it can be integrated into our clinical practice. The focus of this work is on building the automation framework. Further improvements in segmentation and plan quality can be achieved by implementing future state-of-the-art segmentation and planning methods. The workflow was submitted to the AUTO-RTP challenge, presented at the 2023 ESTRO Physics Workshop on automation.

This challenge asked participants to generate treatment plans as-automated-as-possible, based only on the planning CT and a description of the treatment protocol[7].

2 Materials and Methods

In the proposed fully automated workflow, the segmentation is done in Mirada WorkflowBox 2.6.2. (Mirada Medical, Oxford, UK) and plan optimization is done in Pinnacle 16.2 (Philips Radiation Oncology, Madison, WI), using our automation framework FAST[8].

2.1 Challenge data, protocol and scoring

Planning CT scans of ten prostate cancer patients were made available via the AUTO-RTP challenge website. The cases were a mix of prostate only, prostate with nodes, and prostate bed with nodes cancer patients. Organs at risk (bowel bag, bladder, rectum, and femoral heads) and CTVs (prostate(bed), seminal vesicles and lymph node regions) were all to be auto-segmented according to published guidelines. Dose prescription, DVH constraints and PTV expansions were also provided, broadly based on the PIVOTAL trial[9].

After submission to the AUTO-RTP challenge, submissions were scored on a single plan quality measure (range 0 – 100%). This measure was calculated based on DVH parameters of the submitted RTdose when evaluated using consensus contours from three experts. The consensus contour was not available for participants. For details, see the AUTO-RTP grand challenge website[7].

2.2 Image acquisition, DICOM routing and segmentation

The workflow starts with a planning CT that is sent to the general DICOM server. The DICOM series description contains the clinical protocol name. The DICOM server forwards all planning scans to the Mirada WorkflowBox (WFB), a server that performs automated processing. The Mirada WFB contains routing rules that configure how to process incoming DICOM objects that meet certain criteria. In this workflow, the WFB responds to the clinical protocol name, which is read from the DICOM series description of the planning CT.

As the protocol name was not available in the DICOM objects of the ESTRO challenge, a data-preprocessing step

was added where we entered the protocol name in the series description of the CT.

The Mirada WFB is configured to segment the images using Mirada DLCExpert[10]. As two separate DLC modules had to be run to obtain all required structures, a step was added to the Mirada WFB to combine the two resulting structure sets (RTSS). Next, the CTV-PTV expansions are performed as configured in the WFB, and the CT and complete RTSS are automatically sent to Pinnacle.

A FAST listener was built that triggers on DICOM files arriving in the Pinnacle DICOM pool. Based on the plan protocol in the series description, automatic planning is started using the corresponding FAST protocol. An existing clinical protocol was modified to adjust to the treatment prescriptions as described in the challenge. After performing the planning FAST automatically exports the RTPlan and RTDose DICOM objects. Automated protocol-based DVH analysis and plan evaluation documentation is generated and sent as pdf to the record and verify system for verification. More details on FAST planning will be given in the next section.

The entire workflow is fully automatic. It responds to a CT scan arriving on the DICOM server and concludes with an export of RTSS, RTPlan and RTDose and evaluation pdf. A summary of the automated workflow, compared to the current clinical workflow, is shown in Figure 1.

2.2 FAST Planning

The treatment plans are generated using Pinnacle, including the Auto-Planning module, and our in-house developed automation framework FAST.

FAST was designed to automate the entire planning process. In a typical clinical workflow, FAST is started manually from Mirada RTx and it performs planning headless in the background. If needed the planning process can be monitored on a dashboard. Once planning is done, the RTT and physician can evaluate the treatment plan in Pinnacle. Alternatively, all FAST functionality is also available from a user menu within the TPS. FAST was clinically released in February 2015. Since then, it has been clinically used to create over 20.000 plans.

The core of FAST consists of 1) planning protocols, 2) modules that interact with Pinnacle and 3) a workflow engine.

The protocols are xml-based and contain all information that is required to generate a plan. This includes the definition of auxiliary structures, beam configurations, dose calculation settings, planning optimization objectives and planning constraints for evaluation. An important advantage of the protocols is that they separate the configuration from the planning functionality.

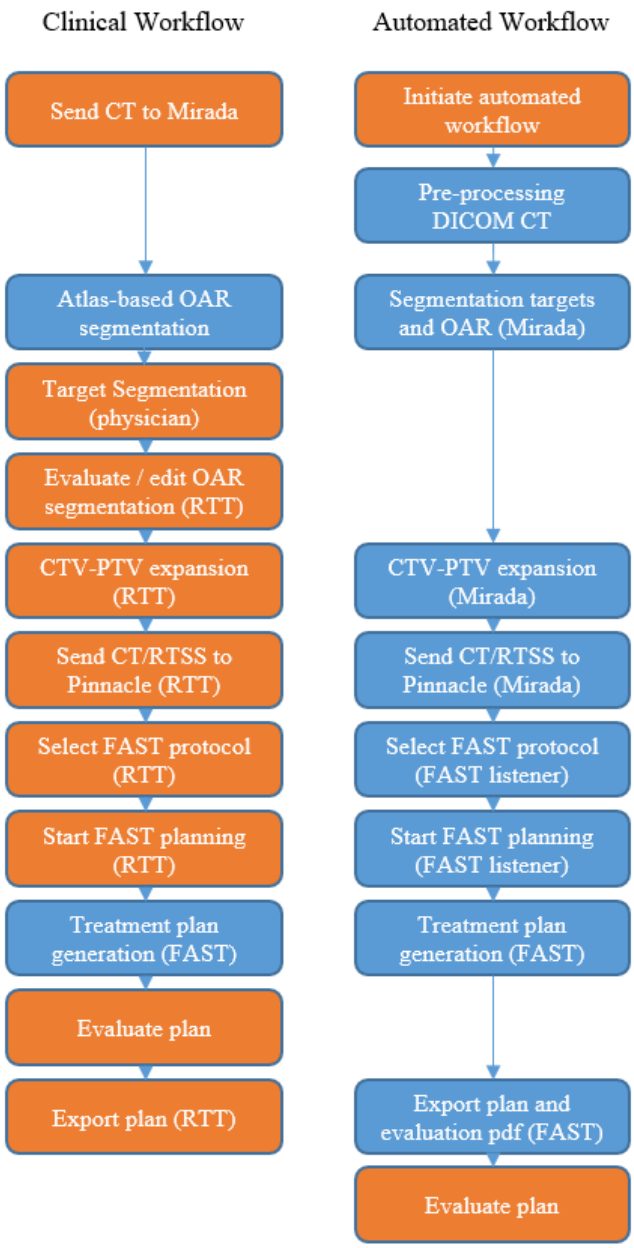


Figure 1: A comparison of the current clinical (left) and fully automated workflow (right). The orange blocks are manual, the blue blocks are automated.

Modules were built for each step in plan generation. A module executes its tasks based on the assigned protocol and current status in the TPS. Modules communicate with Pinnacle using an in-house developed Pinnacle API that enables direct use of Pinnacle scripting from Python.

The workflow engine consecutively starts all the modules. For the standard planning workflow, the order of the modules is determined by the natural order of the steps in the planning process.

The modules were designed to be robust and multi-purpose. In case something has to be changed in an early step of the planning, all modules following that step can simply be re-executed. FAST continuously monitors the

status of the TPS and therefore can trigger on unexpected events or a module which is not executed properly.
The workflows are started by sending an URL to the FAST webservice. This URL crucially contains the DICOM UIDs of the CT and RTSS to be used for the plan, and the protocol name. FAST will create a patient record in Pinnacle, add the DICOM files and the workflow engine will start the planning.
Once the plan is done, a DVH report is generated and saved as pdf. A DICOM export module is available to export any DICOM objects. In the clinical workflow, this is not used as the planners always evaluate the plan before export.
To generate optimal dose distributions, the plan optimization is done using Pinnacle Auto-Planning. For each treatment protocol, an automation template was developed that contains the required target dose as well as the priorities in OAR sparing and dose conformity[11]. The Pinnacle Auto-Planning template was encoded in the FAST protocol and Auto-Planning is started by a FAST module.
In the clinic, the acceptance rate of the automatically generated prostate plans is 75%[9].

3 Results

The fully automatic workflow ran successfully for all 10 patients in the challenge. The total runtime was approximately 30 minutes (5 minutes for segmentation, 25 minutes for planning) per prostate cases and took about 60 minutes for cases involving nodes (10 for segmentation, 50 for planning).
The primary aim of the challenge was to gauge the state of automation in the field. Our submission was fully automated (‘zero click’) and from all participants, the closest to clinical practice as all key components are currently in clinical use.
In the challenge we achieved a 70% score based on DVH evaluation. This resulted in 7th place out of 13 participants.
In one out the ten plans all clinical constraints were met. In the remaining nine, most constraints were met except for the D95% target coverage. The reason was that the autosegmentation missed a small part of the target compared to the consensus contour, immediately resulting in insufficient target coverage when evaluated on the consensus contour. The objectives were met on our own auto-segmentations.
For prostate cases we scored average plan quality measures of 76%, for prostate plus nodes 71% and for prostate bed plus nodes 58%. The low score for prostate bed cases is mostly due by the fact that we did not have an auto segmentation method available for the prostate bed and used our available prostate model instead.

4 Discussion

In this work we showed that a fully automatic radiotherapy planning workflow is feasible using clinically available auto-segmentation and auto-planning methods.
While a lot of automation literature focusses on the specific task of segmentation and treatment plan optimization, a fully automated workflow is considerably more than that, since all auxilliary steps taken by a user in clinical practice also need to be automated.
Automation of the segmentation part of plan preparation was achieved by adding extra rules for DICOM routing and Boolean operations on auto-segmented structures in WFB. The dose optimization part of planning was performed by our existing automation framework FAST. WFB and FAST were connected by a newly developed FAST listener. Possible alternatives include either starting FAST explicitly from WFB or WFB from FAST, or adding an overarching workflow manager that first starts the segmentation and then the planning.
Even though plans could be prepared fully automatically, they were not always of acceptable quality when evaluated on the challenge golden standard segmentation. This was mostly due to deviations in the target segmentation, which was strongly penalized in the scoring. However, the autosegmentaion models used were not optimized on this guideline and inter-observer variation in segmentation is always present. Still, this shows that in a fully automated workflow, high quality target segmentation is a crucial bottleneck. As long as this is not near flawless, it is sensible to at least evaluate target segmentations before starting an automated plan preparation. With improved segmentation methods, however, it becomes feasible to automate the entire planning and check the segmentation only once the plan is done.
A major factor in the implementation of automated treatment preparation is trust. The acceptance rate of the resulting plans should be high enough to make it perceived as useful to skip intermediate evaluations. In case an aspect of a plan has to be improved, all steps of the workflow depending on the improved factor also have to be re-executed. This requires extra time as well as flexibility of the involved software. This factor becomes of less importance as the execution time of segmentation and planning methods decrease. Whereas our current implementation is still relatively slow, speed increase is expect e.g. due to GPU-based dose computation[12]. The applied software should be designed such that plans can easily be rerun and the overview of data and plans remains clear.
In this work, we showed automation from CT to plan. However, after RTplan export additional steps are often required. These could include import of the rtplan in the Record and Verify system and pre-treatment planQA. Automation of these steps might also be (partially)

possible, but were outside the scope of the challenge and therefore also not taken into account in this work.

5 Conclusion

Full automation of the treatment preparation workflow is possible and requires embedding auto segmentation and auto treatment planning into an automation framework. Crucial aspects regarding this workflow are that all required input and information should be present when the preparation starts, all data transfer and starting of applications or modules should be automated, and a framework is required in which the next component in the workflow starts or is started at the right moment. Arguably, the most crucial aspect of creating the appropriate trust in the automated workflow is the auto-segmentation of the targets to be near flawless.

References

[1] A. Barragán-Montero, U. Javaid, G. Valdés et al. "Artificial intelligence and machine learning for medical imaging: A technology review". Phys. Med. 83 (2021), pp 242-256. DOI: 10.1016/j.ejmp.2021.04.016.

[2] C. Cardenas, J. Yang, B. Anderson et al. "Advances in Auto-Segmentation". Semin Radiat Oncol 29.3 (2019), pp 185-197. DOI: 10.1016/j.semradonc.2019.02.001.

[3] K. Nealon, P. Balter, R. Douglas et al. "Using Failure Mode and Effects Analysis to Evaluate Risk in the Clinical Adoption of Automated Contouring and Treatment Planning Tools". Practical Radiation Oncology 12 (2022), pp 344-353. DOI: 10.1016/j.prro.2022.01.003.

[4] M. Hussein, B. Heijmen, D. Verellen et al. "Automation in intensity modulated radiotherapy treatment planning—a review of recent innovations". Br J Radiol 19.1092 (2019), pp. 20180270. DOI: 10.1259/bjr.20180270.

[5] P. Meyer, M. Biston, C. Khamphan et al. "Automation in radiotherapy treatment planning: Examples of use in clinical practice and future trends for a complete automated workflow". Cancer rad. 25.6-7 (2021), pp. 1278-3218. DOI: 10.1016/j.canrad.2021.06.006

[6] C. Glide-Hurst, P. Lee, A. Yock et al. "Adaptive Radiation Therapy (ART) Strategies and Technical Considerations: A State of the ART Review From NRG Oncology". Int. J. Radiat. Oncol.* Biol.* Phys. 109.4 (2021), pp. 1054-1075. DOI: 10.1016/j.ijrobp.2020.10.021.

[7] ESTRO working group on AI "Fully Automated Radiotherapy Treatment Planning Challenge". Available at: https://auto-rtp.grand-challenge.org/ (Accessed 18 December 2023).

[8] E. Damen, "Automated treatment planning at NKI: benefits and challenges". ESTRO 37 (2018), DOI: 10.1016/S0167-8140(18)30955-1

[9] Dearnaley D, Griffin CL, Lewis R, et al. Toxicity and Patient-Reported Outcomes of a Phase 2 Randomized Trial of Prostate and Pelvic Lymph Node Versus Prostate only Radiotherapy in Advanced Localised Prostate Cancer (PIVOTAL). Int J Radiat Oncol Biol Phys 2019; 103: 605–17.

[10] T. Lustberg, J. van Soest, M. Gooding, D. Peressutti, P. Aljabar, J. van der Stoep, W. van Elmpt, A. Dekker, Clinical evaluation of atlas and deep learning based automatic contouring for lung cancer, Radiotherapy and Oncology, Volume 126, Issue 2, 2018, Pages 312-317, https://doi.org/10.1016/j.radonc.2017.11.012.

[11] G. Wortel, D. Eekhout, E. Lamers et al. "Characterization of automatic treatment planning approaches in radiotherapy". Phys Imaging Radiat Oncol, 19 (2021), pp 60-65. DOI: 10.1016/j.phro.2021.07.003.

[12] F. Fracchiolla, E. Engwall, M. Janson et al. "Clinical validation of a GPU-based Monte Carlo dose engine of a commercial treatment planning system for pencil beam scanning proton therapy". Phys. Med. 88 (2021), pp 226-234. DOI: 10.1016/j.ejmp.2021.07.012.

Development of Bolus Optimization System for Electron Conformal Therapy

Sang-Won Kang[1], Woong Cho[2], Dongsan Kang[3], Moo-Jae Han[1], and Jin-Beom Chung[1]

[1]Department of Radiation Oncology, Seoul National University Bundang Hospital, Seongnam, Korea
[2]Department of Radiation Oncology, Seoul National University Borame Medical Center, Seoul, Korea
[3]Department of Biomedical Engineering and of Biomedicine & Health Sciences, The Catholic University of Korea, Seoul, South Korea

Abstract Aim of this study is to develop a bolus optimization system for electron conformal therapy (ECT). To develop the optimization system, the electron dose calculation engine and an optimization algorithm should be built. Thus, Hogstrom's dose calulation algorithm and backtracking line search were implemented and applied in the optimization system. Accuracy of the developed dose calculation algorithm was verified by a 94.97% passing-rate in a gamma evaluation with a commercially available program. With evaluating the dose of the target-line by applying the optimzed bolus, the average dose and root mean square error were 95.23% and 15.17, respectively. In this study, we evaluated the developed bolus optimization system for ECT. Although some limitations are presented, it has shown feasibility of the opimization system to apply in clinical practice. In future study, we will imporve the performance of the system and incorporate 3D printing technology into the optimization system.

1 Introduction

Electron radiotherapy is commonly used for the skin cancer or malignant tumors within 7.5 cm of the patient's surface.[1-3] However, when the patient's surface is not flat or the tumor has a complex morphology, it is not an effective treatment due to it's physical limitations. To overcome these limitations, the concept of electron conformal therapy (ECT) using a bolus compensator [4-7] or an electron-collimator [8-11] has been proposed, but practical clinical utilization has not been achieved due to the failure to control the electron scattering effect and the cost of manufacturing.

Recently, the clinical application of ECT with bolus has been facilitated by the availability of 3D printing technology. However, there is no commercially optimization technique for bolus shape as the target, so this study aims to develop the bolus optimization system for ECT.

2 Materials and Methods

At first, to develop a bolus optimization system, the electron dose calculation algorithm refered by Hogstrom's was builted in this study. [12] Secondly, an optimization algorithm was also built. In general, multi-criteria optimization method is usually used for optimization in radiotherapy, but the bolus optimization system applied the backtracking line search method to the prescribeed dose-line (target-line) for dose optimization.

The electron dose calcualtion algorithm assumes that overall calculation was performed in the infinitesimal area and the electron beam are parallel beams. Thus, the percentage depth dose (PDD) of a pencil beam falling on a plane is the same shape in the calculation. In addition, when an electron reacts in a medium and transfers energy to a distant location, the direction of transfer is assumed to be perpendicular to the incident direction.

S(x', y', z') is the relative amount of electron fluence when an electron of initial mono-energy arrives at position (x', y', z'). An electron arriving at (x', y', z') will undergo multiple scattering processes and transfer energy to point (x, y, z) as a kernel. Define the rate of this transfer as K(x' →x, y'→y, z'→z). To summarize the above concept, the dose arriving at (x, y, z) is expressed as the sum of energy contributions from electrons falling everywhere.

$$D(x,y,z) = \iint S(x',y') \cdot K(x' \to x,\ \ y' \to y, z' \to z)\ \ dx'dy'$$

S(x', y') is a kind of off-axis electron fluence ratio, and it is preferentially set to 1 for simplicity of calculation in this study. To solve for K(x' →x, y'→y, z'→z), it is assumed that the vertically incident pencil beam has a Gaussian distribution in the horizontal direction and a spread function similar to that of a water droplet. Assuming that the vertical function is G(z) and the horizontal function is F(x,y,z), and that the horizontal function takes the form of a Gaussian distribution, it can be expressed as follows.

$$\mathrm{dose}_{\mathrm{Pencil}}(x,y,z) = F(x,y,z) \cdot G(z)$$

F(x,y,z) is presented for multiple Coulomb scattering (MCS) function according to the Fermi-Eyges theroy, and it indicates that σ_{MCS}^2 becomes more diffuse in the lateral direction as it grows.

$$F(x,y,z) = \frac{1}{2\pi\sigma_{MCS}^2} \exp\left(-\frac{x^2+y^2}{2\pi\sigma_{MCS}^2}\right)$$

G(z) is a function of the depth direction within a single pencil beam, represented by the following equation.

$$G(z) = G_{water}(Z_{eff}) \cdot \left(\frac{SSD + Z_{eff}}{SSD + z}\right)^2$$

Using the above electron dose calculation engine, the bolus optimization system that can match the desired target-line was built using backtracking line search algorithm. The objective function is the mean square error (mse) of the dose within the target-line according to the change of the

thickness of bolus, and the parameters α and β were set as 0.5 and 0.9, respectively.

3 Results

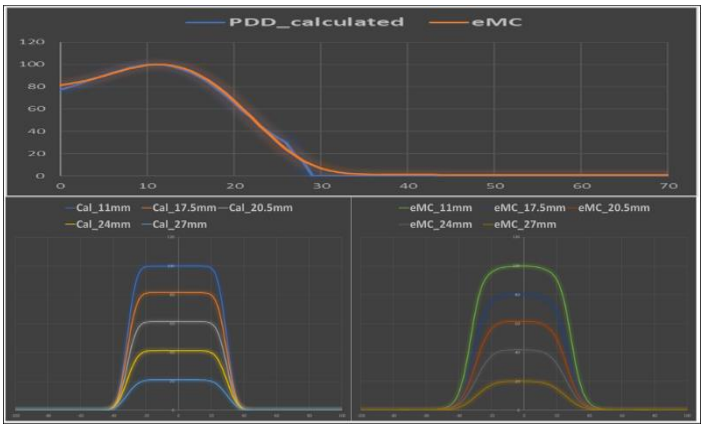


Figure 1: Comparison of percentage depth dose and profile calculated with the our electron beam dose calculation program and the commercialized dose calculation algorithm (eMC).

Figure 1 shows the comparison of PDD and profile calculated by eMC, which is a Monte-Carlo simulation-based electron dose calculation algorithm within Eclipse, and the our developed electron dose calculation engine. For PDD, the root mean square error (RMSE) analysis showed an error rate of 2.07. For profile, the RMSE was calculated to be 3.10, 3.89, 3.34, 2.30, and 1.54 at depths of 11, 17.5, 20.5, 24, and 27 mm, respectively.

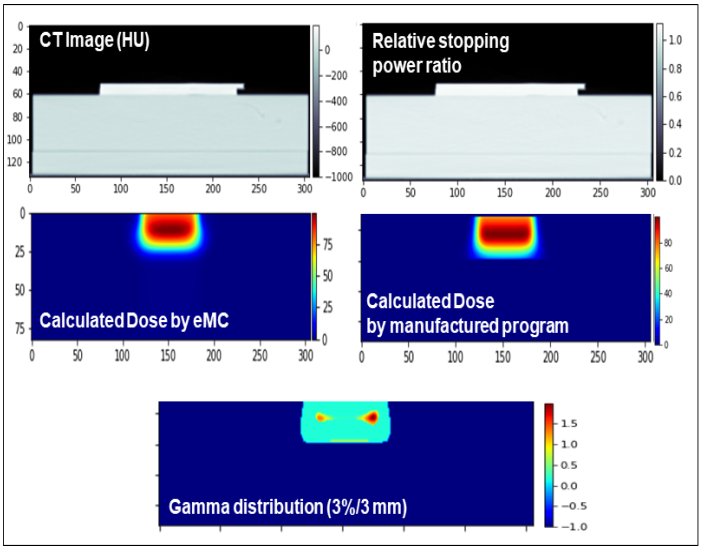


Figure 2. Dose and gamma distribution calculated with the developed dose calculation system and commercialized dose calculation algorithm (eMC) at the center axial within the phantom.

A result of the gamma evaluation between the dose distribution calculated by eMC and our dose calculation engine is 94.97% within 3%/3 mm criterion.

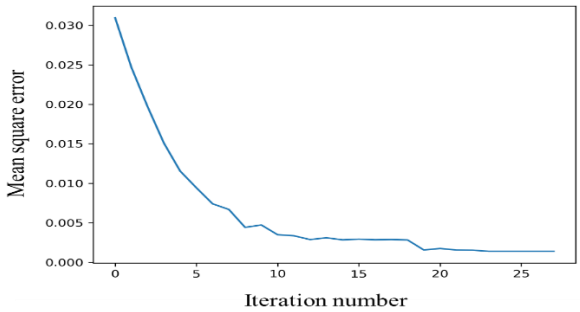


Figure 3. Mean square error variation as the number of iterations.

As shown in figure 3, after more than 25 iterations, the MSE converged to about 0.0014.

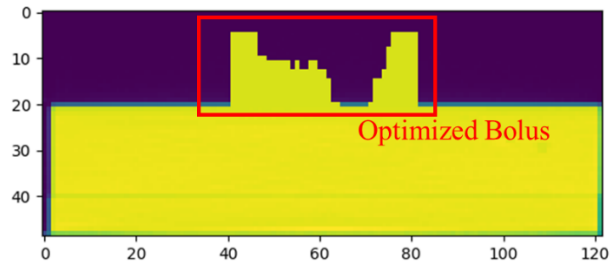


Figure 4. Optimized bolus shape

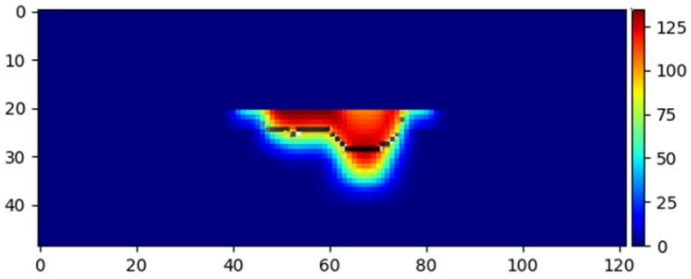


Figure 5. Dose distribution inside the body. The black dot represents the target line, which is the 95% dose fraction.

When the target line is normalized to the 95% prescribed dose, the Dmax is 135% within the target, and the mean value and the root mean square error of the dose on the target line is 95.23% and 15.17, respectively.

4 Discussion

Based on the results, it is shown that our developed electron radiation dose calculation engine can be utilized in optimization and is somewhat consistent with commercially available algorithms. However, there is a difference with the eMC result due to the calculation assuming S(x', y') = 1 and ignoring the beam divergence. Although the computational algorithm is less accurate than commercially available algorithms, the calculation algorithm was enough to be used as a dose calculation algorithm for bolus optimization.

The current optimization algorithm has limited simulation of scatter effects in the lateral direction to improve optimization speed. This causes a difference between the actual final calculation and the optimization stage. If a secondary optimization with more precise calculations is added after the first optimization, the accuracy of the optimization algorithm can be improved. Therefore, we will try to improve the dose calculation algorithm and optimization algorithm as future research.

The above studies have not yet been validated using measurements. Therefore, as a future study, we would like to manufacture the optimized bolus using a 3D printer and verify its clinical use through measurement.

5 Conclusion

In this study, we developed the electron dose calculation engine and optimization system to build a bolus optimization system for ECT and evaluated its feasibility.

References

[1] N. D. Tapley, Clinical Applications of the Electron Beam (John Wiley & Sons, New York, 1976).

[2] R. R. Million, J. T. Parsons, F. J. Bova, and K. J. Kalbaugh, “"Electron beam: The management of head and neck cancer,"" The Role of High Energy Electrons in the Treatment of Cancer: 25th Annual San Francisco Cancer Symposium (Frontiers of Radiation Therapy and Oncology, San Francisco, 1991), pp. 107–-27.

[3] K. R. Hogstrom, “"Electron beam therapy: Dosimetry, planning, and techniques,"" Principles and Practice of Radiation Oncology, edited by C. Perez et al. (Lippincott, Williams, and Wilkins, Baltimore, 2003), pp. 252–-282.

[4] Low DA, Starkschall G, Bujnowski SW, Wang LL, Hogstrom KR. Electron bolus design for radiotherapy treatment planning: bolus design algorithms. Med Phys. 1992;19(1):115–24. DOI: https://doi.org 10.1118/1.596885

[5] Perkins GH, McNeese MD, Antolak JA, Buchholz TA, Strom EA, Hogstrom KR. A custom three-dimensional electron bolus technique for optimization of postmastectomy irradiation. Int J Radiat Oncol Biol Phys. 2001;51(4):1142–51. DOI: https://doi.org/10.1016/S0360-3016(01)01744-8

[6] Kudchadker RJ, Antolak JA, Morrison WH, Wong PF, Hogstrom KR. Utilization of custom electron bolus in head and neck radiotherapy. J Appl Clin Med Phys. 2003; 4(4):321–33. DOI: https://doi.org/10.1120/jacmp.v4i4.2503

[7] Low DA, Starkschall G, Sherman NE, Bujnowski SW, Ewton JR, Hogstrom KR. Computer-aided design and fabrication of an electron bolus for treatment of the paraspinal muscles. Int J Radiat Oncol Biol Phys. 1995;33(5):1127–38 DOI: https://doi.org/10.1016/0360-3016(95)00257-x

[8] Klein EE. Modulated electron beams using multi-segmented multileaf collimation. Radiother Oncol. 1998;48(3):307–11. DOI: https://doi.org/10.1016/S0167-8140(98)00050-4

[9] Ma CM, Pawlicki T, Lee MC, et al. Energy- and intensity-modulated electron beams for radiotherapy. Phys Med Biol. 2000;45(8):2293–311. DOI: https://doi.org/10.1088/0031-9155/45/8/316

[10] Blomquist M, Karlsson MG, Zackrisson B, Karlsson M. Multileaf collimation of electrons--clinical effects on electron energy modulation and mixed beam therapy depending on treatment head design. Phys Med Biol. 2002;47(7):1013–24. DOI: https://doi.org/10.1088/0031-9155/47/7/30

[11] Hogstrom KR, Boyd RA, Antolak JA, Svatos MM, Faddegon BA, Rosenman JG. Dosimetry of a prototype retractable eMLC for fixed-beam electron therapy. Med Phys. 2004;31(3):443–62. DOI: https://doi.org/10.1118/1.1644516

[12] Hogstrom KR, Mills MD, Almond PR. Electron beam dose calculations. Phys Med Biol. 1981;26:445–459. DOI: https://doi.org/10.1088/0031-9155/26/3/008

6 Acknowledgement

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (NRF-2022R1F1A107219912 and NRF-RS-2023-00212193).

A Python Script to Analyse Treatment Planning Throughput

Ezhilalan Ramalingam

Department of Radiotherapy Physics, Cork University Hospital, Cork, Ireland

Abstract: The aim of this study is to estimate radiotherapy treatment planning throughput from the data extracted from our Oncology Information System (Mosaiq, OIS). We have developed an in-house Python script[1] to analyse the data exported from the OIS. The Quality Check List (QCL) data captured between April 2021 and March 2022 and from April 2022 to March 2023 by our treatment planning team has been used in the data analysis. The QCL data includes but not limited to planning categories such as radical, palliative, and re-planning. The OIS data has been exported in an Excel file format, preprocessed to eliminate extraneous information and saved as a Comma Separated Value (CSV) file. Tasks like image fusion, contouring, and the reconstructions of previous treatments are inherent components of the planning process and have been excluded from the data analysis. Measure of central tendency and dispersion such as mean, median, mode, and standard deviation have been documented for the daily, weekly, and monthly treatment planning throughput.

While the throughput distributions of individual planners deviate from normality, the collective probability distributions of all planners in the study display nearly normal (symmetrical) distributions. The selection of the specific time period (April to March) for analysis aligns with our center's leave calendar, ensuring the inclusion of annual leave taken by the planners involved in the study.
From April 2022 to March 2023, an average of 3.9 planners generated a mean of 9.5 treatment plans per day. Similarly, the mean number of treatment plans produced per week by an average of 6.8 planners is 45.6, and the mean number produced per month by an average of 7.9 planners is 197.7. With a total of 249 working days, averaging 4.8 days per week, the daily throughput can be approximately scaled to weekly ($9.5 \times 4.8 = 45.6$), and the weekly figure can be approximately scaled to monthly ($45.6 \times 4.33 = 197.5$). From April 2021 to March 2022, an average of 3.9 planners generated a mean of 7.7 treatment plans per day. Similarly, the mean number of treatment plans produced per week by an average of 7.3 planners is 35.9, and the mean number produced per month by an average of 8.8 planners is 146.8. With a total of 230 working days (with approximately 3 weeks lost due to a cyber-attack), averaging 4.7 days per week, the daily throughput can be approximately scaled to weekly, and the weekly figure can be approximately scaled to monthly, as outlined earlier. Due to variations in individual planner's work patterns, their throughput also needs to be analysed independently.

1 Introduction

Due to recent advancements in radiotherapy treatment techniques, treatment planning process in general has become more complicated and time-consuming. As a result, managing the treatment planning workload coming into the department is crucial for efficiency and smooth functioning. A well-organised planning department, equipped with efficient workflows and adequate staffing, minimises bottlenecks in day-to-day operations. However, in situations where the workload significantly increases or when there is a shortage of staff, managing the planning workload can be challenging, especially if it’s uncertain how many plans the currently available staff can safely handle. Nonetheless, with knowledge of treatment planning throughput, extreme fluctuations in workload can be efficiently managed by implementing proper control measures such as assurance of adequate staffing levels and means to monitor workload periodically. In this study, we aim to estimate our treatment planning throughput through the analysis of data recorded in the Mosaiq OIS.

2 Materials and Methods

For fast data extraction and analysis, an Excel spreadsheet can be used. However, handling large amounts of data in Excel becomes impractical without the use of extensive Visual Basic (VB) codes running in the background or resorting to laborious manual data manipulation to derive meaningful insights from the collected data. To overcome this hurdle, a Python script (version 3.8) was written to read and process the Excel data extracted from the OIS. For the data analysis, a Python package called pandas (version 1.2.3) was used. The package Pandas is a versatile, powerful, and efficient open-source Python library that can be used to read, manipulate, and analyse data from a CSV file. Our Python script includes the pandas and matplotlib packages for data analysis and plotting, respectively. The script reads the data from a CSV file (created from an Excel file), forms a Data Frame (DF), and manipulates the data depending on the user input. The Excel file exported from the OIS was

preprocessed to include only relevant data columns such as the date of task capture, patient identification (ID), task captured, and planner initials. The filtered data was then saved as a CSV file. The Python script utilises the pandas library to read data from the ‘date’, 'task' and 'planner’s initial (Dosimetrist-dos) columns, storing them as a Data Frame (DF) for further manipulation. The DF is then processed to a specific set of tasks in the 'task' column (e.g., '1B Planning' & '2B Planning') for a given number of planners in the 'dos' column. The script then calculates various parameters, including the total number of days, weeks and months during which the specified tasks were captured. In addition, the total number of tasks (plans) per unique planner is also reported. Metrics such as the total number of unique planners present per day, per week, and per month are also reported. The script is capable of generating bar charts to visualise the number of tasks captured per day, per week, and per month. Three separate scripts were developed for daily, weekly, and monthly data analysis. Descriptive statistical metrics such as minimum, maximum, mean, mode, standard deviation, skewness, and kurtosis have been calculated. To validate the script, it was initially tested on a small data file collected for a month from the OIS. The results were then compared against manually calculated figures directly in an Excel file.

3 Results

The table below (Table 1) presents the central tendency statistics calculated for the data collected over the two-year period mentioned earlier. Planning throughput for daily, weekly, and monthly periods is depicted, aggregating contributions from all planners in the study. However, it's important to note that, due to varied work patterns, the throughput for individual planners can only be accurately calculated from the specific data collected for each planner. Table 1 provides a basis for estimating the combined throughput from the entire dataset.

Central tendency and dispersion statistics	April 2022 to March 2023			April 2021 to March 2022		
	daily	weekly	monthly	daily	weekly	monthly
Mean number of plans	9.5	45.6	197.7	7.7	35.9	146.8
Median number of plans	9.0	44.0	187.5	7.0	35.0	163.5
Mode number of plans	10.0	33.0	(n/a)	5.0	46.0	(n/a)
Std. deviation	4.3	13.4	43.9	3.7	11.1	46.1
Minimum number of plans	1.0	16.0	149.0	1.0	8.0	47.0
Maximum number of plans	21.0	79.0	294.0	20.0	59.0	216.0
Skew	0.6	0.2	1.1	0.6	-0.2	-1.1
Kurtosis	0.1	0.1	0.6	0.4	-0.3	1.1
Mean number of planners	3.9	6.8	7.9	3.9	7.3	8.8
Total number of plans	2372.0	2372.0	2372.0	1761.0	1761.0	1761.0
Total time (days, weeks, months)	249.0	52.0	12.0	230.0	49.0	12.0

Table-1, the central tendency and dispersion of the throughput data calculated for the two consecutive years.

The figures (Figure 1 and Figure 2) display the histogram and the Quantile-Quantile (QQ) plot of the weekly planning throughput for the period April 2022 to March 2023 respectively.

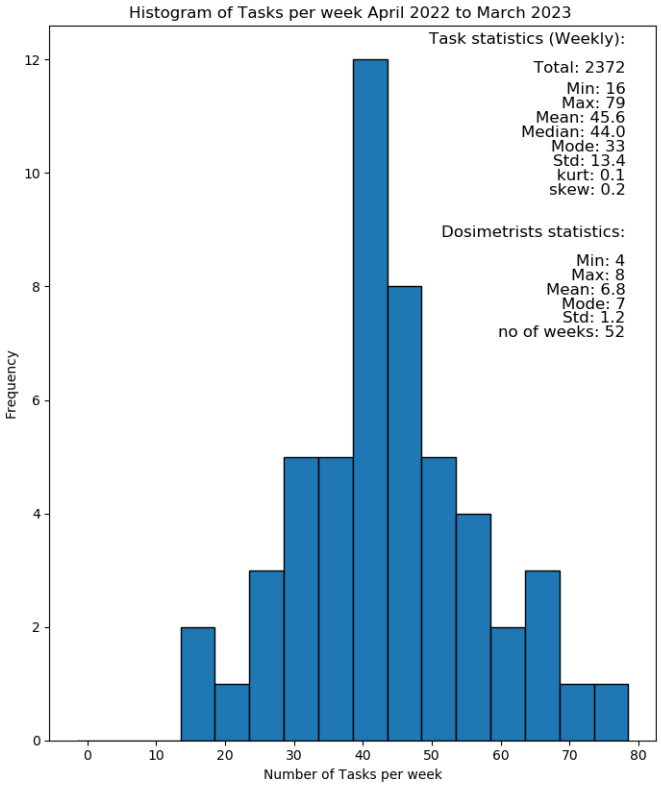


Figure-1, weekly throughput histogram

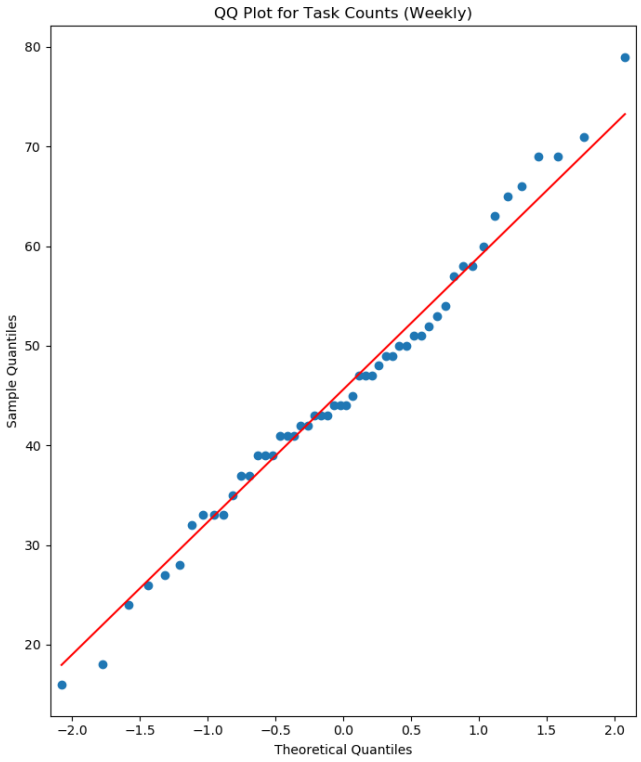


Figure-2, weekly throughput QQ plot

4 Discussion

The throughput of radiotherapy treatment planning can be influenced by numerous factors, including planner experience, plan complexity, the efficiency of the planning system used, and the collaborative influence of other Multi-Disciplinary Team (MDT) members involved in the process. The rate of treatment plan production is likely to be affected by all of these aforementioned factors. By analysing the data captured in the Oncology Information System (OIS), treatment planning throughput can be reasonably estimated. In our study, the estimated planning throughput was particularly influenced by factors such as planner work patterns and experience. To mitigate this impact, we exclusively included experienced planners in the study, examining both combined and independent throughput for all planners involved. While individual planner statistics are not presented in this overview, such data can be easily obtained by applying appropriate filters within the Python code.

5 Conclusion

From our study, it can be concluded that treatment planning throughput can be consistently estimated from the data collected from an OIS using an independent computer programming language such as Python. The time required to develop such a system is a one-time effort, and an appropriately validated script can be applied repeatedly for any length of data. Additionally, this method can be employed periodically to monitor workload, allowing for proactive measures to be taken as needed.

References

[1] Rex A. Cardan, PhD, Elizabeth L. Covington, PhD ∗, John B. Fiveash, MD, Richard A. Popple, PhD. " Using a whiteboard web application for tracking treatment workflow metrics for dosimetrists and physicians". Medical Dosimetry 45 (2020) 73–76. DOI: 10.1016/j.meddos.2019.07.002

Development of dynamic collimator mixed beam radiotherapy to explore the dosimetric benefits of non-coplanar and mixed photon-electron beam setups

Chengchen Zhu[1], Gian Guyer[1], Jenny Bertholet[1], Silvan Mueller[1], Hannes A Loebner[1], Marco FM Stampanoni[2], Michael K Fix[1], and Peter Manser[1]

[1]Division of Medical Radiation Physics and Department of Radiation Oncology, Inselspital, Bern University Hospital, and University of Bern, Bern, Switzerland
[2]Institute for Biomedical Engineering, ETH Zürich and PSI, Villigen, Switzerland

Abstract Dynamic collimator mixed beam radiotherapy (colli-DYMBER) is a novel treatment technique that uses non-coplanar table angles, dynamic gantry and collimator rotation, mixed photon-electron beams/arcs, and intensity modulation.
The aim of this work is to solve the beam angle optimization (BAO) problem dosimetrically using a direct aperture optimization (DAO)-based pathfinding strategy that iteratively adds promising apertures to form paths. We show how the developed dosimetrically motivated pathfinding strategy can be used to investigate the dosimetric benefit of using non-coplanar beam setups and/or mixed photon-electron modality by comparing colli-DYMBER plans with non-coplanar photon-only plans (dynamic collimator trajectory radiotherapy, colli-DTRT), coplanar mixed beam plans (coplanar-DYMBER), and coplanar photon-only plans (coplanar-VMAT) for three clinically motivated cases. The dosimetric benefit observed by incorporating non-coplanar/mixed beam setups varied for the considered cases.

1 Introduction

Ongoing research in radiotherapy aims to improve treatment quality by developing novel techniques that maximize organ-at-risk (OAR) sparing while maintaining treatment efficacy. Current clinic practices include intensity modulated radiation therapy (IMRT), where intensity modulation is performed using a photon multileaf collimator (pMLC) [1], and volumetric modulated arc therapy (VMAT) where the dynamic gantry rotation during beam on improves delivery efficiency without compromising the dosimetric plan quality from IMRT [2]. In recent years, non-coplanar radiotherapy and mixed beam photon-electron beam radiotherapy have been explored to further improve the dosimetric plan quality [3-7]. Bertholet et al. explored dynamic collimator trajectory radiotherapy (colli-DTRT), where non-coplanar photon-only paths were determined using dosimetrically motivated pathfinding based on iterative fluence map optimization (FMO) and beam elimination [4]. In the work of Guyer et al., coplanar mixed photon-electron arcs (Arc-MBRT) were manually set for breast cases to improve delivery efficiency without compromising the dosimetric plan quality compared to Static-MBRT (with static electron fields) [5]. However, the benefit from non-coplanarity or mixed photon-electron treatments varies for different treatment sites and differences in the pathfinding strategy or beam setup raise difficulties in discerning the origin of the outcomes.

A novel treatment technique called dynamic collimator mixed beam radiotherapy (colli-DYMBER) exploits different degrees of freedom (DoFs) offered by the clinical machines: dynamic gantry rotation, multiple different table angles, dynamic collimator rotation, and mixed photon and electron beams, both intensity-modulated with pMLC. The beam angle optimization (BAO) problem for colli-DYMBER is solved with a dosimetrically motivated pathfinding approach that simultaneously optimizes photon and electron contributions based on hybrid-direct aperture optimization (H-DAO) [6].

In this work, we show how the pathfinding approach for colli-DYMBER can be used to investigate the dosimetric benefit of non-coplanarity/mixed beam modality on the treatment plan. The dosimetric plan quality was compared to variations of colli-DYMBER by enabling/disenabling the non-coplanarity or mixed beam feature using the same pathfinding strategy for three clinically motivated cases.

2 Materials and Methods

The dosimetrically motivated pathfinding approach for colli-DYMBER is illustrated in Fig. 1.

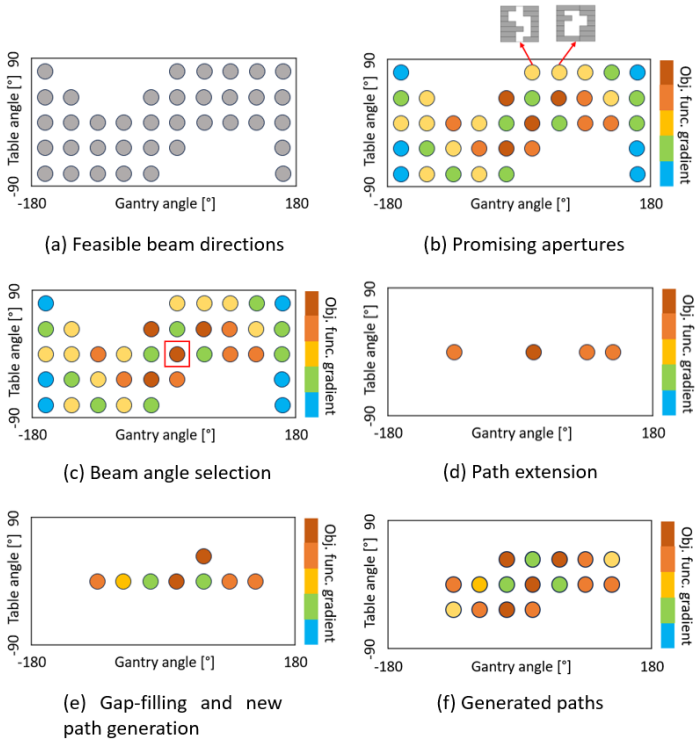


Fig. 1 Pathfinding approach for colli-DYMBER. Favorable beam directions have steeper objective function gradients (red) while unfavorable beam directions have shallower gradients (blue).

First, the user defines the resolution (10° used in this work) along the gantry and table axes, setting up a grid of map points of different gantry-table angle combinations for different beam qualities: 6 MV (photon), 6, 9, 12, 15, 18, and 22 MeV (electron). The collimator angle is aligned with the superior-inferior axis of the patient (constant for 0° table angle). A fixed source-to-surface distance (SSD) of 80 cm is used for electron beams/arcs to limit electron in-air scattering resulting in non-isocentric electron arcs. Beam directions leading to collisions are excluded by a collision prediction tool [7], as well as beams coming through the end of the CT stack. Electron beams with insufficient range to reach the target or impinging on the table before reaching the patient are eliminated from the set of feasible beam directions (Fig. 1a). Each beam is discretized into beamlets of 5 x 5 mm^2 and the beamlet dose calculation is performed using voxel grids of 5 x 5 x 5 mm^3. A column generation algorithm is used to iteratively generate promising apertures for all feasible beams and select the beam direction with the most promising aperture (Fig. 1b, 1c) based on the gradient of the objective function, defined as a weighted sum of dose-volume and generalized equivalent uniform dose objectives. The selected beam directions are formed into paths and the maximum number of paths allowed is defined by the user. In this path formation process, non-adjacent beams can also be selected (Fig. 1d) and the gap between them is subsequently filled with intermediate beam directions to form connected paths (Fig. 1e). The path formation terminates when a user-defined total gantry angle range is achieved, or no more promising apertures are available (Fig. 1f).

The colli-DYMBER paths are resampled to increase the gantry angle resolution to 5° and a second beamlet dose calculation (with finer voxel grids of 2.5 x 2.5 x 2.5 mm^3) is performed. A second H-DAO is performed to determine the intensity modulation along the resulting colli-DYMBER paths, optimizing the aperture weights and shapes [6]. A final dose calculation considering the exact MLC model is then performed with subsequent monitor unit (MU) weight re-optimization using a limited-memory Broyden-Fletcher-Goldfarb-Shanno algorithm.

Both beamlet and final dose calculations are performed using the Swiss Monte Carlo Plan (SMCP) integrated beam models and dose calculation algorithms VMC++ [8] and macro Monte Carlo [11-13] for photon and electron beams, respectively.

The treatment planning process is used to create plans for a TrueBeam system (Varian, a Siemens Healthineers Company, Germany) and three clinically motivated cases have been considered in this study: one brain (30 x 2 Gy), one breast (16 x 2.65 Gy), and one pelvis (28 x 1.8 Gy) case. All prescriptions referred to the median dose to the planning target volume (PTV). For each case, four plans were created to investigate the potential benefit of using non-coplanar beam setups or mixed photon-electron beams in improving OAR sparing while maintaining similar target coverage:

- colli-DYMBER plan (non-coplanar, mixed beam)
- colli-DTRT plan (non-coplanar, photon beams only)
- coplanar-DYMBER plan (coplanar, mixed beam)
- VMAT plan (coplanar, photon beams only) as reference

For colli-DYMBER and its variations, a maximum of 7 arcs and a total gantry angle range of 720° were used for the brain and pelvis cases. For the breast case, a maximum of 10 arcs for a 1080° total gantry angle range were used. VMAT plans with the same total gantry angle range as the corresponding colli-DYMBER plans were created for the brain (2 arcs), breast (4 arcs), and pelvis (4 arcs) cases. Dose distributions and differences in dosimetric endpoints were evaluated and compared. Paddick conformity index (CI) and homogeneity index (HI = $(D_{2\%} - D_{98\%})/D_{50\%}$) were calculated for all plans.

3 Results

For the mixed photon-electron plans, the photon and electron contributions to the mean PTV dose are reported in Table 1. The electron contribution increases when restricting to use only coplanar beam setups. The highest electron contribution observed is 39.2% in the coplanar-DYMBER plan for the pelvis case.

		colli-DYMBER	coplanar-DYMBER
Brain case	X	97.6%	75.6%
	E	2.4%	24.4%
Breast case	X	94.5%	88.9%
	E	5.5%	11.1%
Pelvis case	X	61.0%	60.8%
	E	39.0%	39.2%

X = photon contribution to mean PTV dose; E = electron contribution to mean PTV dose.

Table 1 Photon and electron contributions to mean PTV dose for colli-DYMBER and coplanar-DYMBER for the considered cases.

Similar target conformity and homogeneity are reported for all cases as shown in Table 2.

		colli-DYMBER	colli-DTRT	coplanar-DYMBER	VMAT
Brain case	CI	**0.499**	0.496	0.497	**0.499**
	HI	**0.088**	0.089	0.100	0.100
Breast case	CI	**0.498**	**0.498**	0.497	**0.498**
	HI	**0.074**	0.079	0.079	0.079
Pelvis case	CI	0.493	0.495	**0.496**	0.486
	HI	**0.059**	0.073	**0.059**	0.078

Table 2 Conformity and homogeneity indexes for the considered cases. The best values for each row are highlighted in bold.

Fig. 2 illustrates the difference in dosimetric end-points for the considered cases relative to VMAT for colli-DYMBER, colli-DTRT, and coplanar-DYMBER.

Substantial dose reductions to OARs were observed for the non-coplanar techniques compared to VMAT for the brain case (Fig. 2, top). Mean doses to the OARs were reduced by 1.2-4.7 Gy for colli-DYMBER and colli-DTRT compared

to VMAT. $D_{2\%}$ were reduced by 1.1-14.6 Gy for serial organs. coplanar-DYMBER also reduced D_{mean} and $D_{2\%}$ for the considered organs by 0.4-3.0 Gy and 1.1-9.8 Gy, respectively, compared to VMAT, except for the ipsilateral optic nerve where $D_{2\%}$ was increased by 7.1 Gy.

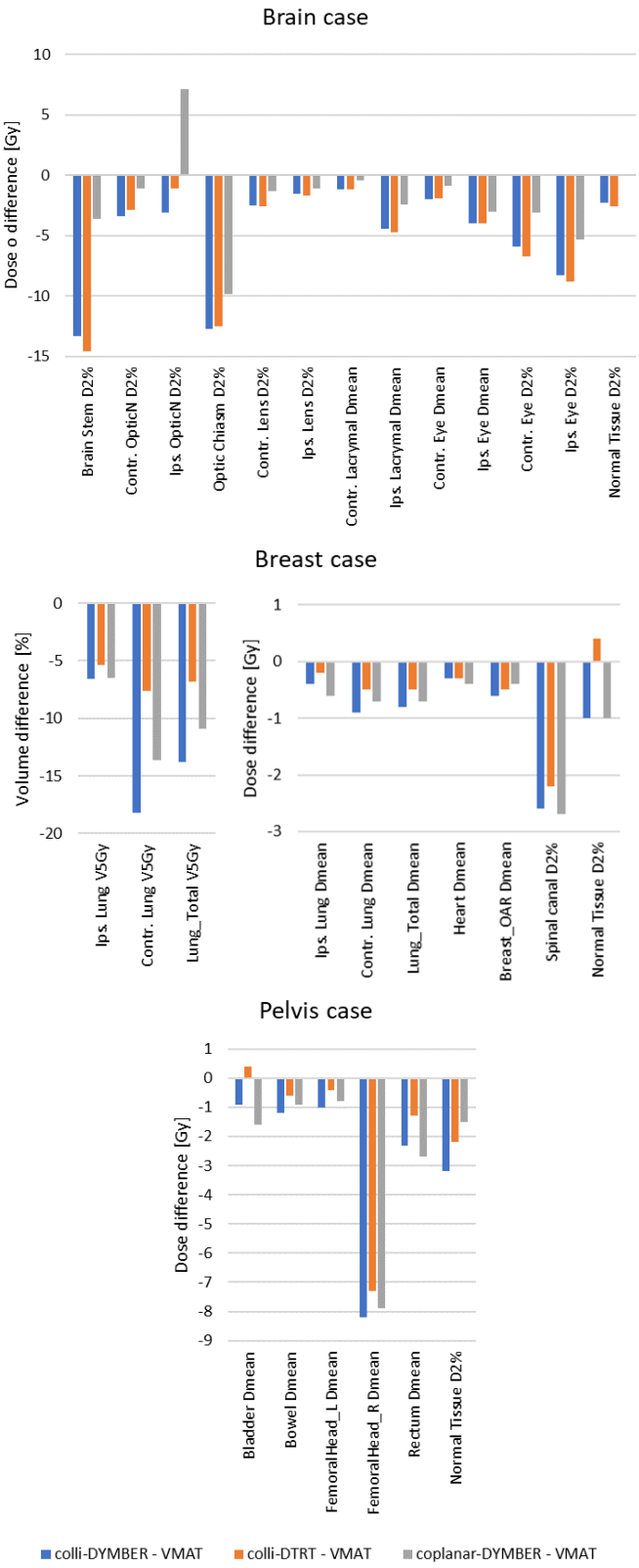


Fig. 2 Differences in dosimetric end-points for colli-DYMBER (blue), colli-DTRT (orange), and coplanar-DYMBER (gray), compared to the VMAT plan for the brain (top), breast (middle), and pelvis (bottom) case. Negative values indicate a reduction in OAR dose compared to VMAT.

For the breast case (Fig. 2, middle), a larger reduction in V_{5Gy} to both lungs is observed for coplanar-DYMBER compared to colli-DTRT, while colli-DYMBER obtained slightly lower doses to the contralateral lung compared to coplanar-DYMBER. Reductions in mean doses to the OARs are below 1.0 Gy for the non-coplanar/mixed beam techniques compared to VMAT.

Greater reductions in mean doses to OARs were also achieved by coplanar-DYMBER compared to colli-DTRT for the pelvis case (Fig. 2, bottom), except for the near maximum dose to the normal tissue, which is lower when non-coplanar beams were used. Further dose reductions to the bowel and femoral heads were achieved by the colli-DYMBER plan compared to coplanar-DYMBER, albeit at the cost of increased mean doses to the bladder and rectum. Fig. 3 illustrates the axial dose distributions for the pelvis case.

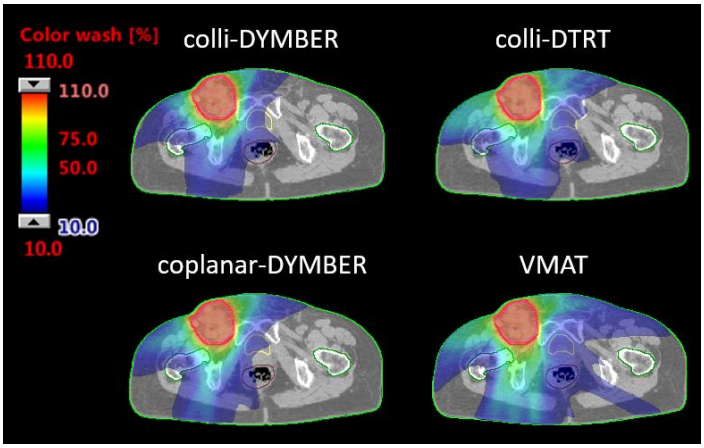


Fig. 3 Dose distributions on an axial plane for the pelvis case.

4 Discussion

In this work we explored the dosimetric benefits of non-coplanar photon/electron fields/arcs with a novel treatment technique called colli-DYMBER, combining non-coplanarity, mixed beam, and dosimetrically motivated BAO. The increased DoF was expected to improve the dosimetric plan quality owing to the increased solution space. Bertholet et al. investigated colli-DTRT using an FMO-based dosimetrically motivated pathfinding strategy. However, mechanical constraints for the delivery were not taken into account during the path determination process and this could lead to degradation in the dosimetric plan quality for the final plan. Mueller et al. investigated DYMBER which combined dynamic photon trajectories with static electron fields [6]. The addition of electrons to dynamic photon trajectories brought dosimetric benefits in OAR sparing, but only static electron fields were considered which were manually set up. Mixed photon/electron arcs have also been investigated by Guyer et al. in a technique called Arc-MBRT, where the use of manually defined

coplanar photon and electron arcs outperformed the photon-only plans dosimetrically [7].

In these studies, the differences in the pathfinding strategy used or in the planner's experience to set up beams manually pose difficulties in comparing the different techniques and identifying the origin of the benefit observed (non-coplanarity or combination of photons and electrons). With colli-DYMBER, we can "enable/disable" the non-coplanarity or mixed beam feature using the same optimization algorithm and dosimetrically motivated BAO. The benefit of non-coplanarity and mixed beam modality can be identified conclusively from the comparison between variations of colli-DYMBER enabling the comparison between the different techniques.

For the cases analyzed in this study, the increased DoF improved the dose distributions compared to coplanar photon treatments overall. For the brain case, non-coplanarity brought a larger benefit in improving the dose distribution compared to the inclusion of electrons, suggesting that the use of electrons might not be needed if favorable non-coplanar beam directions can be found. On the other hand, for the breast case, the mixed beam modalities brought the largest improvement compared to VMAT. The additional benefit of non-coplanarity was not substantial. For the pelvis case, mixed beam was also revealed to be beneficial in improving the dosimetric plan quality. This might be explained by the fact that electrons become more important to spare critical organs for cases where the collision-free space is limited (Fig. 2, Table 1). A study with a larger cohort is needed to generalize the findings of this work, as well as the assessment of delivery times for the different techniques.

5 Conclusion

We developed a dosimetrically motivated BAO for colli-DYMBER which enabled equitable comparisons with its variants by selectively disabling one of the features. By combining both non-coplanarity and mixed beam qualities, the dosimetric plan quality can be improved compared to coplanar photon-only treatments. For the considered cases, a reduction of doses in critical organs has been generally observed for the proposed method, suggesting that colli-DYMBER is a promising treatment technique for future radiation therapy treatments.

Acknowledgements

This work was supported by Varian, a Siemens Healthineers Company and SNSF grant 200021_185366. Dose calculations and optimizations were performed on UBELIX (www.id.unibe.ch/hpc), the high-performance cluster at the University of Bern.

References

[1] Bortfeld T. IMRT: A review and preview. Phys Med Biol 2006;51. https://doi.org/10.1088/0031-9155/51/13/R21.

[2] Otto K. Volumetric modulated arc therapy: IMRT in a single gantry arc. Med Phys 2007;35:310–7. https://doi.org/10.1118/1.2818738.

[3] Xu Y, Ma P, Hu Z, Tian Y, Men K, Wang S, et al. Non-coplanar Volumetric Modulated Arc Therapy for Locoregional Radiotherapy of Left-sided Breast Cancer Including Internal Mammary Nodes. Radiol Oncol 2021;55:499. https://doi.org/10.2478/RAON-2021-0045.

[4] Bertholet J, Zhu C, Guyer G, Mueller S, Volken W, Mackeprang PH, et al. Dosimetrically motivated beam-angle optimization for non-coplanar arc radiotherapy with and without dynamic collimator rotation. Med Phys 2023:1–14. https://doi.org/10.1002/mp.16899.

[5] Guyer G, Mueller S, Mackeprang PH, Frei D, Volken W, Aebersold DM, et al. Delivery time reduction for mixed photon-electron radiotherapy by using photon MLC collimated electron arcs. J Phys D Appl Phys 2023;68. https://doi.org/10.1088/1361-6560/ad021a.

[6] Guyer G, Mueller S, Koechli C, Frei D, Volken W, Bertholet J, et al. Enabling non-isocentric dynamic trajectory radiotherapy by integration of dynamic table translations. Phys Med Biol 2022;67:175003. https://doi.org/10.1088/1361-6560/AC840D.

[7] Guyer G, Mueller S, Wyss Y, Bertholet J, Schmid R, Stampanoni MFM, et al. Technical note: A collision prediction tool using Blender. J Appl Clin Med Phys 2023;24:1–7. https://doi.org/10.1002/acm2.14165.

[8] Kawrakow I, Fippel M. VMC++, a fast MC algorithm for Radiation Treatment planning. Use Comput Radiat Ther 2000:126–8. https://doi.org/10.1007/978-3-642-59758-9_46.

[9] Neuenschwander H, Born EJ. A macro Monte Carlo method for electron beam dose calculations. Phys Med Biol 1992;37:107–25. https://doi.org/10.1088/0031-9155/37/1/007.

Online adaptive proton therapy for esophageal cancer: adaptation rate and automated strategies comparison

Camille Draguet[1,2], Pieter Populaire[2,3], Macarena Chocan Vera[1], Karin Haustermans[2,3], John A. Lee[1], Edmond Sterpin[1,2], Ana M. Barragán-Montero[1]

[1] UCLouvain, Institut de Recherche Expérimentale et Clinique, Molecular Imaging Radiotherapy and Oncology (MIRO), Brussels, Belgium
[2] KU Leuven, Department of Oncology, Laboratory of Experimental Radiotherapy, Leuven, Belgium
[3] University Hospitals Leuven, Department of Radiation Oncology, Leuven, Belgium

Abstract Adaptive radiotherapy shows potential in enhancing patient outcomes through the acquisition of longitudinal images during treatment, facilitating the management of anatomical changes that happen during treatment. Given the resource and time cost of adaptive treatment, careful consideration is put to the selection of patients who might benefit from it. The goal of this work is twofold. First, it investigates the sensitivity of proton therapy adaptation rate to various residual setup errors in patients with esophageal cancer. Second, four different fully automatic strategies for online adaptive planning are implemented and compared. These strategies combine deformable registration and deep learning based methods, both for contouring and planning. Our preliminary results serve as a proof of concept for the feasibility of the presented planning strategies, which are compared to a fully manual approach in terms of timing and target coverage (D98%). We show that, on the single patient used for this preliminary study, the four strategies achieve similar D98% as the manual approach while being much faster. The clinical significance of these results still needs to be confirmed on a larger cohort of patients.

1. Introduction

Adaptive radiotherapy (ART) aims at ensuring an optimal treatment for the patient despite anatomical changes, by acquiring longitudinal images (e.g., weekly or daily) and adjusting the treatment plan to the new anatomy (1). ART has shown potential for improving patient outcomes (2) although its implementation is known to be resource intensive. When performed offline, clinicians spend some time (i.e., a few hours) to prepare the new treatment. In contrast, online ART requires highly optimized workflows that ideally generate plans in a few minutes, since the patient is waiting on the treatment couch. Commercial solutions for streamlined online adaptive intensity-modulated radiotherapy (IMRT) are becoming available, achieving adapted plans within 15 min up to an hour (3). These solutions typically integrate artificial intelligence (AI) (4,5) and deformable registration (5) for automatic contouring, as well as fast plan optimization algorithms through template-based approaches (6).

In the case of protons (PT), online adaptive solutions are lagging behind (4,7). Indeed, treatment planning is typically slower and more challenging in PT, since it entails robust optimization to be safe against range uncertainties in addition to setup uncertainties, delineation errors, etc. already addressed in conventional radiotherapy (XT). Therefore, clinical goals are first often evaluated on the initial plan recomputed on the new anatomy, here referred as the NA dose, and if these goals are not met, only then the patient should be considered for adaptation. After the patient gets selected for adaption, the adaptation process should be streamlined for the sake of fast clinical workflow and high patient throughput, while still minimizing the intra-fractional motion. Consequently, in this work, we present a comparison of online adaptive strategies for proton therapy (OAPT), all relying on fully automated methods. Since AI-guided treatment planning and contouring has recently gained popularity, the proposed methods partly rely on such techniques.

Esophageal cancer (EC) is the 8th most common cancer type worldwide and is associated with a severe reduction in the patients quality of life. Locally advanced EC is typically treated with neoadjuvant chemoradiotherapy followed by surgery (8). PT has shown to improve toxicity by better sparing organs-at-risk (OARs) like lungs and heart.

This study aims at investigating online adaptive strategies for PT from the perspective of treatment plan optimization and evaluation. First the adaptation rate for several EC patients will be evaluated, in agreement with the PROTECT clinical trial guidelines (9). The benefits of using adaptation for this cohort of patients are also reported. Secondly, the first results obtained by comparing automated strategies, including AI-models for OARs contouring and dose prediction, will be presented.

2 Materials and Methods

Patient cohort and planning strategies

Between 2016 and 2018, a cohort of 40 patients diagnosed with locally advanced EC underwent neoadjuvant chemoradiation therapy followed by surgery at the University Hospital of Leuven (UZLeuven). We received planning CT scans (pCT) and corresponding manually generated contours for these 40 patients. To ensure database consistency and quality, all patients were systematically planned for proton beam scanning (PBS), adhering to the treatment planning guidelines outlined in PROTECT (5). The prescribed treatment regimen involved a total dose of 50.4Gy delivered over 28 fractions.

Among the patient cohort, 18 out of 40 individuals underwent a repeated CT scan (rCT) after approximately 3 weeks

of treatment, corresponding to about 15 treatment fractions. New contours were delineated on these rCT scans.

Adaptation rate

NA doses were generated by recomputing the initial plan on the rCT scans. When evaluating this NA dose on the new anatomy, it is also important to address its robustness against treatment delivery errors. Different setup error strategies have been investigated with residual setup uncertainties of 2, 4, and 5 mm, respectively. We may justify the use of these different strategies by assuming that the new image realizes a greater or lesser part of the intra-fraction residual setup error, which is added to an isocenter positioning error or contouring errors. Corresponding adaptation rates are reported.

Benefits of adaptation for EC

Additionally, the no-adaptation strategy (NA dose) was compared with a fully manual approach (MA dose) in terms of dose volume histograms (DVH) that are important in the PROTECT protocol. Robustness evaluation was also conducted with a 7mm setup error for the manual strategy and 2, 4 and 5 mm of residual setup error for the no-adaptation strategy.

OAPT strategies

Four different OAPT strategies were compared with the fully manual approach. These different strategies were implemented on the rCT scans from our 18 patients and are represented in Figure 1.

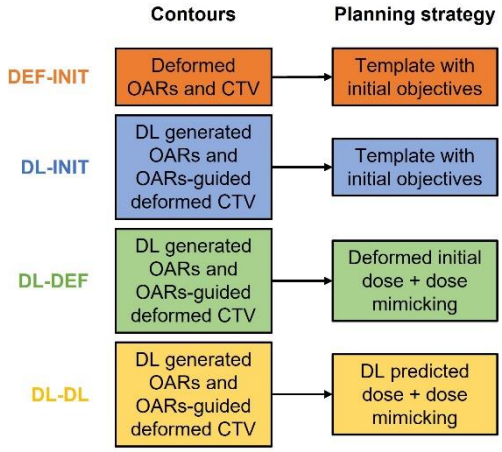


Figure 1: Workflow illustrating the four OAPT strategies.

The strategies were implemented on a Raystation 2023B Software by using scripted actions, apart from deep learning (DL) models, which have been trained and used for inference on GPUs hosted by the *Consortium des Equipements de Calculs Intensifs (CECI).*

a. Contour delineation

In DEF-INIT, contours of OARs and CTV were mapped from the pCT scan to the rCT scan using deformable image registration (DIR).

In DL-INIT, DL-DEF and DL-DL a nnUNet model *(10)* is used for automated OARs contouring. The nnUnet model was trained on our database of 40 EC patients with pCT

scans, using a k-fold cross-validation strategy. Careful selection of models used for inference on the 18 rCTs was conducted, to avoid selecting a model already trained on corresponding patient's pCT. The model predicts contours of 9 OARs per patient in XX seconds on a TeslaA100_80 GPU. Subsequently, DL generated OARs, identified as structures influencing the shape of the iCTV, are used to guide a DIR which later serve for iCTV mapping from the pCT scan to the rCT scan.

b. Planning strategies

In DEF-INIT and DL-INIT, initial optimization objectives, originally specified in the pCT scan, are copied in a newly generated plan in the rCT scan, prior to running an optimization for 200 iterations. These objectives are defined on the deformed and nnUNet-generated contours, respectively.

In DL-DEF, the initial dose, originally defined in the pCT scan, is mapped to the rCT scan using DIR. A dose mimicking strategy is then applied to this deformed dose. It uses isodoses-based and voxelwise least square error optimization. Additionally, conventional weighted objectives on the CTV are added to the optimization problem, before running the optimization process for 60 iterations.

In DL-DL, a HD-Unet *(11)* is cascaded with the previously described nnUnet. This second model is used for inference of dose distribution on the rCTs and corresponding contours previously generated by the nnUnet. The HD-UNet was trained with a k-fold cross-validation on the 40 pCTs and corresponding contours. Again, we performed a careful selection of the model used for inference on the rCTs. More information about the model training can be found in *(11)*. Our HD-UNet predicts a dose distribution for a patient in 3.76 seconds on a TeslaA100_80 GPU. The dose mimicking strategy is applied on the HD-UNet generated dose distribution.

3 Results

Adaptation rate

Figure 2 provides the adaptation rate for each patient and shows the PROTECT criteria for adaptation after 3 weeks of treatment in the nominal and worst-case scenarios. As observed in Figure 2, three patients (17%) require adaptation based on the nominal criteria only. When considering worst-case scenario criteria (V95% iCTV > 95%), two more patients require adaptation with a 5mm residual setup error (28% in total) and only one additional patient for 4mm and 2mm residual setup errors (22%). The

PROTECT criteria on OARs are not shown in this Figure since they did not trigger any adaptations.

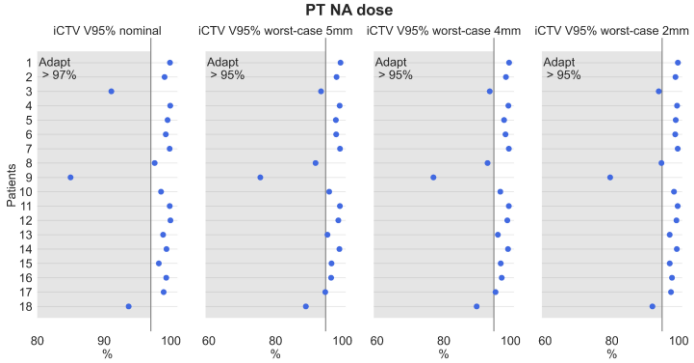


Figure 2: Criteria for adaptation on the CTV in the nominal case and worst-case scenario computed with a 2, 4 or 5mm residual setup errors. The range uncertainty for plan evaluation is 2.6%.

Benefits of adaptation for EC

The second experiment emphasizes how patients would benefit from adaptation. This is depicted in Figure 2. In the nominal scenario, the MA dose tends to have a higher target coverage than the NA dose, both in XT and PT, with an average difference between V95% to iCTV of more than 2%.

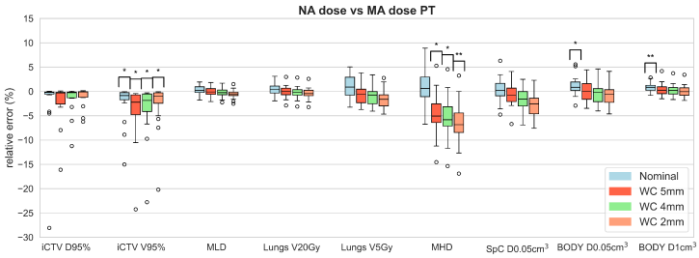


*Figure 3: Comparison between NA dose and MA dose. Boxplots show the error (NA dose - MA dose) relative to the prescription (50.4Gy) or to the structure's volume. Worst-case (WC) scenarios are computed with residual setup errors of 2, 4 or 5mm for the NA dose and a setup error of 7mm for the MA dose. * indicates a statistically significant difference (p = 0.05) between NA dose and MA dose, ** indicates a highly statistically significant difference (p = 0.001) between NA dose and MA dose.*

Moreover, the nominal dose to most organs at risk is on average higher in the NA dose than in the MA dose. However, we can observe a general trend while diminishing the residual setup error with a NA dose having a decreasingly lower dose to OARs in the worst-case scenario. Sometimes, this dose to OARs becomes even lower than the MA dose. For the mean heart dose (MHD), the different worst-case scenarios present a NA dose which is significantly lower than the MA dose (p = 0.05).

Comparison of OAPT Strategies

The different OAPT strategies were compared to the manual approach in terms of D98% to the iCTV. Additionally, timings for contours and plan dose generation are also reported in Table 1. These results were observed on a single patient and are preliminary results of a bigger study integrating our 18 patients with a rCT scan.

In the nominal scenario, all 4 OAPT strategies achieve a D98% which is comparable to the manual approach. It should be noted that the strategy involving dose deformation (DL-DEF) does not reach a D98% of 95% of the prescription (47.88 Gy). Regarding timings, the bottleneck step remains the planning strategy, and the associated time depends on the chosen number of iterations.

	D98%		Timing		
	Nom. (Gy)	*WC (Gy)*	*Contours*	*Planning strategy*	*Total*
Manual	49.07	48.98	Hours	Hours	Hours
DEF-INIT	48.95	47.94	80 s	1055 s	1135 s ~ 19 min
DL-INIT	49.09	48.23	60 s + 70 s	1055 s	1185 s ~ 20 min
DL-DEF	49.04	47.86	60 s + 70 s	360 s	490 s ~ 8 min
DL-DL	49.36	48.43	60 s + 70 s	3.8 x 5 folds + 455 s	585 s ~ 10 min

Table 1: Comparison of the manual adaptation strategy with the proposed OAPT strategies. In DL-INIT, DL-DEF and DL-DL, the time needed for DL OARs segmentation is summed with the time needed for OARs-guided iCTV mapping. In DL-DL, 5 folds are used for dose prediction. The corresponding prediction time is thus summed with the time needed for dose mimicking.

4 Discussion

Adapting a plan during treatment requires significant computational and time resources. Therefore, assessing the need for adaptation is important. This study reports the adaptation rate for different robust evaluation scenarios, with different residual setup errors. The choice of the residual setup error for evaluation must be conducted carefully as this will influence the number of patients requiring adaptation. However, the rationale behind the different options for residual setup uncertainties lies in the fact that intra-fractional motion (other than breathing) is not or at least not totally realized and observed in the newly acquired image. Moreover, unpredicted isocenter shifts and delineation errors need to be accounted for. Our results suggest that robustly optimized PT plans for EC patients may lead to treatments that are often robust to anatomical changes. Indeed, we observe that few patients require adaptation while treated with PT and also that the sensitivity of PT dose to the choice of the residual setup error is relatively low, with only one extra patient being selected for adaptation while varying the residual setup error. However, these results were obtained on a small cohort of 18 patients. Different conclusions could be drawn from a larger batch of patients.

Additionally, results of the second experiment demonstrate that in the nominal scenario, adaptation may lead to a safer target coverage and to lower dose to OARs for EC patients. However, only a few of these improvements reach statistical significance and results for worst-case scenarios are sometimes worse for the MA dose. This might be caused by the fixed margins used for optimization and evaluation of adapted dose in this work. Existing research suggests that ART with reduced margins may enhance organ sparing without compromising target coverage (12), leading to a potential reduction in normal tissue complication probability (NTCP) (13). These findings indicate that, despite evidence for dose distribution improvements, ART with fixed margins may not provide significant clinical benefits and clinical resources might be put to better use.

Preliminary results for OAPT strategies comparison suggest that close to manually optimized D98% can be obtained with all four strategies. However, DL-generated contours associated with initial objectives optimization (DL-INIT) or with DL dose prediction and dose mimicking (DL-DL) exhibits a larger robustness to treatment uncertainties than the other two strategies. It is however necessary to validate this intuition on a larger cohort of patients. This work complements the work of (14) and a broader comparison with other OAPT strategies from the literature could be of interest.

The advantage of the proposed OAPT methods is that they require no manual intervention. All actions were scripted and could be obtained with a single click. In methods integrating a dose mimicking step on a reference dose, fewer iterations were used for optimization, shortening the time needed for the whole process. However, it should be noted that timings for data transfer and conversion between the CECI GPU and RayStation software were not included. It was considered that with a thoughtful end-to-end workflow and collaboration to integrate DL-models in RaySearch software, these steps would not be necessary.

5 Conclusion

To conclude, this work is a step towards fast and automated online plan adaptation and gives us insight about adaptation rates for EC patients. Our results show that the choice of the residual setup error will slightly affect the number of patients requiring adaptation in PT. It serves also as a proof-of-concept for the implementation of 4 fully automated online adaptive proton therapy strategies.

References

1. Sonke JJ, Aznar M, Rasch C. Adaptive Radiotherapy for Anatomical Changes. Semin Radiat Oncol. 2019 Jul 1;29(3):245–57.
2. Møller DS, Lutz CM, Khalil AA, Alber M, Holt MI, Kandi M, et al. Survival benefits for non-small cell lung cancer patients treated with adaptive radiotherapy. Radiother Oncol. 2022 Mar;168:234–40.
3. Qiu Z, Olberg S, den Hertog D, Ajdari A, Bortfeld T, Pursley J. Online adaptive planning methods for intensity-modulated radiotherapy. Phys Med Biol. 2023 May 21;68(10):10TR01.
4. Albertini F, Matter M, Nenoff L, Zhang Y, Lomax A. Online daily adaptive proton therapy. Br J Radiol. 2020 Mar;93(1107):20190594.
5. Ghose S, Holloway L, Lim K, Chan P, Veera J, Vinod SK, et al. A review of segmentation and deformable registration methods applied to adaptive cervical cancer radiation therapy treatment planning. Artif Intell Med. 2015 Jun 1;64(2):75–87.
6. Calmels L, Sibolt P, Åström LM, Serup-Hansen E, Lindberg H, Fromm AL, et al. Evaluation of an automated template-based treatment planning system for radiotherapy of anal, rectal and prostate cancer. Tech Innov Patient Support Radiat Oncol. 2022 Jun;22:30–6.
7. Borderías-Villarroel E, Taasti V, Van Elmpt W, Teruel-Rivas S, Geets X, Sterpin E. Evaluation of the clinical value of automatic online dose restoration for adaptive proton therapy of head and neck cancer. Radiother Oncol. 2022 May;170:190–7.
8. Eyck BM, Van Lanschot JJB, Hulshof MCCM, Van Der Wilk BJ, Shapiro J, Van Hagen P, et al. Ten-Year Outcome of Neoadjuvant Chemoradiotherapy Plus Surgery for Esophageal Cancer: The Randomized Controlled CROSS Trial. J Clin Oncol. 2021 Jun 20;39(18):1995–2004.
9. University of Aarhus. PROton Versus Photon Therapy for Esophageal Cancer - a Trimodality Strategy (PROTECT) A Multicenter International Randomized Phase III Study of Neoadjuvant Proton Versus Photon Chemoradiotherapy in Locally Advanced Esophageal Cancer [Internet]. clinicaltrials.gov; 2022 Nov [cited 2023 May 24]. Report No.: NCT05055648. Available from: https://clinicaltrials.gov/ct2/show/NCT05055648
10. Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat Methods. 2021 Feb;18(2):203–11.
11. Draguet C, Barragán-Montero AM, Vera MC, Thomas M, Populaire P, Defraene G, et al. Automated clinical decision support system with deep learning dose prediction and NTCP models to evaluate treatment complications in patients with esophageal cancer. Radiother Oncol. 2022 Nov;176:101–7.
12. Oud M, Breedveld S, Giżyńska M, Kroesen M, Hutschemaekers S, Habraken S, et al. An online adaptive plan library approach for intensity modulated proton therapy for head and neck cancer. Radiother Oncol. 2022 Nov 1;176:68–75.
13. van de Water S, van Dam I, Schaart DR, Al-Mamgani A, Heijmen BJM, Hoogeman MS. The price of robustness; impact of worst-case optimization on organ-at-risk dose and complication probability in intensity-modulated proton therapy for oropharyngeal cancer patients. Radiother Oncol. 2016 Jul 1;120(1):56–62.
14. Borderias-Villarroel E, Fredriksson A, Cvilic S, Perri DD, Longton E, Pierrard J, et al. Dose mimicking based strategies for online adaptive proton therapy of head and neck cancer. Phys Med Biol. 2023 May 2;68(10):105002.

AI-guided unattended plan generation

Christian Velten[1,2], Michael Bowers[3], William Martin[1], Madhur K. Garg[1], Julie Shade[3], Todd McNutt[3,4], and Wolfgang A. Tomé[1,2]

[1]Department of Radiation Oncology, Montefiore Medical Center, Bronx, NY, USA
[2]Institute for Onco-Physics, Albert Einstein College of Medicine, Bronx, NY, USA
[3]Oncospace, Inc., Baltimore, MD, USA
[4]Department of Radiation Oncology, Johns Hopkins University, Baltimore, MD, USA

Abstract. Dose volume histogram predictions from a custom random forest ensemble model were integrated into an automated heuristic planning engine for unattended plan generation. Optimization objectives are derived from the prediction of the 5% quantile doses at given volumes. Two optimization iterations were performed separated by an intermediate dose calculation. IMRT and VMAT plans were generated in batch mode for fifty previously treated prostate cancer patients and compared to the clinically approved and treated plans. Doses covering large portions of bladder and rectum were substantially lower compared to reference, with 9 Gy and 7 Gy reductions in D80%, respectively. Prescription dose coverage of target volumes was not compromised, and homogeneity was comparable. Combining personalized dose predictions with an automated planning engine has the potential to efficiently generate clinically acceptable patient-specific treatment plans in a standardized way. This would decrease planning staff workloads and have the potential to increase radiotherapy access in low- and middle-income countries, and underserved areas.

1 Introduction

Therapeutic radiology and radiation oncology are technology-driven medical sub-specialties and over the last two decades have become increasingly and pervasively computerized. Advanced radiation therapy (RT) planning and delivery would be infeasible without the use of advanced software for image processing and treatment planning.

Artificial intelligence (AI) driven models for automatic plan generation and dose prediction provide excellent opportunities to increase efficiency and standardization of the treatment planning workflow while increasing the use of individualized medicine [1]. Of those, substantial progress has been made in knowledge-based planning by researchers and vendors alike [2–9]. AI predictions of individual dose/volume metrics [10], dose volume histograms (DVH) [5,7], and entire 3D dose distributions [11] can provide both clinicians and planning staff with varying levels of information on achievable plan metrics. DVH prediction methods, perhaps counterintuitively, provide the most robust results, as they only predict the statistical description of dose distributions given the input data, allowing to capture clinical tradeoffs in the dataset.

Here, we utilized a custom random forest ensemble model (Oncospace Inc., Baltimore, MD) to predict the best achievable individual patient's organ-at-risk (OAR) DVHs as an initialization to a heuristic automated planning engine built using the Varian Eclipse (Varian, Palo Alto, CA) Scripting API (ESAPI). The results were evaluated for both Volumetric Modulated Arc Therapy (VMAT) and static field Intensity Modulated Therapy (IMRT) geometries on two different delivery platforms, a conventional C-arm linac (TDS) and a ring-gantry based linac (RDS).

2 Materials and Methods

Unattended planning (UPG) using predicted best achievable OAR DVHs was performed for fifty patients who previously underwent conventionally fractionated radiotherapy for prostate cancer. The clinically approved and treated plans served as reference for evaluation.

The initialization of plan optimization began with the DVH predictions of best achievable doses. The random forest ensemble prediction model was trained with 1336 patient treatment plans and was constructed to support the full range of treatment protocols for prostate cancer. For this study, the focus was on the initial plan of the whole pelvis PTV given its complexity.

The prediction model utilizes a multi-target to OAR geometry categorization. Input features include 2D and 3D overlap volume histograms between each OAR and target; shape features to characterize target size, complexity, and concavity; and prescriptions to identify the intended targets

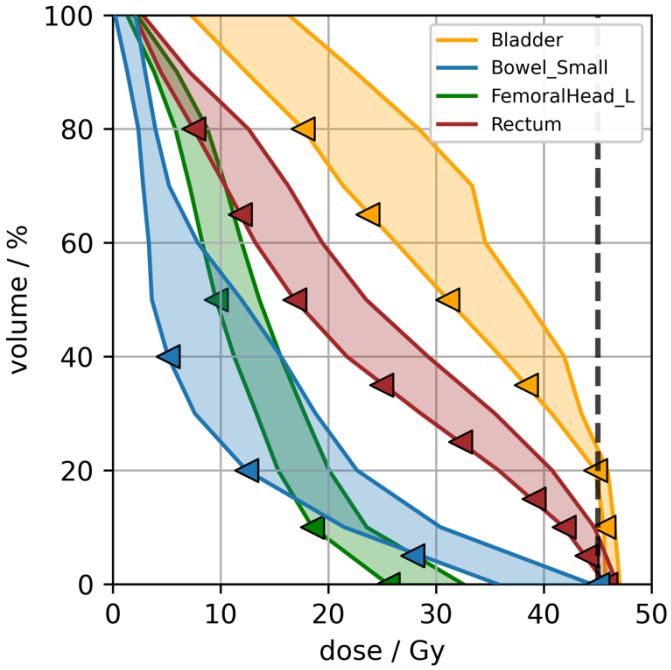


Figure 1. Example model predictions (q5% to mean) of OAR DVHs with associated optimization objectives derived from q5% dose values at a given volume. Structures are (left-to-right): small bowel (blue), femoral head (green), rectum (brown), bladder (yellow).

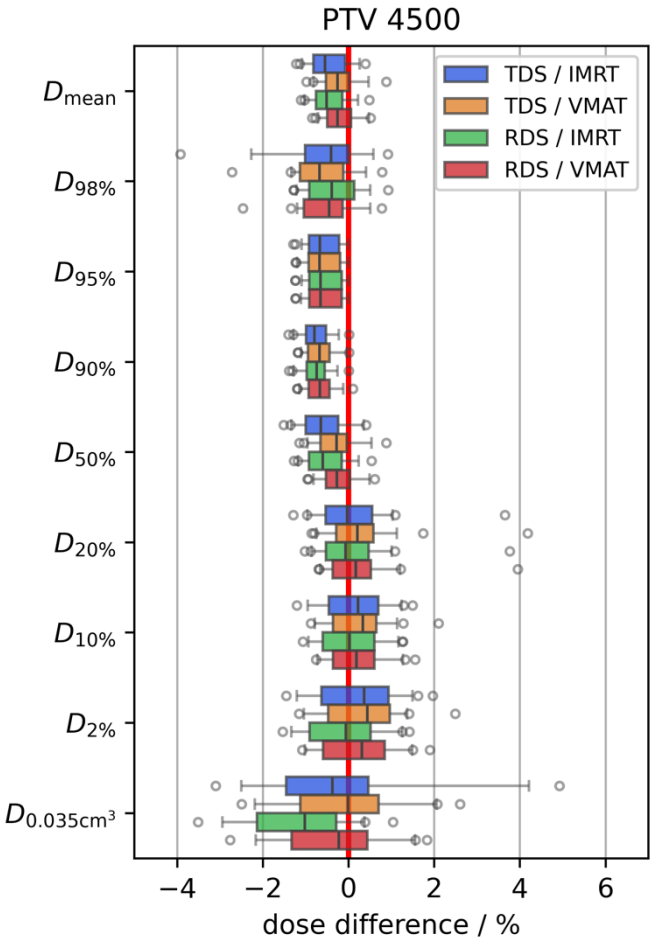


Figure 2. 45 Gy planning target volume relative differences for mean dose, doses covering 0.035 cm^3, and 98% to 2% of the volume. UPG plans were normalized to the clinical minimum goal of $V_{100\%}$=95%, whereas clinical plans often achieved ≥ 95%, leading to a systematic difference of < 1 Gy for $D_{95\%}$. Boxes indicate the interquartile range, while whiskers include 95% of the data points.

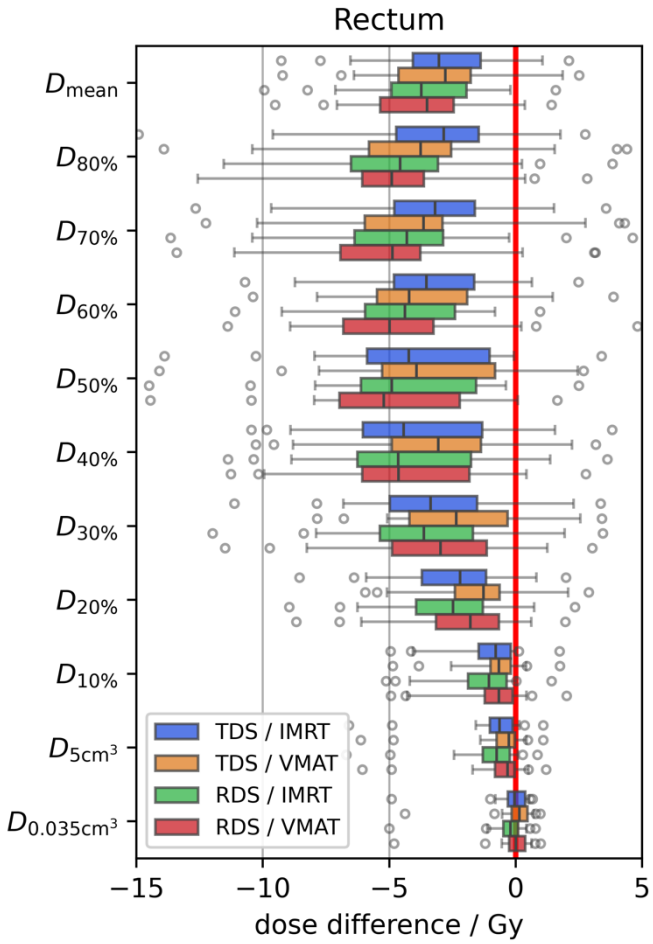


Figure 3. Rectum organ-at-risk differences for mean dose, doses covering 5 cm^3, 0.035 cm^3, and 80% to 10% of the structure.

and relative dosimetry for the plan. The model is trained to predict both the mean and 5% quantile dose-volume metrics to generate a range representing an above average quality plan. Optimization objectives are derived from the prediction of the 5% quantile dose at a given volume value of the OAR's DVH distribution to push the solution towards this best achievable result. Plan doses within the mean to 5% quartile range are identified as meeting the expected doses for a given patient and, thus are considered better than average.

Planning procedures were generalized and implemented in C# using the Varian Eclipse Scripting API (ESAPI) to generate binary plug-ins and standalone executables for batch execution. Plan generation was preceded by structure set duplication and generation of generalized planning structures. Beam arrangements were fifteen equidistant static IMRT fields and two to three full VMAT arcs. IMRT collimator angles were optimized using ESAPI methods, while for VMAT 0º and 90º collimator angles were used with the latter split into overlapping superior/inferior arcs when otherwise exceeding an aperture of 17 cm. The isocenter was defined by the target volume's geometric center.

Optimization and dose calculation were performed using version 16.1 of the Photon Optimizer and AAA algorithm. For VMAT, continuing optimization was restarted at MR level 3 and for RDS VMAT plans convergence mode was enabled (*on*). All plans utilized the normal tissue objectives (NTO) with customized parameters (priority 100, distance from target 0.5 cm, initial dose level 105%, final dose level 40%, fall-off 0.05). Ring objectives were added for each target to avoid high dose spillage for areas of concave target volumes where NTO can struggle to achieve conformality.

To improve dose homogeneity within the target volume, areas exceeding 105% of the prescription dose were outlined after initial optimization and additionally restricted in the second (continued) optimization. Finally, the dose at 6.5 cm (∥)and 1.5 cm (⊥) distance perpendicular from the largest (by volume) target was limited to 50% of the highest prescription dose. All UPG plans were normalized to cover 95% of the target volume with the 45 Gy prescription dose. Batch UPG was executed on a Varian Eclipse TPS v16.1 system in research mode with two distributed calculation servers without GPU. DVH data were subsequently extracted for both reference and UPG plans. Evaluated plan metrics included mean dose and doses covering relative and absolute volumes for both OARs and targets.

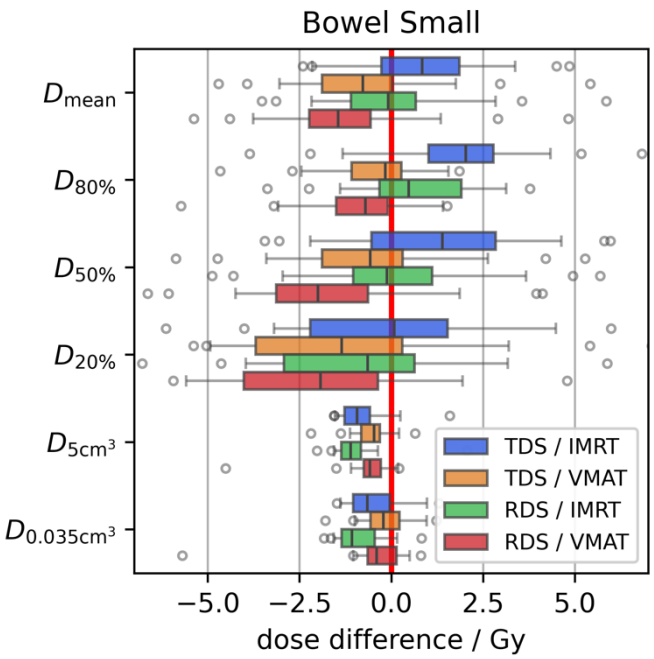


Figure 4. Small bowel organ-at-risk differences for mean dose and doses covering 5 cm^3, 0.035 cm^3, 20%, 50%, and 80% of the structure.

3 Results

Doses covering the 90% to 98% of the 45 Gy target volume (PTV 4500) as well as mean doses were up to 1% lower in most cases (fig. 2). Doses covering smaller volumes (0 to 20%) of the target, often called *hot spots*, were comparable on average deviating less than ±1%, except for $D_{0.035cm^3}$, chosen to be a clinically relevant surrogate for maximum point dose. Subtle differences between the two techniques (IMRT vs. VMAT) were observed for most metrics with IMRT scoring better (negative values) with respect to target mean dose and low volume (≤ 50%) doses. Differences between the two machines (TDS vs. RDS) at the same technique were generally unobservable, except for $D_{2\%}$ and $D_{0.035cm^3}$, with RDS scoring better.

Rectum OAR maximum point doses (fig. 3) were comparable to reference plans with negligible differences, while mean doses were 2.5 Gy to over 5 Gy lower. With increasing volume UPG plans achieved increasingly lower doses covering the OAR. For example, D_{5cm^3} and $D_{10\%}$ had median reductions less than 1 Gy, while median reductions in $D_{20\%}$ were around 2.5 Gy increasing to about 5 Gy for $D_{50\%}$ to $D_{80\%}$. The interquartile ranges for mean dose and D20% to D80% were about 2.5 Gy to 5 Gy and less than 2 Gy for volumes ≤ 10%. Similar results were observed for the bladder and sigmoid colon.

For the small bowel (delineated as bowel bag / potential space) UPG achieved lower doses to small volumes (0.035 cm^3 and 5 cm^3), while doses to larger volumes as well as mean dose varied with respect to treatment machine and technique (fig. 4). UPG VMAT plans yielded overall higher doses, most notably for D_{mean}, $D_{80\%}$, and $D_{50\%}$. Using the same techniques, UPG TDS doses also yielded higher doses.

4 Discussion

This study developed an unattended plan generation workflow utilizing ESAPI and Oncospace DVH predictions for optimization objective generation and applied the workflow to fifty prostate cancer datasets. The results show that plans were both clinically acceptable and comparable to the reference plans. Differences in PTV coverage metrics ($D_{90\%}$ to $D_{98\%}$) are due to most clinical plans achieved prescription dose coverage fractions exceeding 95%, which was the setting for this study. It is important to note that even at a higher UPG coverage goal, it is unlikely to see $D_{20\%}$ to $D_{2\%}$ and $D_{0.035cm^3}$ increase substantially. This is because the optimization setup is built to attempt to compensate for higher doses at these smaller volumes in the second step. A coverage goal of 97% or 98% may thus be feasible for plans like these. For OARs, UPG results were overall more favorable achieving at times significant reductions in doses covering large portions of the OARs: reducing the dose covering 80% of the rectum and bladder by up to 7 Gy (fig. 3) and 9 Gy, respectively. These reductions may help reduce the incidence and/or severity of proctitis and cystitis further [12–14].

While it is possible for dosimetrists and physicists to achieve similar improvements through manual iterative optimization and by utilizing assistive tools, it would require substantially more resources. For example, we found that setting up an individual patient's treatment protocol for prediction required at most 5 minutes, often less. DVH predictions subsequently take less than 5 minutes. Finally, unattended plan generation took 2.5 min to 40 min (without GPU acceleration) depending on the technique used and the number of concurrent calculations. However, since UPG does not require user interaction its actuarial time requirement may be negligible. In the author's experience, manual generation of clinically acceptable plans without AI-guided optimization objectives requires no less than 60 min of undivided attention by the planning staff.

5 Conclusion

Personalized dose-volume objectives were predicted and integrated into an automated heuristic planning solution. Unattended plan generation successfully created plans that were comparable and often superior to clinical plans with respect to OAR sparing. Deployment of this workflow could substantially reduce planning staff workload or allow staff to focus on quality assurance of data and plans. Finally, an automated and unattended workflow as presented here has the potential to increase access to radiotherapy care in low- to middle-income countries and underserved areas.

References

[1] Vandewinckele L et al. Overview of artificial intelligence-based applications in radiotherapy: Recommendations for implementation and quality assurance. Radiotherapy and Oncology 2020; 153: 55–66. doi:10.1016/j.radonc.2020.09.008.

[2] Hussein M et al. Automation in intensity modulated radiotherapy treatment planning-a review of recent innovations. Br J Radiol 2018; 91: 20180270. doi:10.1259/bjr.20180270.

[3] Cornell M et al. Noninferiority Study of Automated Knowledge-Based Planning Versus Human-Driven Optimization Across Multiple Disease Sites. International Journal of Radiation Oncology*Biology*Physics 2020; 106: 430–9. doi:10.1016/j.ijrobp.2019.10.036.

[4] Scaggion A et al. Reducing inter- and intra-planner variability in radiotherapy plan output with a commercial knowledge-based planning solution. Physica Medica 2018; 53: 86–93. doi:10.1016/j.ejmp.2018.08.016.

[5] Tol JP et al. Can knowledge-based DVH predictions be used for automated, individualized quality assurance of radiotherapy treatment plans? Radiation Oncology 2015; 10: 234. doi:10.1186/s13014-015-0542-1.

[6] Wortel G et al. Characterization of automatic treatment planning approaches in radiotherapy. Phys Imaging Radiat Oncol 2021; 19: 60–5. doi:10.1016/j.phro.2021.07.003.

[7] Wu B et al. Cross-institutional knowledge-based planning (KBP) implementation and its performance comparison to Auto-Planning Engine (APE). Radiotherapy and Oncology 2017; 123: 57–62. doi:10.1016/j.radonc.2017.01.012.

[8] Yoo S et al. Clinical Experience With Machine Learning-Based Automated Treatment Planning for Whole Breast Radiation Therapy. Adv Radiat Oncol 2021; 6: 100656. doi:10.1016/j.adro.2021.100656.

[9] Ahmed S et al. A method for *a priori* estimation of best feasible <scp>DVH</scp> for organs-at-risk: Validation for head and neck <scp>VMAT</scp> planning. Med Phys 2017; 44: 5486–97. doi:10.1002/mp.12500.

[10] Brodin NP et al. Organ-at-risk dose prediction using a machine learning algorithm: Clinical validation and treatment planning benefit for lung SBRT. J Appl Clin Med Phys 2022; 23:. doi:10.1002/acm2.13609.

[11] Nguyen D et al. 3D radiotherapy dose prediction on head and neck cancer patients with a hierarchically densely connected U-net deep learning architecture. Phys Med Biol 2019; 64: 065020. doi:10.1088/1361-6560/ab039b.

[12] Emami B et al. Tolerance of normal tissue to therapeutic irradiation. International Journal of Radiation Oncology*Biology*Physics 1991; 21: 109–22. doi:10.1016/0360-3016(91)90171-Y.

[13] Viswanathan AN et al. Radiation Dose–Volume Effects of the Urinary Bladder. International Journal of Radiation Oncology*Biology*Physics 2010; 76: S116–22. doi:10.1016/j.ijrobp.2009.02.090.

[14] Michalski JM et al. Radiation Dose–Volume Effects in Radiation-Induced Rectal Injury. International Journal of Radiation Oncology*Biology*Physics 2010; 76: S123–9. doi:10.1016/j.ijrobp.2009.03.078.

Impact of artificial intelligence solution aiming to accelerate MRI acquisition on the therapeutic care in oncology in radiology and radiotherapy departments

Raphaëlle Lemaire[1], Charlotte Raboutet[2], Thomas Leleu[3], Cyril Jaudet[1], Loise Dessoude[3], Fernand Missohou[3], Yoann Poirier[2], Pierre-Yves Deslandes[4], Alexis Lechervy[5], Joëlle Lacroix[2], Ilyass Moummad[1,6], Stéphane Bardet[7], Juliette Thariat[3], Dinu Stefan[3] and Aurélien Corroyer-Dulmont[1,8]

[1] Medical Physics Department, Centre François Baclesse, 14000 Caen, France
[2] Radiology Department, Centre François Baclesse, 14000 Caen, France
[3] Radiotherapy Department, Centre François Baclesse, 14000 Caen, France
[4] Informatics Department, Centre François Baclesse, 14000 Caen, France
[5] UMR GREYC, Normandie Univ, UNICAEN, ENSICAEN, CNRS, 14000 Caen
[6] IMT Atlantique, Lab-STICC, UMR CNRS 6285, 29238 Brest, France
[7] Nuclear Medicine Department, Centre François Baclesse, 14000 Caen, France
[8] Université de Caen Normandie, CNRS, Normandie Université, ISTCT UMR6030, GIP CYCERON, 14000 Caen, France

Abstract: The objective of this prospective study was to evaluate the impact of commercially available solution (SubtleMR®), which halve the acquisition time on the detectability of brain lesions in radiology and radiotherapy in a clinical routine way.
Methods: T1-MRI of 33 patients with brain metastases were analyzed. Images acquired quickly have a matrix divided by two which halve the acquisition time. The visual quality and lesion detectability of the AI images were evaluated by radiologists and radiation oncologist as well as pixel intensity and lesions size.
Results: Analysis of the lesion detectability shows a specificity of 1 and a sensitivity of 0.92 and 0.77 for radiology and radiotherapy respectively. Undetected-lesions on IA image were with a diameter <4mm and statistically low Gadolinium-enhancement-contrast.
Conclusions: It is possible to half MRI acquisition times by using the commercial algorithm and obtain good specificity and sensitivity for lesions with a diameter greater than 4 mm.

1 Introduction

Magnetic resonance imaging (MRI) is an important medical imaging tool for the diagnosis and treatment of cancer. Despite the presence of more than 800 MRIs on French territory, the average time to obtain an appointment for this type of examination is about 32 days, higher than the waiting time target of 20 days set by the 2014-2019 Cancer Plan. Reducing the acquisition time would increase the availability of MRIs and reduce the time to obtain an appointment. There is therefore a real need to reduce this acquisition time to improve patient care. However, reducing acquisition time is only possible by reducing the acquisition matrix or by image averaging, which reduces image quality [1]. Therefore, solutions are needed to restore image quality after acquisition time reduction.
Artificial intelligence (AI), and in particular deep learning, a subfield of AI, has seen tremendous progress and application in medicine in recent years [2]. NYU Langone and Facebook AI Research collaborated on a research project called "fastMRI", which consists of open datasets and benchmarks for accelerating MRI sequence acquisition. They have focused on brain and knee [3-5], and more recently on prostate data [6]. Other applications of these algorithms exist, ranging from image denoising to improving the rendering of examination images [7] to super-resolution [8]. One super-resolution algorithm, Subtle MR™, has been validated by the Food and Drug Administration for routine clinical use. This solution has reduced acquisition time by 40% in spinal MRI [9] and by 60% in brain MRI in multi-centre and multi-MRI studies [10]. To our knowledge, only one study has evaluated the use of the AI solution in the assessment of brain lesions, using a cohort of 25 patients and achieving a 45% reduction in acquisition time with half the number of phase-encode lines [11]. Despite these findings, no studies have addressed how the associated decrease in resolution due to the shorter acquisition time could potentially affect tumour delineation. In addition, there is no information on the lesion size threshold for detectability when using the AI solution. As this is a resampling problem for the algorithm, there may be a problem with removing or creating some lesions, especially the smaller ones. Therefore, in this prospective study, we evaluated the impact of the AI solution, which halves the acquisition time, on the therapeutic management of small lesions such as brain metastases.

2 Materials and Methods

For this prospective study thirty-three patients with brain metastases (BM) and meningioma (MG) referred to our oncology centre between August 2022 and March 2023 were included. MRI was performed for initial diagnosis or treatment efficacy follow-up. MR-004, a French national institution (INDS) that sets guidelines for the conduct of health research, was used for this study.

MRI was performed on an AREA SIEMENS 1.5 Tesla magnet. Tumour gadolinium enhancement was detected with a post-Gd T1 brain sequence. T2 FLAIR sequences were also used to assess oedema. Each sequence was acquired clinically (reference, longer) and then accelerated (half the time, with two times lower spatial resolution). Images from the accelerated sequences were then processed using the commercial AI solution.

The commercially available deep learning model SubtleMR™, was used for resampling. SubtleMR™ is based on a U-Net deep convolutional neural network backbone that was previously trained on a large number of paired low- and high-resolution images acquired from a variety of vendors, field strengths, and institutions. A new series with the same nominal resolution as the standard of care sequence was generated after applying the AI-enhanced algorithm. This model is used for inference, without any local fine-tuning or adaptation.

Visual analysis: Reference and AI reconstructed images were reviewed by one experienced radiologist and three experienced radiation oncologist who have attributed a global, whole-image quality (IQ) score to each MRI series: 1=poor; 2, 3=moderate; 4=good and 5=very good.

Semi-quantitative analysis:
Radiology: As part of routine image review, one radiologist evaluated the maximum length of each lesion and comparison was made between reference and AI reconstructed images.
Radiotherapy: Following a routine workflow in radiotherapy department, radiation oncologist delimited in 3D volume of interest (VOI) the volume of each lesion. Structural similarity index measure (SSIM) and Jaccard index were used to compare the VOI delimited by the radiation oncologist on reference and AI reconstructed images.

Global analysis: First order intensity evaluation was performed using mean, standard deviation, min, max, skewness, kurtosis, signal to noise ratio and absolute contrast. Specificity and sensitivity were evaluated to control potential radiologist and radiation oncologist failure to detect lesions after AI reconstruction.

Student's t-test was used to compare the different quantitative metrics between reference and AI reconstructed images. All statistical analyses were performed using Python [12] and the SciPy library. All python codes used in the analysis are available at https://github.com/AurelienCD/Impact-of-AI-MRI-solution-on-clinical-routine.

3 Results

Visual analysis:
As presented in Figure 1, AI reconstruction improves some details in the image, with the large lesion (white arrow) easily distinguishable, but the smaller lesion (black arrow) suffering from loss of detail.

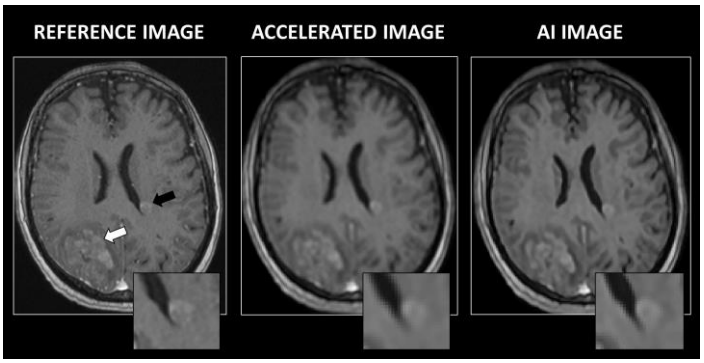


Figure 1: Representative MRI of reference image, accelerated image and AI reconstructed image in brain presenting brain metastases

Radiologist IQ scores are lower for AI images compared to reference images (mean score 4.79 ± 0.43 vs. 1.86 ± 0.77, ***p<0.001 for reference and AI images, respectively). Radiation oncologist IQ were similar.

Semi-quantitative analysis:
First order signal intensity analysis of the whole image was performed and showed a decrease in mean pixel value in accelerated and AI reconstructed images compared to reference images (p<0.001). No difference was found for minimum, maximum and skewness values. In order to assess the ability of the AI solution to restore details lost in the accelerated images, difference maps between accelerated and reference and AI and reference were computed. The subtraction map between the reference and accelerated images shows a 60% difference in pixel values in the contrast-enhanced area. Interestingly, the AI solution was able to restore the shape of the lesion with pixel value differences of less than 30%.
SSIM was then evaluated to compare the similarity of the accelerated and AI images to the reference. SSIM was significantly higher in the AI images compared to the accelerated images (SSIM 0.5921 ± 0.0577 vs. 0.5878 ± 0.0588, **p<0.01 for AI and accelerated images, respectively).
Impact of AI solution on radiology department:
Lesion max diameters delimited by the radiologist were similarly drawn in the reference and in the AI images. No statistical difference were observed in lesion max diameter. However, two lesions were not detected by the radiologist in the AI images. This result leads to a specificity of 1 and a sensitivity of 0.92 in comparison to reference images. However, if we analyse the lesions missed by the radiologist in the AI images, it appeared that these lesions are the smaller in dimension with max lengths of three and four millimetres respectively.

Impact of AI solution on radiotherapy department:
No significant differences were observed between volumes of interest drawn by the radiation oncologist on the reference and the AI images. However, five lesions were missed by the radiation oncologist in the AI images in comparison to the reference images. Specificity and

sensitivity were respectively 1 and 0.77. As for the radiology part, lesions missed were smaller in dimension, with volumes below 0.12 cm^3. It can also be noted that these lesions have a significant lower contrast enhancement in comparison to the other lesions (absolute contrast 0.007 ± 0.0028 vs 0.0154 ± 0.0074 for AI and reference images respectively, $p<0.05$).

4 Discussion

Excessive delays in MRI acquisition result in long waiting times for patients who urgently need medical intervention. Many research groups, MRI manufacturers or even start-ups are developing solutions to enable faster acquisition without losing the diagnostic quality of the images. In our study, we analysed one of these solutions, SubtleMR™. Specifically, we evaluated whether this solution is able to reconstruct an image acquired two times faster than clinical routine acquisition time, without significant loss of image quality, as required for routine clinical management of these patients.

Regarding visual analysis, Bash and colleagues [9] showed similar image quality for reference and AI reconstructed images. In our study, radiologists and radiation oncologists found lower image quality for AI images. This difference with the literature may be due to the fact that the matrix is only divided by two on the phase encoding axis, whereas we chose to divide it on the x and y axes to achieve a significant reduction in acquisition time. However, with respect to the cited study, we show that the AI solution is able to recover image details lost during the accelerated acquisition process.

Regarding the semi-quantitative analysis, in literature studies [7], the signal intensity values collected on the AI image are generally closer to those of the reference image than to those of the accelerated image. This phenomenon is also present in our study. This observation was also true for standard deviation, kurtosis and signal-to-noise ratio. As for the study by Jin and colleagues [8], our results show that SSIM with reference images is higher in AI images compared to accelerated images.

Regarding the effect of the AI solution on lesion detectability and delineation, the maximum diameters and three-dimensional delineations by the radiologist and radiation oncologist, respectively, were not significantly different using reference or AI images. This conclusion is in line with the literature showing that AI solutions need to be improved but are sufficient for safe use in clinical routine [2]. As we have shown in a previous study using an AI solution with PET imaging [13]), missed lesions correspond only to small lesions with low radiotracer uptake, which will have a limited impact on clinical therapeutic care and patient outcome.

Finally, it is important to compare the performance of an AI solution applied to final MRI images with an AI algorithm developed using a raw MRI signal as K-space. In the literature, several articles use K-space images as input with better results because more information can be found in this image than in the final MRI image [14]. However, in a clinical situation, it is often not possible to routinely access k-space, except for directly embedded AI solutions from the manufacturer.

5 Conclusion

In a busy clinical MRI department, the AI solution was used to increase patient throughput without strong significant loss of resolution. This was demonstrated in the comparable assessment of brain lesions > 4 mm in diameter by both oncologists and radiologists. Future multicentre studies including other MRI sequences and tumour sites are warranted.

References

[1] R. Obuchowicz, A. Piórkowski, A. Urbanik et al., Influence of Acquisition Time on MR Image Quality Estimated with Nonparametric Measures Based on Texture Features., Biomed Res Int. 2019 (2019) 3706581. https://doi.org/10.1155/2019/3706581.

[2] P.M. Johnson, M.P. Recht, F. Knoll, Improving the Speed of MRI with Artificial Intelligence, Semin Musculoskelet Radiol. 24 (2020) 12–20. https://doi.org/10.1055/s-0039-3400265.

[3] J. Zbontar, F. Knoll, A. Sriram, et al., fastMRI: An Open Dataset and Benchmarks for Accelerated MRI, (2018) 1–35. http://arxiv.org/abs/1811.08839.

[4] F. Knoll, J. Zbontar, A. Sriram, M.J. et al., fastMRI: A Publicly Available Raw k-Space and DICOM Dataset of Knee Images for Accelerated MR Image Reconstruction Using Machine Learning, Radiol Artif Intell. 2 (2020) e190007. https://doi.org/10.1148/ryai.2020190007.

[5] M.J. Muckley, B. Riemenschneider, A. Radmanesh, et al., Results of the 2020 fastMRI Challenge for Machine Learning MR Image Reconstruction., IEEE Trans Med Imaging. 40 (2021) 2306–2317. https://doi.org/10.1109/TMI.2021.3075856.

[6] R. Tibrewala, T. Dutt, A. Tong, L. et al., FastMRI Prostate: A Publicly Available, Biparametric MRI Dataset to Advance Machine Learning for Prostate Cancer Imaging., ArXiv. (2023).

[7] I. Moummad, C. Jaudet, A. Lechervy, S. et al., The Impact of Resampling and Denoising Deep Learning Algorithms on Radiomics in Brain Metastases MRI, Cancers (Basel). 14 (2021) 36. https://doi.org/10.3390/cancers14010036.

[8] Z. Jin, Q.-S. Xiang, Improving accelerated MRI by deep learning with sparsified complex data., Magn Reson Med. 89 (2023) 1825–1838. https://doi.org/10.1002/mrm.29556.

[9] S. Bash, B. Johnson, W. Gibbs, et al., Deep Learning Image Processing Enables 40% Faster Spinal MR Scans Which Match or Exceed Quality of Standard of Care : A Prospective Multicenter Multireader Study., Clin Neuroradiol. 32 (2022) 197–203. https://doi.org/10.1007/s00062-021-01121-2.

[10] S. Bash, L. Wang, C. Airriess, et al., Deep Learning Enables 60% Accelerated Volumetric Brain MRI While Preserving Quantitative Performance:

A Prospective, Multicenter, Multireader Trial., AJNR Am J Neuroradiol. 42 (2021) 2130–2137. https://doi.org/10.3174/ajnr.A7358.

[11] J.D. Rudie, T. Gleason, M.J. Barkovich, D.M. et al., Clinical Assessment of Deep Learning-based Super-Resolution for 3D Volumetric Brain MRI., Radiol Artif Intell. 4 (2022) e210059. https://doi.org/10.1148/ryai.210059.

[12] Anaconda Software Distribution, (2020). https://docs.anaconda.com.

[13] K. Weyts, C. Lasnon, R. Ciappuccini, et al., Artificial intelligence-based PET denoising could allow a two-fold reduction in [^{18}F]FDG PET acquisition time in digital PET/CT., Eur J Nucl Med Mol Imaging. 49 (2022) 3750–3760. https://doi.org/10.1007/s00259-022-05800-1.

[14] D.J. Lin, P.M. Johnson, F. Knoll, Artificial Intelligence for MR Image Reconstruction: An Overview for Clinicians., J Magn Reson Imaging. 53 (2021) 1015–1028. https://doi.org/10.1002/jmri.27078.

The Automatic Body Composition (ABC) Toolkit

Dónal McSweeney[1,2], Craig Jones[1,3], Michael D. Brown[1], Ashwin Sachdeva[1,3], Noel Clarke[3], Marcel van Herk[1,2], and Alan McWilliam[1,2]

[1]Division of Cancer Sciences, University of Manchester, Manchester, M13 9PL, UK
[2]Radiotherapy Related Research, The Christie NHS Foundation Trust, Manchester, M20 4BX, UK
[3]Department of Surgery, Christie Hospital NHS Foundation Trust, Manchester, M20 4BX. UK

Abstract Body composition, muscle mass in particular, is an established prognostic factor in oncology as it provides complementary information to performance status about patient fitness. Manual assessment is time-consuming so current practice does not use information about body composition, even though it could be extracted from routine CT or MRI scans. Instead, clinical practice still relies on broad and subjective criteria, like performance status and BMI, for patient stratification and monitoring. To facilitate large-scale research and promote (eventual) integration into routine clinical care we have developed the Automatic Body Composition (ABC) Toolkit. ABC was designed with flexibility in mind; it provides a browser-based front end for use by those without programming experience and exposes a set of APIs for simple integration into research and clinical workflows. It includes AI-based segmentation and efficient quality assurance. ABC is currently available for collaborative projects upon request.

1 Introduction

Skeletal muscle plays a key role in regulating several physiological systems and is therefore vital for responding to stressors (e.g., cancer and its treatments) [1]. Sarcopenia, the loss of muscle mass and function due to ageing, is a central component of frailty, and can be compounded by cancer driven muscle loss through cachexia. Sarcopenic cancer patients are known to have poorer outcomes. Across disease sites and treatment modalities, low muscle mass is associated with both decreased survival and increased toxicities [2]. This effect is not just limited to cancer patients: burn victims and patients in intensive care presenting with low muscle mass also experience worse survival and increased recovery times [3]. However, as imaging biomarkers of muscle health are time-consuming to measure, this information is not yet used in clinical practice. Instead, clinicians rely on broad and subjective criteria, like BMI and performance status, to inform treatment decisions.

A common approach to muscle mass estimation involves segmentation of the muscle compartment on a CT scan. Typically, this is performed on a single slice at the L3 vertebral level as this region has been shown to be highly correlated with total body muscle mass [4]. However, even finding and delineating a single slice is a time-consuming process and reliance on L3 excludes all scans where this level is not imaged. Radiotherapy planning and image-guided radiotherapy offers a wealth of imaging data that could be used to measure, monitor, and intervene based on body composition and its dynamic behaviour. However, current approaches to measuring body composition make this unfeasible.

Automatic muscle segmentation is common in literature, but few methods are publicly available. Comp2Comp from Stanford's Machine Intelligence in Medical Imaging lab is an open-source toolkit for measuring body composition from CT imaging [5]. It enables measurements of aortic calcifications, vertebral body density estimation (between T12-L5), muscle and fat quantification, as well as estimates of liver, spleen and pancreas volume and density. Similarly, the Opportunistic Screening Consortium in Abdominal Radiology (OSCAR) developed a toolkit to extract the same features with the omission of the pancreas and the addition of the kidneys [6]. Although, both Comp2Comp and OSCAR provide detailed insights into body composition, they are restricted to CT scans of the abdomen, which limits their use in radiotherapy where imaging is generally only available around the target volume. Although relatively simple to use, both lack a graphical interface and are aimed at researchers with some programming experience.

A commercial alternative is the Data Analysis Facilitation Suite (DAFS) by Voronoi Health Analytics. It provides a wide range of models for full-body tissue and organ segmentation and can be applied to scans with varying fields of view. DAFS is a powerful tool aimed at facilitating large-scale retrospective analyses. It can provide detailed insight into patient anatomy and can be applied to a wider range of scans than Comp2Comp or the OSCAR toolkit. However, it has its limitations. DAFS does not save model predictions, so segmentations cannot be corrected, and further (e.g. radiomic) analyses cannot be performed. Reliance on volumetric segmentation of tissues and organs increases execution time and computational and data transfer requirements. In terms of body composition, there is no evidence to suggest a benefit from using volumetric measurements compared to single-slice measurements [7]. DAFS can only be operated via its graphical interface and data must be organised in a specific manner, which limits integration with research and clinical workflows. Finally, its cost, which includes a per-scan fee, hinders analyses of large-scale, longitudinal data - one of the remaining gaps in the field.

We present the Automatic Body Composition (ABC) toolkit, a tool for fast and accurate measurement of skeletal muscle and fat from routinely collected CT imaging (including cone-beam CTs of the head and neck). The toolkit can be applied

to scans of varying fields of view, with segmentation models extending from neck to thigh. It provides a user-friendly interface for those without programming experience, as well as an API for simple integration with scripting languages. Built-in DICOM communication also enables integration with research databases and clinical picture archiving and communication systems (PACS).

2 Materials and Methods

2.1 System Architecture

A schematic of the ABC Toolkit is shown in Figure 1. The toolkit itself is composed of several Docker (v24.0.7) containers managed by Docker Compose (v2.21.0). The task controller exposes a set of APIs for executing specific pipelines. Processing is performed by separate worker nodes such that the task controller can handle asynchronous requests. This feature is especially useful when performing large-scale batch analyses, as it does not block incoming requests. Deep-learning inference is composed of two main tasks: spine labelling, which consists of automatically labelling all vertebrae in the CT scan; and tissue segmentation, which automatically segments the muscle and fat compartments at the desired level(s) as identified by spine labelling. Estimates of muscle/fat area and density are then calculated and pushed to the database.

The web server communicates with the task controller to display web pages for simplified processing. Importantly, it exposes a quality assurance (QA) page to quickly check model predictions. The user identifies failed segmentations and can leave a note explaining their decision. These notes can then be used to highlight consistent errors that need to be addressed.

A core design choice was to make the toolkit flexible and amenable to different use cases, some of which are shown in Figure 1. By exposing a set of APIs, standardised HTTP methods can be used to make requests and incorporate different elements of the toolkit into other workflows (scripting mode). For example, the spine labelling pipeline can be executed alone, and the results used to initialise a separate registration pipeline. For large-scale analyses, we have built-in communication with XNAT, an open-source imaging platform that is widely used for hosting, organising and sharing imaging data. The ABC toolkit can display available XNAT projects and batch process all the CT scans in a selected project (batch mode). Finally, communication with clinical DICOM interfaces (e.g. Conquest [REF]) enables real-time inference when the interface posts a trigger to the toolkit (real-time mode). The trigger could be the result of an incoming C-STORE, from the clinical PACS for example, with the resulting report pushed back into the PACS.

2.2 Deep-learning

We have chosen to omit many of the technical details for conciseness but thought it important to describe the implementation of the deep-learning pipelines. Spine labelling is based on the method first developed by Payer et al. for the VerSe Challenge at MICCAI 2020 [8]. This is a multi-stage approach in which a 3D convolutional neural network first locates the spine on the full CT scan. The prediction is then used to crop the image around the spine and a second model labels the vertebrae. These steps are computationally expensive since the 3D CT volume is processed, as a result a GPU is needed for inference. We aim to decrease computational complexity in future to facilitate deployment and improve inference speed.

The segmentation pipeline consists of several 2D convolutional networks optimised for different anatomical locations. Currently, these include models for C3, T4, T9, T12, L3, L5 and the thigh. U-Net architectures with MixNet encoders are used as they are lightweight and accurate [9, 10]. Following training, the models are compiled using the Open Neural Network Exchange (ONNX) to reduce memory footprint and improve inference speed . As a result, fast inference can be performed on CPU alone [11]. In addition, reducing the parameter count allows models to be trained with fewer training examples, which accelerates development of new models. We have previously shown that there are minimal benefits to training on more than 25 examples, and in practice models were trained on approximately 40 cases [12]. We would like to extend our models to additional anatomical sites, but data collection is time-consuming.

3 Results

3.1 Benchmarking

The following tests were performed with an AMD Ryzen Threadripper 2950X 16-Core processor running Ubuntu 22.04. An NVIDIA GeForce RTX 2080 Ti with 12 Gb of VRAM was available but only used for spine labelling. Table 1 shows the timings of each step in the processing pipeline. This was performed on a single CT thorax, abdomen and pelvis (large field of view). Total processing time was 58.07 seconds. A formal analysis has not yet been performed to measure average QA time but, from experience analysing thousands of predictions, we estimate this to be between 1-5 seconds per scan. There remain improvements to be made, notably in terms of vertebral labelling, but these results show feasibility of use in large-scale analyses as well as time-sensitive settings.

3.2 Validation

The ABC toolkit was then validated against both Comp2Comp and the OSCAR toolkit, as they are the openly

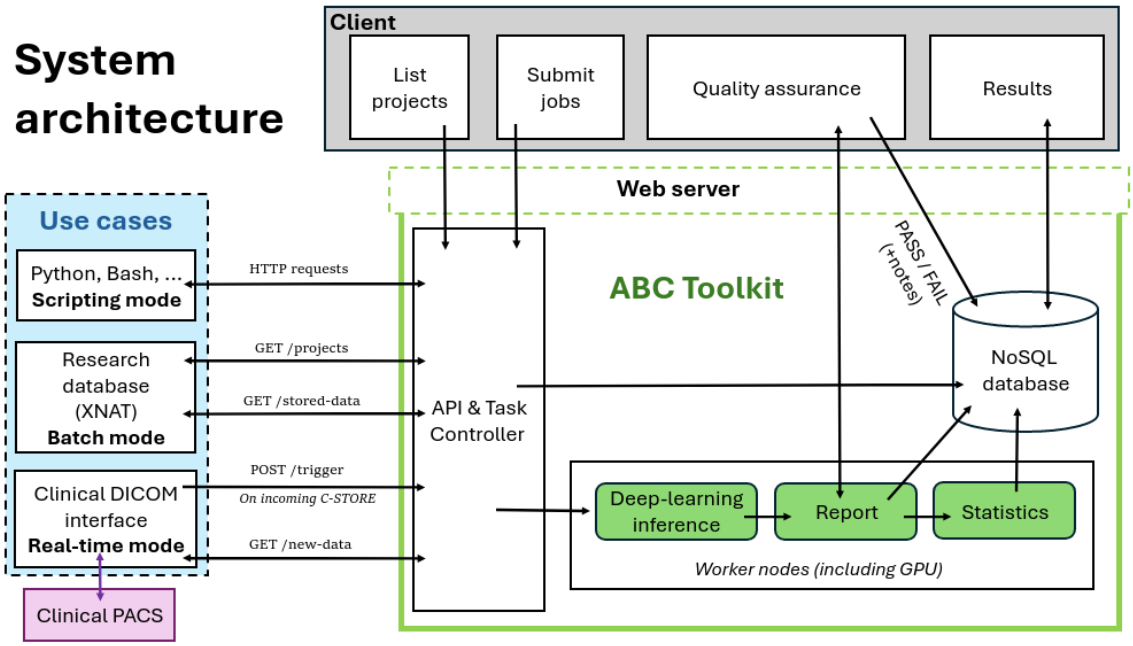


Figure 1: Schematic overview of the ABC Toolkit. The task controller exposes a set of APIs which makes the tool amenable to a variety of use cases. A web server enables use without programming experience and presents a QA page to quickly check model predictions.

Table 1: Table showing model size and processing times for each step in the pipeline. In total, the pipeline takes 58.07 seconds per example. The example used was a CT thorax, abdomen and pelvis of a man treated for prostate cancer. Note post-processing time includes writing segmentation outputs, which is the most time-consuming element of the segmentation pipeline.

	Model Size (Mb)	**Pre-processing** (s)	**Inference** (s)	**Post-processing** (s)	**Total** (s)
Spine locating	72	9.36	3.18	0.02	12.56
Vertebrae labelling	72	0.45	22.58	13.71	36.74
Muscle/fat segmentation	20	1.80	0.37	6.6	8.77

available alternatives. Note that this comparison was performed at the L3 vertebral level since this is the reference site used throughout the literature, and both Comp2Comp and OSCAR focus on analysing abdominal CTs.

First, segmentation accuracy was evaluated against manual delineations of the muscle compartment at the L3 vertebral level. Contouring was performed at two timepoints on 21 cases by a single trained annotator (CJ) with 11 years of experience. Note that CJ was not involved in model development, models were not re-trained following this comparison. Contouring was performed across three slices spanning the L3 vertebra. The 95% Haussdorf distance and mean distance-to-agreement were used to assess accuracy. Importantly, both Comp2Comp and the ABC toolkit separate skeletal muscle from intramuscular adipose tissue, whereas the human annotator did not make this distinction. For clarity, here, we compared total skeletal muscle area or, the union of the skeletal muscle mask and the intramuscular adipose tissue mask. Results are shown in Figure 3. ABC consistently outperforms both Comp2Comp and OSCAR when compared to manual delineations.

4 Discussion

We present the Automatic Body Composition (ABC) toolkit: a flexible, accurate and easy-to-use tool for the estimation of muscle and fat mass from routinely collected CT imaging. Importantly, ABC was designed with radiotherapy in mind and is therefore applicable to scans of varying fields of view including cone-beam CTs. However, we have only trained a head and neck CBCT model so far. A browser-based frontend supports fast QA and allows users with no programming experience to run analyses, while a set of APIs enable simple integration with other research and clinical workflows.
Comp2Comp and the OSCAR toolkit are comparable tools, but both are restricted to analysing abdominal CTs and are therefore unsuited to many applications in radiotherapy. Figure 3 shows that the ABC toolkit outperforms both Comp2Comp and OSCAR in terms of segmentation accuracy. Regardless, these tools extract additional measures of body composition including bone mineral density and estimates of aortic calcifications that are not yet implemented in ABC. A commercial alternative is DAFS by Voronoi Health Analytics. DAFS is aimed at facilitating large-scale retrospective analyses by enabling full-body tissue and organ segmentation. As a result, it is not easily integrable into other pipelines. Moreover, model predictions are not saved so external validation

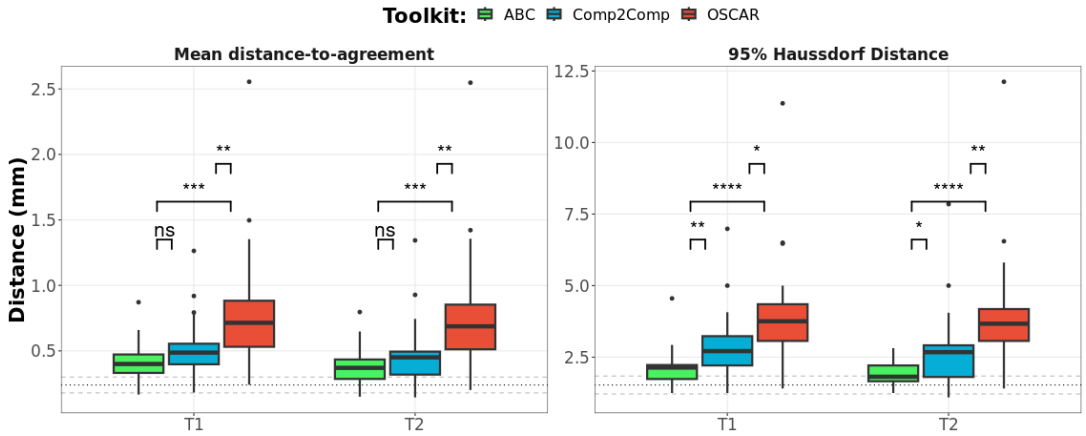


Figure 2: Validation of the ABC Toolkit against openly available toolkits. Models were compared to manual annotations performed at two time-points (T1 & T2). Dotted lines represent mean and standard deviations of the intra-observer agreement. ABC consistently outperformed both Comp2Comp and OSCAR.

cannot be performed.

The ABC toolkit is currently limited in that it only segments fat compartments in the abdomen. At all levels, the muscle compartment is segmented alongside a body mask that provides an estimate of total body size. In future, we aim to integrate more models with the capacity to segment both fat and muscle. Moreover, an aim of the ABC toolkit was to be lightweight and fast. The vertebral labelling module lacks in this aspect as it is still relatively slow (Table 1) and still requires a GPU for inference. Future work will seek to optimise this aspect to ensure the toolkit can be easily deployed on a range of hardware.

The ABC toolkit presents a step towards implementing muscle mass assessment in radiotherapy clinics. By developing a tool that is amenable to both large-scale research and on-demand requests from clinical PACS we hope to promote further research in the field and facilitate the use of body composition in patient stratification, patient monitoring and the development of targeted interventions.

5 Conclusion

We developed a toolkit for automatically measuring body composition from routinely collected CT scans. The ABC toolkit's unique selling point is flexibility. It is set up as a service whereby it listens for incoming processing requests and returns the results to the requester. ABC is suited to both large-scale analyses and real-time processing, which facilitates integration into both research and clinical workflows. This proof-of-concept presents a potential road towards clinical implementation.

References

[1] W. R. Frontera and J. Ochala. "Skeletal muscle: a brief review of structure and function". *Calcified Tissue International* 96 (2015), pp. 183–195.

[2] S. S. Shachar, G. R. Williams, H. B. Muss, et al. "Prognostic value of sarcopenia in adults with solid tumours: A meta-analysis and systematic review". *European Journal of Cancer* 57 (2016), pp. 58–67.

[3] C. T. Pereira, R. E. Barrow, A. M. Sterns, et al. "Age-dependent differences in survival after severe burns: a unicentric review of 1,674 patients and 179 autopsies over 15 years". *Journal of the American College of Surgeons* 202.3 (2006), pp. 536–548.

[4] W. Shen, M. Punyanitya, Z. Wang, et al. "Total body skeletal muscle and adipose tissue volumes: estimation from a single abdominal cross-sectional image". *Journal of Applied Physiology* 97.6 (2004), pp. 2333–2338.

[5] L. Blankemeier, A. Desai, J. M. Z. Chaves, et al. "Comp2comp: Open-source body composition assessment on computed tomography". *arXiv preprint arXiv:2302.06568* (2023).

[6] P. J. Pickhardt, R. M. Summers, and J. W. Garrett. "Automated CT-based body composition analysis: a golden opportunity". *Korean Journal of Radiology* 22.12 (2021), p. 1934.

[7] I. Anyene, B. Caan, G. R. Williams, et al. "Body composition from single versus multi-slice abdominal computed tomography: Concordance and associations with colorectal cancer survival". *Journal of Cachexia, Sarcopenia and Muscle* 13.6 (2022), pp. 2974–2984.

[8] A. Sekuboyina, M. Rempfler, A. Valentinitsch, et al. "Labeling vertebrae with two-dimensional reformations of multidetector CT images: an adversarial approach for incorporating prior knowledge of spine anatomy". *Radiology: Artificial Intelligence* 2.2 (2020), e190074.

[9] O. Ronneberger, P. Fischer, and T. Brox. "U-Net: Convolutional Networks for Biomedical Image Segmentation". *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Vol. 9351. Springer. Springer, 2015, pp. 234–241.

[10] M. Tan and Q. V. Le. "Mixconv: Mixed depthwise convolutional kernels". *arXiv preprint arXiv:1907.09595* (2019).

[11] D. M. McSweeney, E. G. Henderson, and A. F. Green. "COBRA: CPU-Only aBdominal oRgan segmentAtion". *arXiv preprint arXiv:2207.10446* (2022).

[12] D. M. McSweeney, E. G. Henderson, M. Van Herk, et al. "Transfer learning for data-efficient abdominal muscle segmentation with convolutional neural networks". *Medical Physics* 49.5 (2022), pp. 3107–3120.

OSIRIS-RT : a minimum data set for data sharing and interoperability in radiation oncology

Vincent Le Texier[1], Thomas Baudier[2], Sébastien Guihard[3], Thomas Lacornerie[4],

David Pasquier[5], Charlotte Robert[6]

[1]Synergie Lyon Cancer, Platform of Bioinformatics Gilles Thomas, Centre Léon Bérard, Lyon, France
[2]Université de Lyon; CREATIS; CNRS UMR5220; Inserm U1294; INSA-Lyon; Université Lyon 1, Léon Bérard cancer center, Lyon, France
[3]Department of Radiotherapy, Institut de Cancérologie Strasbourg Europe, Strasbourg, France
[4]Centre Oscar Lambret, Service de Physique Médicale, 59000 Lille, France
[5]Département de radiothérapie, centre Oscar-Lambret, Lille, France; Cristal UMR 9189, université de Lille, école Centrale de Lille, CNRS, Lille, France
[6]Université Paris-Saclay, Institut Gustave Roussy, Inserm, Radiothérapie Moléculaire et Innovation Thérapeutique, Villejuif, France

Abstract

Purpose
Radiation oncology is a highly specialized field requiring an accurate and clinically relevant treatment summary. There is a clear need for standardization of radiation therapy data to facilitate care coordination.

Methods
For a year, the OSIRIS (Interoperability and data sharing of clinical and biological data in oncology) group based in France has worked on structuring radiotherapy data and identifying technical solutions for collecting and sharing them. The group used a multidisciplinary approach that included weekly scientific and technical meetings to foster a national consensus on a minimal data set.

Results
The resulting set and event-based data model, which is able to capture radiation therapy treatments, was built with 32 clinical and 60 imaging items. The group made it compatible with the HL7 Fast Healthcare Interoperability Resources (FHIR) format to maximize interoperability. The OSIRIS set was reviewed, approved by a National Plan Strategic Committee, and freely released to the community. A proof-of-concept study is underway in France to put the OSIRIS set and Common Data Model into practice using a national and multi-centric cohort of 10.000 patients.

Conclusion
Using a national and bottom-up approach, the OSIRIS group has defined a model including a minimal set of radiation therapy data that can be used to accelerate data sharing. The model relies on clear and formally defined terminologies and, as such, may also benefit the larger radiotherapy community.

1 Introduction

The lack of cohesion, coordination, and structured data sharing between Radiation Oncology Information Systems (ROIS) and other health information systems is due to the lack of standardized data in the radiation oncology domain and absence of Fast Healthcare Interoperability Resources (FHIR) standards to support data transmission.

Recent national[1] and international[2,3] (e.g. American Society for Radiation Oncology) initiatives to standardize radiotherapy data testify to a growing interest in collecting consistent and comparable data in this field. Identifying a minimum set of data for radiotherapy will help to improve care coordination, reduce data entry and provide a more comprehensive view of the holistic approach required for cancer treatment.

The definition of a minimum data elements for radiation oncology is currently underway with the CodeX project[4] with the participation of the American Association of Physicists in Medicine (AAPM) and major software publishers.

The OSIRIS group has taken over part of the work of the CodeX project to integrate an imaging and radiotherapy module into its data model, which already includes clinical and genomic data.[5]

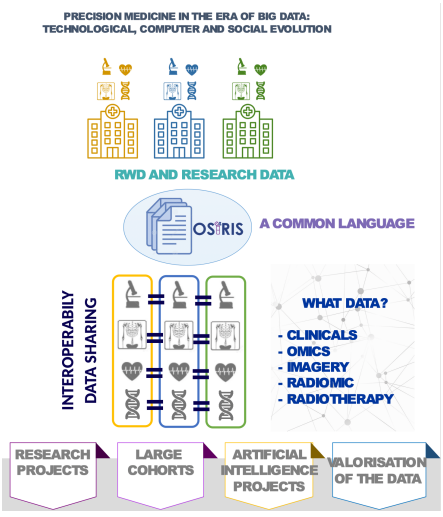


Figure 1 : OSIRIS : French National Initiative on interoperability and data sharing of health data

2 Materials and Methods

The OSIRIS group has proposed to break down the extension of the OSIRIS set into three stages :
1. Determine the important concepts in radiotherapy and how to relate them to the OSIRIS clinical data model,

2. Determine a minimum set of attributes (CDEs, Common Data Elements) by comparing national and international initiatives in the field,
3. Associate a list of possible values (CLs, Code Lists) for each CDE using international terminologies. For example, we could use web portals reflecting biomedical ontologies (such as BIOPORTAL[6] from NCBO-National Center for Biomedical Ontology or UMLS-Unified Medical Language System - NIH National Library of Medicine).

To define a minimum set at national level, our approach will be based on the following steps: (i) identification of the most relevant CDEs based on local partner initiatives; (ii) standardization of the list of possible CDE values using appropriate international terminologies where possible.
The RadioTransNet network (Réseau National de Recherche Préclinique en Radiothérapie)[7] and the Unitrad group[8] (Groupe de recherche en radiothérapie oncologique) have joined the OSIRIS group to define a minimum set of standardized data for radiotherapy.

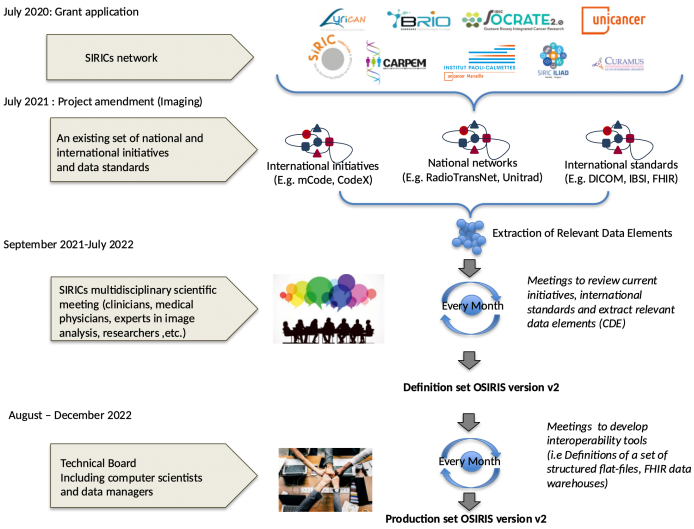


Figure 2 : Overall methodology used between July 2021 and December 2022 to deliver a minimum Common Data Elements (CDEs) for radiation oncology.

3 Results

OSIRIS RT-CDM

A first version of the RT-CDM (v1) was published in december 2022 and freely available to the community on GitHub[9]. This version formed the basis for discussion between the integrated cancer research sites in France (SIRICs), the Unitrad and RadioTransNet groups, and, later, for technical council meetings. Several institutions tested the model (i.e. level of completeness and inconsistency rate), and the group proposed and implemented substantial revisions through functional analysis and data integration feedback. The result is a reliable, robust and comprehensive model for describing radiotherapy treatments.

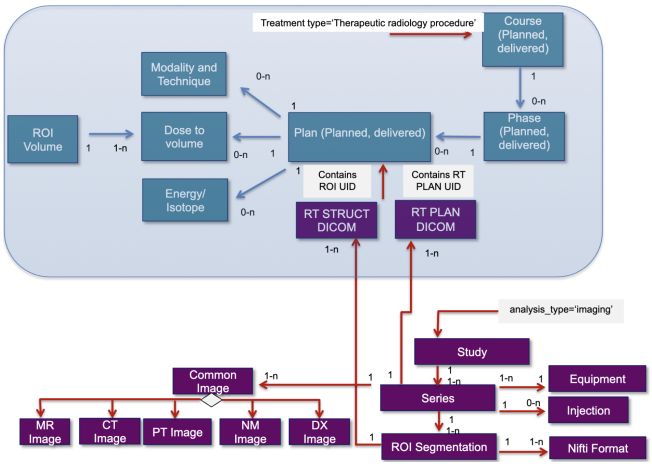


Figure 3 : OSIRIS RT-Common Data Model (CDM)

The *figure 3 (in blue)* shows the radiotherapy data model with the CDEs and the relationships between them. The RT-CDM is connected to the DICOM-Imaging CDM *(in purple)* and is an extension of the OSIRIS clinical CDM when the "Treatment_Type" DE value is "Therapeutic radiology procedure". The RT-CDM facilitates the development and management of the course, phases and treatment plans after care has been provided.

Given that health data change over time, the OSIRIS group focused its efforts on modeling an event-based temporal data model. Each DE of the OSIRIS RT-CDM is linked to a particular treatment plan (ie, planned or delivered).

OSIRIS RT set

Once the OSIRIS set reached a steady state in terms of its semantic content, a first version was released in december 2022. This current release, built on an event-based data model, describes a minimum data set that allows capture of the RT course. This release, freely available to the community on GitHub consists of a minimal data set of 32 clinical and 60 imaging items required and validated by the national scientific board[10]. CLs are based on international terminologies.

OSIRIS-RT FHIR Implementation Guides (IG)

To ensure the sustainability, expandability, and interoperability of the OSIRIS RT set with other systems, the first release was made compatible with the growing health information exchange standard HL7 FHIR in 2023. The OSIRIS-FHIR implementation guide is documented on the GitHub[11]. This guide summarizes a set of rules on how to implement OSIRIS with the HL7 FHIR standard. You can install this implementation guide to facilitate its maintenance and evolution.

HYPO-RT Prostate[12] : a national RT project as a proof-of-concept

A national, multi-centric proof-of-concept is underway to validate the OSIRIS RT set and the CDM. The evaluation will be based in real-life situations on 10.000 patients treated for prostate cancer. The aim is to assess the efficacy and safety of moderately hypofractionated and extremely hypofractionated radiotherapy regimens, in terms of recurrence-free survival, metastasis-free survival, overall survival, severe toxicity (digestive or urinary), compared with normofractionated regimens.

4 Discussion

The OSIRIS group has carried out a substantial amount of work, with the support of INCa (French Institut National du Cancer) but also thanks to the desire and effort of a collective mobilized around these issues of data sharing in radiotherapy. The deliverables of the project will be promoted through national use cases in 2024 (e.g. ODH, OncoDS projects[13]) or working discussions at European level (e.g. UNCAN.eu[14]).

5 Conclusion

OSIRIS use cases for collecting and sharing patient data for clinical trials/studies with Real World Data (OSIRIS-RWD) demonstrate the potential for improving research and care. This is being demonstrated by international standards such as FHIR, and by new national projects planned for 2024 all driven in the field by stakeholders.

With these new projects, the use of OSIRIS RT at national level will expand rapidly, and the project leader will face a number of challenges. It will be necessary to establish a transparent and participative governance (steering committee, operational) but also to establish a community of practice in Radiotherapy. The aim of this community would be to set up an open forum to listen to and share requirements and experiences around OSIRIS RT.

The development of technical solutions for data collection in OSIRIS format from ROIS (e.g. MOSAIQ©,ARIA©), documentation (e.g. OSIRIS-FHIR IG, documented use cases, maintenance and evolution of RT-CDM and minimal set) will be necessary for the completion of projects in 2024.

References

[1] DAPB0111: Radiotherapy Data Set [Internet]. NHS Digital [cited 2024 Jan 22] Available from: https://digital.nhs.uk/data-and-information/information-standards/information-standards-and-data-collections-including-extractions/publications-and-notifications/standards-and-collections/dapb0111-radiotherapy-data-set

[2] Hayman JA et al. Minimum Data Elements for Radiation Oncology: An American Society for Radiation Oncology Consensus Paper. Pract Radiat Oncol. 2019 Nov;9(6):395-401. DOI: 10.1016/j.prro.2019.07.017.

[3] Phillips MH, Serra LM, Dekker A, et al: Ontologies in radiation oncology. Phys Med 72:103–113, 2020. DOI: 10.1016/j.ejmp.2020.03.017

[4] CodeX Radiation Therapy [Internet]. HL7 FHIR [cited 2024 Jan 22] Available from : https://build.fhir.org/ig/HL7/codex-radiation-therapy/index.html

[5] Guerin J, Laizet Y, Le Texier V, et al: OSIRIS: A minimum data set for data sharing and interoperability in oncology. JCO Clin Cancer Inform. 5:256-265, 2021. DOI : 10.1200/CCI.20.00094.

[6] Welcome to the NCBO BioPortal | NCBO BioPortal [Internet][cited 2024 Jan 22] Available from: https://bioportal.bioontology.org/

[7] RadioTransNet – Réseau National de Recherche Préclinique en Radiothérapie [Internet][cited 2024 Jan 22] Available from : https://radiotransnet-fr.translate.goog/?_x_tr_sl=auto&_x_tr_tl=en&_x_tr_hl=en&_x_tr_pto=wapp

[8] The unicancer group of translational research and development in radiation oncology (UNITRAD) [Internet][cited 2024 Jan 22] Available from : https://recherche.unicancer.fr/en/experts-groups/the-unicancer-group-of-translational-research-and-development-in-radiation-oncology-unitrad/

[9] GitHub OSIRIS - CDMRadiotherapieOSIRIS.pdf [Internet][cited 2024 Jan 22] Available from : https://github.com/InstitutNationalduCancer/OSIRIS/blob/MED-OSIRIS/documentation/CDMRadiotherapieOSIRIS.pdf

[10] GitHub OSIRIS - ModeleRadiotherapieOSIRIS.pdf [Internet] [cited 2024 Jan 22] Available from : https://github.com/InstitutNationalduCancer/OSIRIS/blob/MED-OSIRIS/documentation/ModeleRadiotherapieOSIRIS.pdf

[11] GitHub OSIRIS – ImplementationGuide_OsisisFHIR [Internet] [cited 2024 Jan 22] Available from : https://github.com/InstitutNationalduCancer/ImplementationGuide_OsirisFHIR/

[12] Health-Data-Hub Unibase [Internet][cited 2024 Jan 22] Available from : https://www-health--data--hub-fr.translate.goog/actualites/unicancer-la-ligue-nationale-contre-le-cancer-et-le-hdh-annoncent-les-laureats-de-la?_x_tr_sl=auto&_x_tr_tl=en&_x_tr_hl=en&_x_tr_pto=wapp

[13] Health-Data-Hub OncoDS [Internet][cited 2024 Jan 22] Available from : https://www-health--data--hub-fr.translate.goog/actualites/decouvrez-les-6-laureats-selectionnes-pour-la-1ere-vague-de-lappel-projets-visant?_x_tr_sl=auto&_x_tr_tl=en&_x_tr_hl=en&_x_tr_pto=wapp

[14] Uncan.eu [Internet][cited 2024 Jan 22] Available from : https://uncan.eu/

Towards a fully automatic workflow for investigating the dynamics of lung cancer cachexia during radiotherapy using cone beam computed tomography

Lars H.B.A. Daenen[1,2], Wouter R.P.H. van de Worp[3], Behzad Rezaeifar[1], Joël de Bruijn[4], Peiyu Qiu[3], Justine M. Webster[3], Ramon C. J. Langen[3], Frank Verhaegen[1*], and Cecile J.A. Wolfs[1*]

[1]Department of Radiation Oncology (Maastro), GROW School for Oncology and Reproduction, Maastricht University Medical Centre+, The Netherlands

[2]Medical Image Analysis group, Department of Biomedical Engineering, Eindhoven University of Technology, The Netherlands

[3]Department of Respiratory Medicine, NUTRIM Institute of Nutrition and Translational Research in Metabolism, Maastricht University, The Netherlands

[4] SmART Scientific Solutions BV, Maastricht, The Netherlands

*these authors contributed equally

Abstract Cachexia is a devastating condition, often occurring during cancer, diminishing cancer treatment effects and increasing mortality. Currently, there are no methods to accurately monitor cachexia during cancer treatment. Therefore, we present a proof-of-concept for monitoring cachexia based on cone beam computed tomography (CBCT) images acquired during radiotherapy treatment.

Data from 140 stage III non-small cell lung cancer patients was used. Two deep learning models, CycleGAN and CUT, were used for unpaired training of CBCT-to-CT conversion. U-Net was used for automatic pectoralis muscle segmentation.

Both CycleGAN and CUT restored the Hounsfield Unit fidelity and reduced artifacts in CBCT, with CUT displaying the best performance. The U-Net model achieved a DSC of 0.88 on a hold-out test set.

This study shows a proof-of-concept for automatic monitoring of the pectoralis muscle area during radiotherapy treatment based on CBCT, which can serve as a surrogate for cachexia progression.

1 Introduction

Lung cancer is the second most prevalent cancer type worldwide and is the leading cause of cancer-related deaths [1]. More than half of lung cancer patients suffer from cachexia: a devastating syndrome characterized by loss of skeletal muscle mass, with or without loss of fat tissue. It is a sign of poor prognosis and is suggested to account for 20% of all cancer-related deaths. Cachexia is associated with diminished effects of chemotherapy, functional impairment and increased mortality [2]. Consequently, it is important to gain insights into the dynamics of cachexia and monitor muscle wasting over time.

Assessment of skeletal muscle mass at the third lumbar vertebrae (L3 level) on computed tomography (CT) imaging is considered the gold standard for evaluating body composition, due to its high accuracy and specificity [2]. However, manual fat and muscle tissue delineation is a tedious process, limiting the widespread use of body composition analysis in clinical decision-making. For this purpose, deep learning (DL) segmentation models have been developed for the segmentation of skeletal muscle tissue, as well as subcutaneous and visceral adipose tissue, in 2D slices of CT images at the L3 level [3][4].

During the course of radiotherapy treatment, part of standard treatment for stage III non-small cell lung cancer (NSCLC) patients, cone beam CT (CBCT) images are acquired before each fraction, to verify patient positioning and check for anatomical changes. As CBCT images are acquired sequentially over time, commonly for up to 30 fractions, they contain valuable information to investigate the dynamics of cachexia during the course of treatment. However, the L3 level is often not visible in CBCT and CT images of lung cancer patients. As an alternative, skeletal muscle mass at the level of the fourth thoracic vertebrae (T4) has been shown to correlate with muscle mass at L3 and is suited as a surrogate for body composition analysis [5]. Furthermore, low pectoralis muscle area (PMA), approximately at T4 level, at the time of diagnosis is associated with worse overall survival of patients with NSCLC [6].

Using CBCT images to monitor muscle mass is challenging, as CBCT images are lower in quality than conventional CT, due to more scattered radiation reaching the detector and more artifacts resulting from breathing motion. This results in suboptimal Hounsfield Unit (HU) fidelity, and thus suboptimal segmentation of fat and muscle tissue. However, DL models are able to translate low-quality CBCT images to their high-quality CT counterparts [7][8], so-called image-to-image translation (I2I).

The purpose of this study is two-fold. First, I2I of CBCT images to high-quality synthetic CT images (sCT) will be performed, on a dataset of images from stage III NSCLC patients. Secondly, a DL segmentation model will be developed to segment the pectoralis major and minor muscles at the first axial slices above the aortic arch in both the synthetic and original CT images. Ultimately, this study seeks to develop an automatic framework to monitor muscle wasting during the course of radiotherapy treatment and gain insights into the dynamics of lung cancer cachexia.

2 Materials and Methods

2.1 CBCT-to-CT conversion

2.1.1 Dataset & preprocessing

The dataset consisted of the planning CT and, on average, 30 CBCT images of 140 patients treated for stage III NSCLC. First, both the CT and CBCT images were

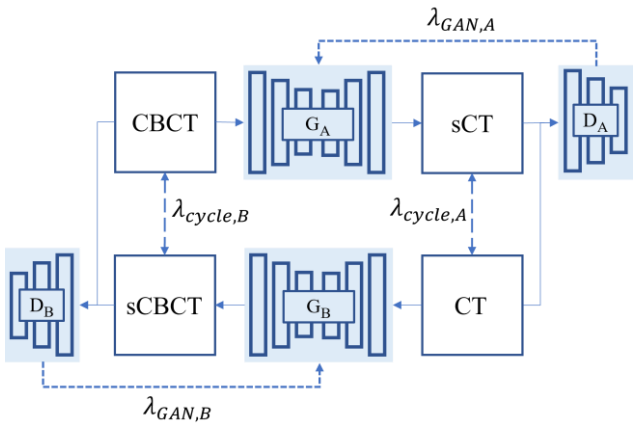


***Figure 1**: Schematic representation of the architecture of the CycleGAN model. G_A and G_B are the generators, and D_A and D_B are the discriminators. λ_{cycle} and λ_{GAN} are the loss functions used. (s)CT and (s)CBCT are the (synthetic) computed tomography (CT) and (synthetic) cone-beam CT images, respectively.*

resampled to a uniform voxel spacing of 0.977 x 0.977 x 3 mm^3. Next, the CT images were cropped to the axial range of the corresponding CBCT images, to ensure that the same anatomical region was present in both the CT and CBCT images. Furthermore, non-anatomical information (e.g. treatment couch and patient immobilization device) was removed using a binary mask and the HU intensities of the CT and CBCT images were clipped between -1000 and 2000 HU. Lastly, the dataset was split with a ratio of 90/10/10 for training (n=112), validation (n=14) and testing (n=14).

2.1.2 CBCT-to-CT models

The goal of image-to-image translation (I2I) is to translate an input image from the source domain (CBCT) to the target domain (CT), such that the structure, i.e. content, of the input image is preserved while matching the appearance of the output image in the target domain. Two I2I models, CycleGAN and CUT, were chosen as these allow for unpaired training, i.e. when corresponding images from two domains cannot be obtained. This is also the case for CBCT-to-CT conversion, as there might be anatomical changes between the planning CT and CBCTs due to breathing.
The state-of-the-art network for unpaired I2I is CycleGAN, first proposed by Zhu et al. [9]. In contrast to standard GAN architectures, CycleGAN comprises two generators and two discriminators as displayed in Fig. 1, which learn both forward and backward mapping of the input and output (i.e. from CBCT to CT and vice versa). The generators seek to generate realistic output images from a given input, while the discriminators seek to distinguish real images from generated ones.
The CUT model by Park et al. [10] uses the same generator and discriminator network as CycleGAN but employs only one generator-discriminator pair as displayed in Fig. 2.

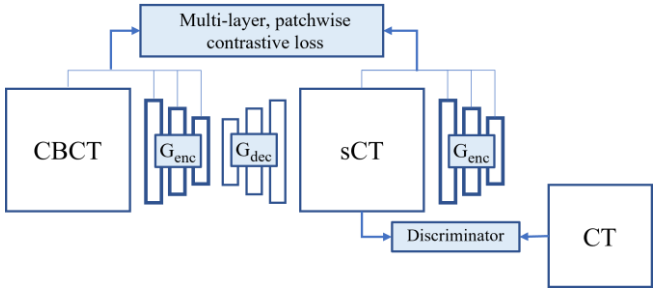


***Figure 2**: Schematic representation of the architecture of the CUT model. G_{enc} and G_{dec} together comprise the generator. CT, sCT and CBCT are the computed tomography (CT), synthetic CT and cone-beam CT images, respectively.*

Instead of learning the backward mapping, it uses a contrastive learning framework, by calculating a multi-layer, patchwise contrastive loss.

2.1.3 Training details

We trained both CycleGAN and CUT with the default configuration of hyperparameters and did not tune these further as this already resulted in a stable training process and satisfactory results. The model was trained with a constant learning rate of 0.0002 for 40 epochs, followed by 40 epochs with a linearly decaying learning rate until zero. Validation based on the Fréchet Inception Distance (FID) was implemented [11]. The best epochs in terms of the FID on the validation set were chosen as the final models.

2.1.4 Evaluation

To obtain a reference CT (rCT) scan for evaluation, image registration of CT to first-fraction CBCT was performed using the Elastix toolbox [12]. A rigid transformation was applied for initial alignment, followed by a multi-resolution B-spline transformation to account for deformational changes between planning CT and CBCT.
The quality of the sCT images was evaluated both on the individual image level, as well as on the group level. On the individual image level, mean absolute error (MAE), structural similarity (SSIM) [13] and peak signal-to-noise ratio (PSNR) were used as image quality metrics.
On the group level, the FID was used, a widely accepted metric for measuring the performance of generative models. It measures the distance between the probability distributions of real and generated images.

2.2 Pectoralis segmentation

2.2.1 Dataset

For pectoralis muscle segmentation, the first axial slice above the aortic arch was selected in the CBCT images from fractions 1, 10, 20, and 30. The best-performing CBCT-to-CT model was used to convert the selected CBCT images

***Table 1**: Median and inter-quartile range of the MAE, PSNR and SSIM for the CBCT and sCT images, compared to the ground truth rCTs. Furthermore, the FID is shown calculated for the CBCT and sCT images, compared to the ground truth CTs.*

Metric	MAE	PSNR	SSIM	FID
CBCT	26.8 ± 12.3	30.5 ± 3.46	0.874 ± 0.053	72.7
CycleGAN	22.8 ± 10.5	30.7 ± 3.58	0.898 ± 0.048	**39.0**
CUT	**22.1 ± 10.8**	**30.8 ± 3.62**	**0.902 ± 0.050**	41.4

to sCT images. Several patients were omitted, as either the first axial slice above the aortic arch was not present in the CBCT volume, or artifacts in the original CBCT were too prominent, rendering the pectoralis muscle indistinguishable. This resulted in a dataset of 326 sCT slices. In these slices, the pectoralis was manually segmented using the Slice-O-Matic software by TomoVision (Magog, QC, Canada).

2.2.2 Model & training details

The U-Net architecture proposed by Ronneberger et al. [14] was used for segmentation of the pectoralis major and minor. The U-Net model consists of an encoding and decoding path, connected via skip connections, to leverage the low- and high-level information within the images. The Combo-Loss (a combination between cross entropy and Dice loss) was used as the loss function [15]. The model was trained for 100 epochs using the Adam optimizer, with a learning rate of $3 \cdot 10^{-4}$ and batch size of 2. On-the-fly data augmentation was performed by randomly scaling (0.95, 1.05), translating in x- and y direction (-25, +25 pixels) and rotating (-5, 5°) the images.

2.2.3 Evaluation

The pectoralis muscle segmentation performance was evaluated based on the Dice Similarity Coefficient (DSC), which calculates the overlap between the segmentations over the total area of the segmentations.

3 Results

3.1 CBCT-to-CT conversion

Figure 3 shows the overall HU distribution of the CT, CBCT and corresponding sCT images generated by the CycleGAN and CUT models. Both CycleGAN and CUT restore the overall HU distribution, as they closely follow the bimodal distribution of the CTs. However, CUT follows the CT HU distribution more closely in the range [-29; 150], which is the range corresponding to muscle tissue.

Table 1 shows the performance of the CycleGAN and CUT model in terms of MAE, PSNR and SSIM. Both CycleGAN and CUT show improvement in terms of MAE, PSNR and SSIM, with CUT displaying the best performance. Both models show increased image similarity between the sCT and CT datasets as a whole, measured by the FID, with

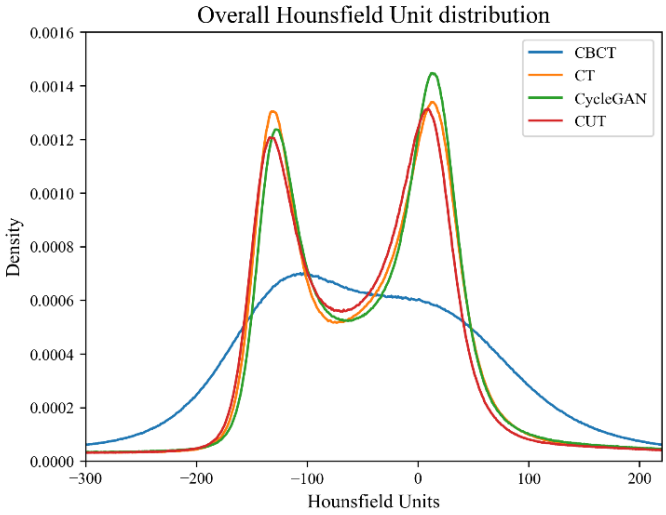


***Figure 3**: Overall Hounsfield Unit distribution across the test set.*

CycleGAN showing the best performance. Ultimately, the CUT model was selected to convert CBCT images to sCT images for pectoralis muscle segmentation.

3.2 Pectoralis muscle segmentation

The U-Net achieved a median DSC of 0.97, 0.94, and 0.88 on the training, validation and test set, respectively (Table 2). Fig. 4 show examples of segmentations with a high and low DSC, respectively. Low DSC is mainly due to oversegmentation at the anterior side of the pectoralis muscle.

Split	Median	Min	Max
Train	0.97	0.90	0.99
Validation	0.94	0.88	0.98
Test	0.88	0.76	0.95

***Table 2**: Median, minimum and maximum DSC for the training (N=260), validation (N=31) and test (N=35) data splits.*

4 Discussion

Both CycleGAN and CUT restore the HU distribution to that of the CT, and reduce artifacts. CUT is chosen as the best performing model as it results in more accurate HU distribution in the range corresponding to muscle tissue, which is relevant for this work and slightly better performance in terms of the quantitative metrics.

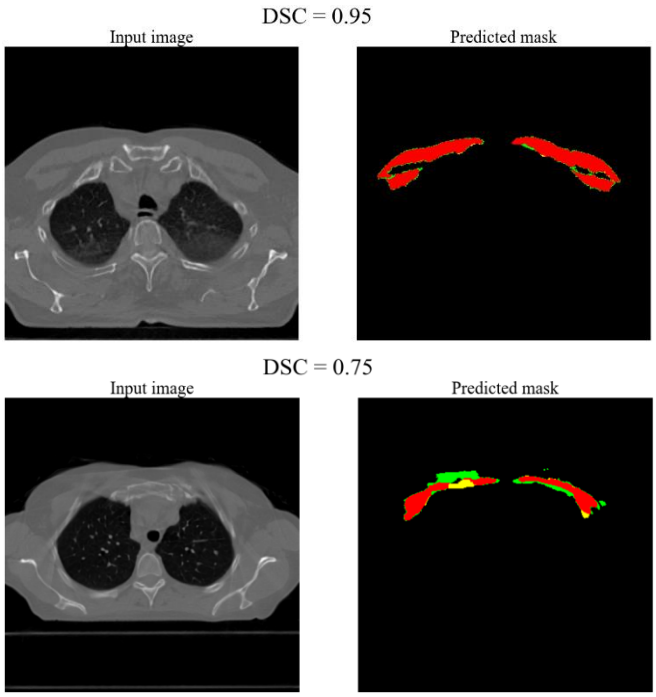


***Figure 4**: High and low DSC segmentation. red, green and yellow areas indicate correct, over- and undersegmentation, respectively.*

The U-Net model shows good performance in segmenting the pectoralis muscle. Fig. 4 shows that artifacts still present after I2I result in a lower quality segmentation. Therefore, it is important to investigate whether the CBCT-to-CT conversion models can be improved, or that artifacts might be too disturbing in some cases.

In order to fully automate the proposed workflow, an additional step needs to be developed to automatically check if the T4 level is present in the CBCT volume and select the corresponding slice.

In addition, robust metrics need to be formulated that indicate performance of CBCT-to-CT conversion and pectoralis muscle segmentation performance during the application of the models in clinical practice, such that the trustworthiness of the final segmentation can be evaluated in a fully automatic workflow, and manual corrections can be performed if needed.

5 Conclusion

This study shows the potential of automatic monitoring of pectoralis muscle area using CBCT, by first restoring HU fidelity and reducing artifacts and then automatically segmenting the pectoralis muscle area. This allows for monitoring muscle wasting during the course of radiotherapy treatment and gaining insights into the dynamics of cachexia.

Funding: This collaborative project is co-financed with a PPP allowance made available by Health~Holland, Top Sector Life Sciences & Health to stimulate public-private partnerships.

References

[1] C. P. Wild, E. Weiderpass, and B. W. Stewart, *World Cancer Report: Cancer Research for Cancer Prevention*. Lyon, France, 2020. [Online]. Available: http://publications.iarc.fr/586

[2] J. Ni and L. Zhang, "Cancer cachexia: Definition, staging, and emerging treatments," *Cancer Manag Res*, vol. 12, pp. 5597–5605, 2020, doi: 10.2147/CMAR.S261585.

[3] L. L. G. C. Ackermans *et al.*, "Deep learning automated segmentation for muscle and adipose tissue from abdominal computed tomography in polytrauma patients," *Sensors*, vol. 21, no. 6, pp. 1–13, 2021, doi: 10.3390/s21062083.

[4] S. Dabiri, K. Popuri, E. M. C. Feliciano, B. J. Caan, V. E. Baracos, and M. F. Beg, "Muscle Segmentation in axial Computed Tomography (CT) Images at the Lumbar (L3) and Thoracic (T4) levels for Body Composition Analysis," *Comput Med Imaging Graph.*, vol. 75, pp. 47–55, 2019, doi: 10.1016/j.compmedimag.2019.04.007.

[5] J. Van den Broeck, M. J. Sealy, C. Brussaard, J. Kooijman, H. Jager-Wittenaar, and A. Scafoglieri, "The correlation of muscle quantity and quality between all vertebra levels and level L3, measured with CT: An exploratory study," *Front Nutr*, vol. 10, 2023, doi: 10.3389/fnut.2023.1148809.

[6] C. M. Kinsey, R. S. J. Estepar, J. van der Velden, B. F. Cole, D. C. Christiani, and G. R. Washko, "Lower Pectoralis Muscle Area is Associated with a Worse Overall Survival in Non- Small Cell Lung Cancer," *Cancer Epidemiol Biomarkers*, vol. 26, no. 1, pp. 38–43, 2017, doi: 10.1158/1055-9965.EPI-15-1067.

[7] X. Wang *et al.*, "Synthetic CT generation from cone-beam CT using deep-learning for breast adaptive radiotherapy," *J Radiat Res Appl Sci*, vol. 15, no. 1, pp. 275–282, 2022, doi: 10.1016/j.jrras.2022.03.009.

[8] J. F. Wynne *et al.*, "Rapid unpaired CBCT-based synthetic CT for CBCT-guided adaptive radiotherapy," *J Appl Clin Med Phys*, vol. e14064, 2023, doi: 10.1002/acm2.14064.

[9] J. Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, 2017, pp. 2242–2251. doi: 10.1109/ICCV.2017.244.

[10] T. Park, A. A. Efros, R. Zhang, and J. Y. Zhu, "Contrastive Learning for Unpaired Image-to-Image Translation," in *European Conference on Computer Vision*, 2020.

[11] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," in *Proceedings of the 31st Conference on Neural Information Processing Systems*, Long Beach, CA, USA, 2017, pp. 6629–6640.

[12] S. Klein, M. Staring, K. Murphy, M. a. Viergever, and J. Pluim, "Elastix: A Toolbox for Intensity-Based Medical Image Registration," *IEEE Trans Med Imaging*, vol. 29, no. 1, pp. 196–205, 2010, doi: 10.1109/TMI.2009.2035616.

[13] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image Quality Assessment: From Error Visibility to Structural Similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004, doi: 10.1109/TIP.2003.819861.

[14] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, vol. 9351, pp. 234–241, 2015, doi: 10.48550/arXiv.1505.04597.

[15] S. A. Taghanaki *et al.*, "Combo loss: Handling input and output imbalance in multi-organ segmentation," *Computerized Medical Imaging and Graphics*, vol. 75, pp. 24–33, 2019, doi: 10.1016/j.compmedimag.2019.04.005.

Enhancing CBCT image quality for radiotherapy using multi-channel cycleGAN for CBCT-to-CT synthesis

Chelsea Sargeant[1], Robert Chuter[1,2], Jane Shortall[1], Andrew Green[3], Jacqui Bridge[2], Phil Fendall Amaro[2], Rachael Bailey[2], and Alan McWilliam[1]

[1]Division of Cancer Sciences - Faculty of Biology, Medicine and Health, The University of Manchester, Manchester, United Kingdom.
[2]The Christie NHS Foundation Trust, Manchester, United Kingdom
[3]EMBL's European Bioinformatics Institute, United Kingdom.

Abstract Deep learning-based image synthesis techniques, like cycle-consistent adversarial networks (cycleGANs), have improved the image quality of cone-beam computed tomography (CBCT) scans for radiotherapy by generating synthetic CTs (sCTs). By leveraging multi-channel input to provide a cycleGAN with a broader range of information upon training, our approach aims to preserve soft tissue contrast and reduce artefacts. The results demonstrate the successful generation of high-fidelity sCT images, with multi-channel models outperforming single-channel input across various image similarity metrics (mean absolute error, peak signal-to-noise ratio and structural similarity index). Furthermore, our approach demonstrates an ability to generalise to an unseen test data set of CBCT scans acquired from prostate cancer patients. A qualitative assessment, conducted by three expert radiographers, shows promising results in terms of overall image quality and artefact reduction and offers valuable insight into the clinical utility of sCTs. We emphasise the need for standardized qualitative assessments and downstream task evaluations to ensure robust insights.

1 Introduction

On-treatment cone-beam computed tomography (CBCT) is integral to adaptive radiotherapy, yet its image quality often suffers from artefacts such as streaking and noise, diminishing its utility compared to planning CT (pCT) scans. Deep learning-based image synthesis techniques, notably cycle-consistent generative adversarial networks (cycleGANs), have shown promise in enhancing CBCT image quality to a level comparable to pCT scans by generating a synthetic CT (sCT). However, challenges persist due to the wide Hounsfield unit (HU) range (-1024 to 2000 HU) in human anatomy CT scans. Soft tissue, confined to a narrow HU range, and high-density features like metal, often exceeding 2000 HU, present particular difficulties. This impacts the performance of image synthesis models, resulting in a loss of information for radiotherapy, such as soft tissue detail and accurate handling of artefacts. To address this, we propose leveraging a multi-channel input, consisting of multiple contrast settings, to provide the network with a broader range of information. By doing so, we aim to emphasise specific image features and enable the network to learn multiple contrast settings separately. This approach seeks to preserve soft tissue contrast and improve artefact reduction, ultimately enhancing the use of CBCT images in the radiotherapy workflow. By integrating observer testing alongside quantitative image similarity metrics, we aim to provide a comprehensive evaluation of the sCT images. This holistic approach aims to ensure that our method not only enhances image quality according to objective measures but also aligns with the perceptual expectations and clinical requirements of radiographers, thereby validating its suitability for adaptive radiotherapy applications. Additionally, we show that image synthesis models trained on data from independent centres can be applied to CBCT images in an unseen centre.

2 Materials and Methods

2.1 Training Data

For training, we used a publicly-available pelvis dataset consisting of CBCT and pCT scans for patients who underwent pelvis radiotherapy at three Dutch institutions: Radboud University Medical Center, University Medical Center Utrecht and University Medical Center Groningen [1]. A total of 180 CBCT-pCT pairs were available for training. For image acquisition parameters, see [1].

2.2 Testing Data

The test set consisted of CBCT and pCT scans from 42 prostate cancer patients who underwent radiotherapy at The Christie NHS Foundation Trust. The pCT images were acquired on Phillips Brilliance Big Bore CT or Siemens SOMATOM Definition AS (Erlangen, Germany) with 120 kVp. The CBCT images were acquired using Elekta Synergy imaging systems (XVI 5.52, Elekta, Crawley, UK) with 120 kVp. The exposure current was 40mA.

2.3 Pre-processing

Before training and inference, all images were resampled to the same voxel spacing: $1 \times 1 \times 2.5\text{mm}^3$ and a rigid image registration between the CBCT and pCT was performed to align the images. Binary masks based on the patient's pCT contour were generated and applied to the images. We then performed multi-channel range selection and normalisation before padding/cropping the images to Nx448x488 voxels, where N is the number of axial slices. All pre and post-processing were applied to 3D images, however, individual

axial slices were extracted from all image volumes (pCT and CBCT pairs) for training. A random 80/20 split was used to create train-time training and validation sets.

2.3.1 Multi-channel range selection and normalisation

To investigate the impact of training with additional information, the pCT and CBCT scans were normalised into either two or three separate channels using windowing to enhance the contrast of structures within the patient (Figure 1). The channels were then independently scaled to the range $[0, 1]$ using min-max normalisation to improve training stability and concatenated for input into the network.

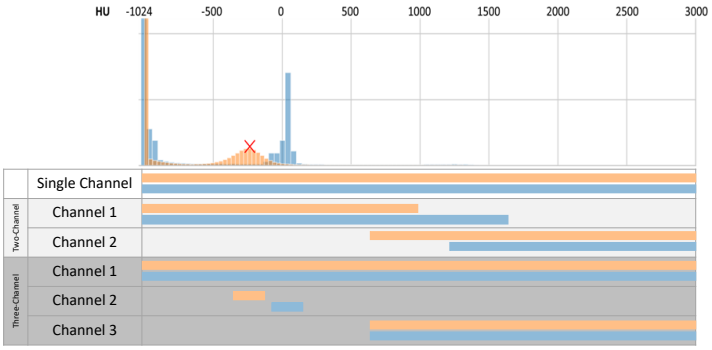


Figure 1: CBCT and pCT image histogram. Below the histogram shows the ranges in HU of each of the three input channels for the three models (single-channel, two-channel, three-channel).

Two channel input In the first channel, the CBCT and pCT were clipped to [-1024, 1000] HU and [-1024, 1600] HU, respectively, while the second channel ranged from [600, 3000] HU for CBCT and [1200, 3000] HU for pCT. Overlapping channels were used to facilitate recombination after sCT generation.

Three channel input In the first channel, we use the full width of the image range $[-1024, 3000]$ HU. In channel two, we use a contrast setting used to view soft tissue structures; for the pCT this was $[-150, 150]$ HU chosen to be centred on the water peak. Due to a lack of calibration of the CBCT images, both within and among centres, an automated peak finder was implemented to set the level to the soft tissue peak, with a fixed window width of ± 150 HU. In the final channel, both the CT and CBCT images were clipped to $[600, 3000]$ HU to capture information about the high-density structures within the patient.

2.4 Image Synthesis Method

Our method is based on a traditional cycleGAN architecture [2] with some specific modifications. CycleGANs consist of two sets of competing generator and discriminator networks trained simultaneously to enable unpaired training. The generators learn to produce CT-like output from CBCT input (a sCT) and vice versa (synthetic CBCT (sCBCT) output from CT input). The discriminators learn to classify CT/sCT and CBCT/sCBCT input as real and fake. Additionally, cycle consistency was enforced by inputting the synthetic image into the reverse generator to reverse the translation back to the original image modality.
The discriminator used in our modified cycleGAN is as described in [2], however, we modified the generator architecture. UNet-like long-range skip connections were added between corresponding down- and up-sampling convolutional levels to preserve contextual information. Finally, attention gates were added to the skip connections to emphasize salient features propagated forward from earlier in the network.
In the multi-channel case, the generator receives multi-channel input and produces multi-channel output. This is performed for ease of implementation of the cycle consistency aspect, where the output of each generator is also sent through the reverse generator.

2.5 Implementation Details

The model was implemented in PyTorch Lightning version 1.5.10 and trained using a single NVidia GeForce RTX 3090 GPU with 24GB VRAM. The batch size was 1. Each model was trained for a maximum of 200 epochs, however, early stopping was initiated if the total generator validation loss did not improve for 20 epochs. The models were trained with Adam Optimiser. The generator learning rate was set to 1e-4 while the discriminator learning rates were set to 2e-4. Upon inference, the trained CBCT-to-sCT generator networks were applied to preprocessed CBCTs to generate a corresponding sCT. The best model was selected according to the total generator validation loss.

2.6 Post-Processing / Channel Fusion

The padding or cropping performed during preprocessing was reversed to restore the sCTs to the original CBCT image size. Next, the HU intensity range of each channel was restored using a linear re-scaling: for example, channel one of the three-channel sCT was rescaled to [-1024, 3000] HU; channel two was re-scaled to [-150, 150] HU and channel three was re-scaled to [600, 3000] HU. For the two-channel sCT, the channels were rescaled to [-1024, 1600] and [1200, 3000] HU, respectively. The information from each channel were then combined into a single representation.

Two-channel fusion Voxels with HU values greater than 1400, as determined by the centre of the overlap region, were allocated the intensity value from the second channel, otherwise, voxels were allocated the value from the first channel.

Three-channel fusion The full-width image (channel one) underwent modifications based on specific conditions: values within the narrow range (-150 to 150 HU) were substituted

with channel two values, and values > 600 HU were replaced with channel three values.

3 Evaluation

Quantitative Evaluation The generated sCTs are assessed according to image similarity metrics, computed between the sCT and the pCT, rigidly registered to the CBCT. We include mean absolute error (MAE), structural similarity index (SSIM), and peak signal-to-noise ratio (PSNR). The Wilcoxon signed rank test was employed to determine whether there were any significant differences between the calculated metrics from each model. A significance level of 0.05 was used.

Qualitative Evaluation To evaluate both the clinical usefulness of synthetic CTs and the impact of multi-channel input, three radiographers, with 7 years or more experience, were asked to score the sCTs a subset of 10 cases from the test dataset (3 sCTs per patient). The radiographers were shown a sCT and corresponding pCT and CBCT, pseudo-randomly to ensure that sCTs generated from the same CBCT were not shown sequentially. They were asked to score the sCT concerning the severity of artefacts present, anatomical accuracy, clinical usability and overall image quality using a five-point Likert scale. The inter-observer variation was assessed by calculating the intra-class correlation coefficient (ICC) based on a mean rating (k=3), two-way random effects, absolute-agreement model.

4 Results

The sCTs generated by our methods are displayed in Figure 3 for an example test patient. Visually, the methods preserve the anatomical structure of the CBCT images and the contrast in soft tissue is enhanced in each of the models. Table 1 summarises the quantitative results obtained by image similarity metrics; MAE, PSNR and SSIM. Our reported image similarity metrics indicate good similarity to the ground truth images for each model, however, both models trained with multi-channel input significantly outperform the single-channel model according to the Wilcoxon signed rank test ($p < 0.05$), with the dual-channel model performing best. There was no significant difference found between the two and three-channel performance for any metric.

Table 1: Image similarity metrics between sCT and pCT.

Input	MAE (HU) ↓	SSIM ↑	PSNR (dB) ↑
One-channel	77.5 ± 10.1	0.91 ± 0.02	29.7 ± 1.5
Two-channel	$\mathbf{62.8 \pm 11.5}$	$\mathbf{0.94 \pm 0.02}$	$\mathbf{30.4 \pm 1.6}$
Three-channel	65.4 ± 10.7	0.94 ± 0.02	30.0 ± 1.4

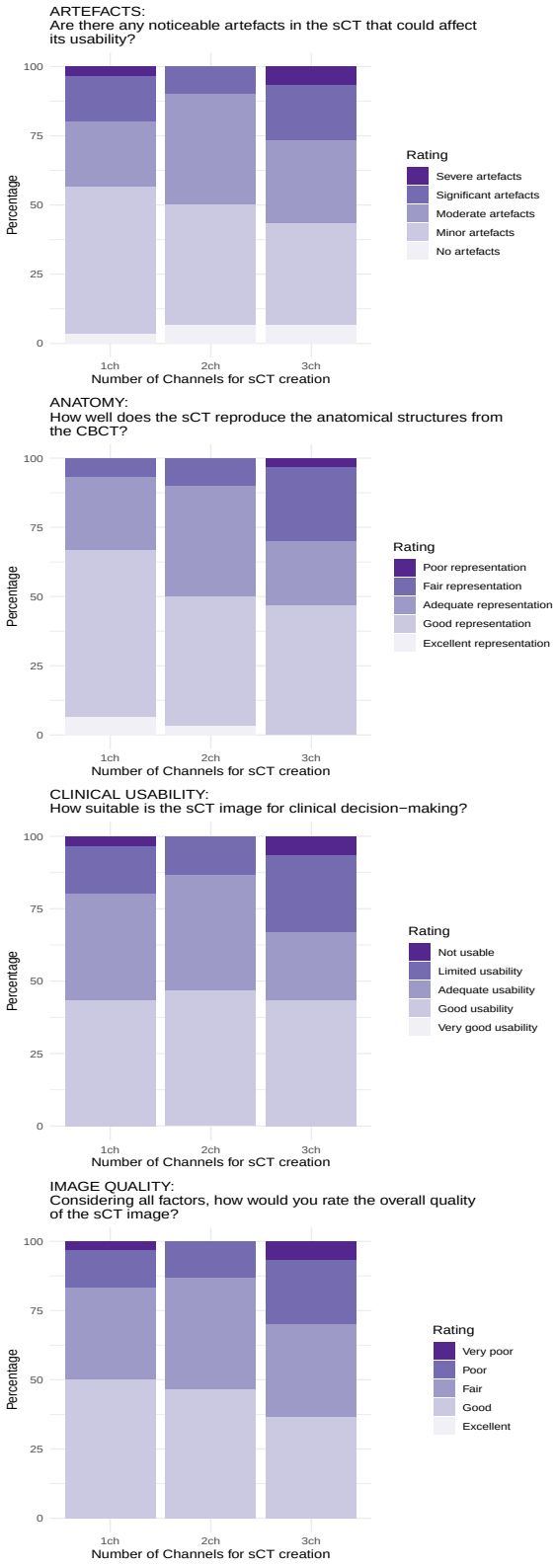


Figure 2: Results from visual grading analysis showed similar scores for single and two-channel approaches across all observers. However, three-channel sCTs performed worse overall.

4.1 Observer Assessment

Observers generally concurred with the image similarity metrics, with the two-channel sCT images receiving the highest ratings for acceptability concerning clinical usability, image quality and artefact reduction although lower in anatomical

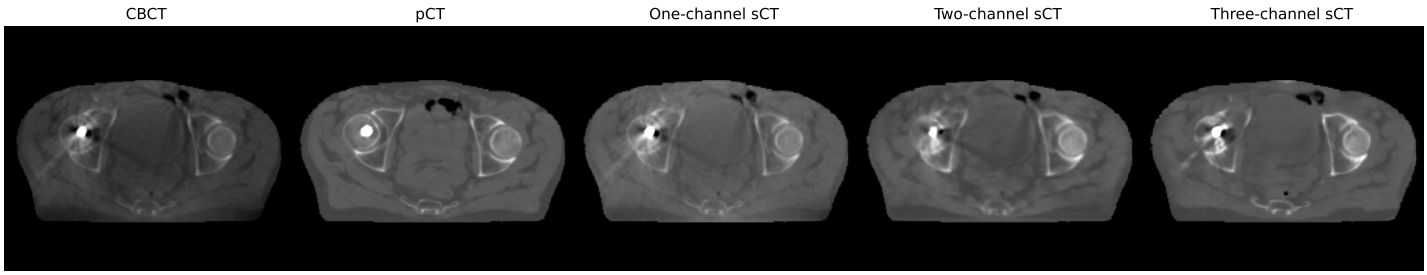


Figure 3: CBCT, pCT and sCTs generated by our methods for example prostate patient with metal implant in the hip. Images are shown with HU scale of [-500,1000] for ease of visualisation.

accuracy compared to the single-channel sCT. Notably, the three-channel sCT was rated less effective, primarily due to the noticeable distortion of organs at risk and the presence of synthetic artefacts, which outweighed any benefits from improved HU accuracy. This suggests the need to improve the ensemble of channels in post-processing. On average, there was minimal difference in score across sCT type for all observers. However, there was limited agreement among observers, with ICC measuring below 0.6 for each category.

5 Discussion

Our approach demonstrates the successful generation of high-quality sCT images from CBCT scans, with over 75% of images rated at least adequate for clinical decision-making for both single and two-channel models. Additionally, more than 75% of the images were rated at least fair in terms of image quality for these models. Incorporating image data at varying windows/levels within a concatenated multi-channel input has enabled us to address challenges inherent in CBCT imaging. Notably, we observed improvements in the suppression of artefacts, clinical usability and image quality with the addition of a second channel. However, in some instances, anatomical distortion was reported, and this was amplified with the three-channel approach. The re-combination of the multi-channel sCT images requires further refinement. Initial attempts at auxiliary fusion networks and other combination methods are ongoing.

Our reported image similarity metrics on the test dataset reveal promising results for the two-channel cycleGAN: a best average MAE of 62.8 ± 11.5 HU, PSNR 30.4 ± 1.6 dB and SSIM of 0.94 ± 0.02, indicate good similarity to the ground truth images, despite the test images being from an institution not included in the training set. This shows a good ability of the multi-center model to generalise. Interestingly, the model's performance in terms of image similarity on this test set surpasses results from the SynthRAD2023 challenge test set, where the three-channel model was tested on data from the same institutions for which it was trained [1, 3].

The image similarity metrics are supported by the observer assessment, where the two-channel model was rated more suitable for radiotherapy. Despite notable reductions in CBCT artefacts, such as streaking shown in Figure 3 with multi-channel models, concerns arise regarding the amplification of synthetic artefacts displaying as discontinuity in the image. This phenomenon may stem from the recombination of channels, leading to unfamiliar artefacts, thereby undermining trust in the images. Moreover, the inter-observer agreement was found to be low. The time taken to complete the evaluation ranged from 30 to 60 minutes indicating differences across observers. At present, there are no established guidelines for the evaluation or use of synthetic images in radiotherapy, and better definitions for the assessment questions may aid future evaluations. Nevertheless, qualitative assessments by experts, like radiographers, are crucial for understanding the potential of CBCT-based sCTs in enhancing adaptive radiotherapy workflows. However, to ensure robust insights, qualitative score sheets should be comprehensive, standardized and task-oriented, given the subjective nature of these evaluations. Both image similarity evaluation and qualitative assessment should also be supported by the evaluation of sCTs on downstream tasks such as dosimetric evaluation and contour propagation.

6 Conclusion

Although the global image metrics show a significant improvement in sCT quality as more information was incorporated into the training protocol, observers found all image reconstructions to be of similar quality. This highlights the need to include human observers in the assessment of image synthesis. Further improvement in combining the multi-channel output may remove the anatomical image discontinuities observed. However, we believe this is unlikely to change the observer rankings significantly.

References

[1] A. Thummerer, E. van der Bijl, A. Galapon, et al. "SynthRAD2023 Grand Challenge dataset: Generating synthetic CT for radiotherapy". *Medical Physics* 50.7 (June 2023), pp. 4664–4674.

[2] J.-Y. Zhu, T. Park, P. Isola, et al. "Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks". *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, Oct. 2017.

[3] C. A. H. Sargeant, E. G. A. Henderson, D. M. McSweeney, et al. "A multi-channel cycleGAN for CBCT to CT synthesis". *arXiv* (2023).

Evaluating Treatment Delivery Performance of SCINTIX Biology-guided Radiotherapy under non-periodic In-treatment Motions

Chenyang Shen, Thomas Banks, Sagar Ghimire, Yang Kyun Park, Andrew Godley, Steve Jiang, Bin Cai

Department of Radiation Oncology, University of Texas Southwestern Medical Center

Abstract The RefleXion SCINTIX system uniquely enables novel biology-guided radiotherapy (BgRT), providing real-time treatment guidance via on-board positron emission tomography (PET). By directly targeting tumor response via PET, SCINTIX has the potential to substantially enhance the accuracy and efficacy of radiotherapy. This technology was only recently released for clinical usage, however, so its performance when dealing with in-treatment patient motion has not been fully explored. In this paper we report a novel study to investigate the performance of SCINTIX under discrete motions using an in-house developed fluorodeoxyglucose phantom. Ion chamber and film were used to measure point dose and 2D dose, respectively, under various scenarios of different discrete motions. The results show good agreement between delivered dose and planned dose, illustrating the efficacy of SCINTIX in resolving discrete motions during treatment delivery.

1 Introduction

SCINTIX radiotherapy (RT), uniquely enabled on the RefleXion X1 RT system (RefleXion Medical Inc, Hayward, CA), is the first commercially available biology-guided radiotherapy (BgRT) treatment platform [1]. It innovatively integrates a ring-gantry LINAC with a positron emission tomography (PET) scanner to provide functional guidance for treatment delivery. With recent FDA clearance for lung and bone cancer treatment, SCINTIX offers great potential to substantially improve the treatment precision and efficacy by adjusting beam fluence based on the PET signal which is updated every 100 milliseconds during beam delivery [2, 3].

To ensure the safe integration of SCINTIX into clinical practice and maximize its potential benefits, it is imperative to thoroughly assess its performance across a diverse range of motion scenarios. This presents a distinctive commissioning challenge. Generally speaking, in-treatment patient motion can be categorized into two main types: discrete non-periodic motion (such as sudden voluntary or involuntary movements) and continuous periodic motion (such as respiratory motion). These motions may occur independently or concurrently during treatment, adding further complexity to the evaluation process. In this paper we focused on investigating the impact of various discrete motions during treatment delivery.

2 Materials and Methods

2.1. Phantom design and fabrication

To assess SCINTIX delivery robustness against in-treatment motions, we developed an in-house fluorodeoxyglucose (FDG) phantom (Fig. 1). Our objective was to fabricate a phantom capable of holding FDG solutions of varying concentrations to simulate realistic FDG distribution within a patient's body. Specifically, the phantom features a dome-shaped volume for holding FDG, representing the PET-avid treatment target, atop an insert that installs inside a cylindrical water-filled phantom. Additionally, this cylindrical phantom can be filled with FDG at a comparatively lower concentration to simulate surrounding normal tissues in a patient. As highlighted in Fig.1, the insert was designed to also hold a custom-sized Gafchromic EBT3 film for 2D dose measurement. A chamber holder accommodating an Exradin A14SL chamber (Standard Imaging, Inc, Middleton, WI) was also included to facilitate precise point dose measurement.

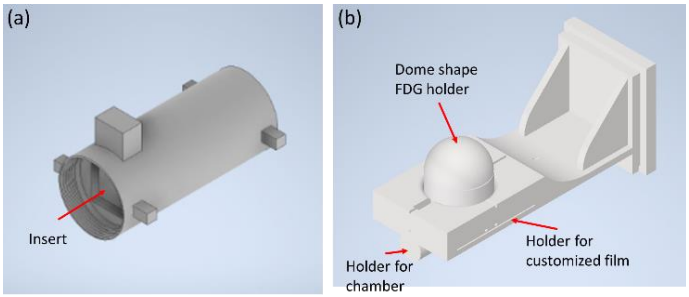


Figure 1. (a) In-house developed FDG phantom. (b) Detailed design for the insert with FDG, film, and chamber holders.

2.2 SCINTIX test plan development

We conducted our motion studies using the aforementioned in-house phantom. To evaluate the dosimetric impact of discrete motions during SCINTIX delivery, two plans were developed for ion chamber (IC) and film measurements, respectively. Specifically, a planning CT scan of the phantom was acquired on our clinical CT simulator with 1 mm in-plane resolution and 2 mm slice thickness. As shown in Fig. 2, the dome-shaped FDG holder was contoured to be the gross tumor volume (GTV) for both plans. Different expansions were then applied to the GTV to create two planning target volumes (PTVs) for the respective IC and film measurements. Note that relatively larger inferior expansions were applied to generate the PTV for the IC plan to ensure the active volume of the chamber remained within the PTV. The biological tracking zone (BTZ), a unique conceptual structure required by SCINTIX planning and treatment, was generated via a uniform expansion of 1 cm from the chamber PTV. The purpose of the BTZ is to filter the real-time PET signal for efficient data processing during treatment delivery. It is required that the PTV always stay within the BTZ during treatment delivery. Note that this BTZ generated from the IC PTV served as the BTZ for both plans since it is sufficiently large for both PTVs. In addition, two cylindrical

pseudo organs at risk (OARs) were contoured anterior and posterior left of the PTVs to make the test plan more realistic.

Prior to initiating SCINTIX planning, an on-board PET image must also be acquired on the RefleXion treatment unit. This PET image is then back-projected to the beam's eye view for each beam angle and convolved with some kernels referred to as “firing filters” to generate fluence maps. The inverse planning process is essentially the optimization process for these firing filters. In our experiment, the in-house phantom was filled with FDG at a target-to-background activity-per-unit-volume ratio of 8:1 for the pre-planning PET scan. This resulted in a target activity of approximately 150 kBq/ml for the preplan PET image, which was subsequently utilized as the planning PET for generating both ion chamber (IC) and film plans. The same experimental setup was maintained across all treatment deliveries.

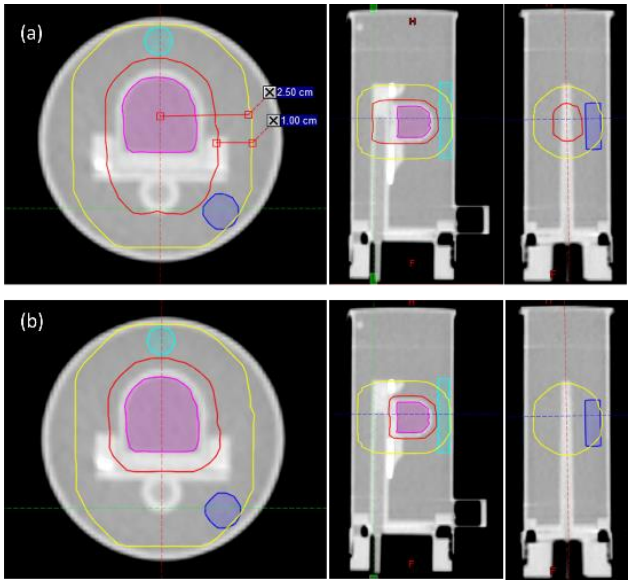


Figure 2. Three views of the planning CT of the in-house phantom for (a) the IC plan and (b) the film plan. Planning structures include the GTV (purple), PTVs (red), BTZ (yellow), and OARs (blue and cyan).

The prescription (Rx) dose to PTV for both plans was set to 100 Gy in 10 fractions with maximal dose objective set as Dmax≤105% of Rx dose. For film plan we considered both OARs and the planning goals were set as Dmax≤ 70 Gy for the anterior OAR and 80 Gy for the posterior left OAR. In contrast, for the IC plan only the anterior OAR was taken into account in the planning process, with an associated planning objective of Dmax≤50 Gy. Snapshots of the dose distributions of the optimized plans are shown in Fig 3. Note that both plans were generated for the purpose of evaluating delivery quality of SCINTIX under different discrete motions, so plan quality was not a major focus.

2.3. Static motion delivery setup

We started the test by verifying point dose accuracy with the IC plan. Specifically, measurements under four different scenarios in the plan delivery were carried out. In the first scenario, the phantom was held still after being set up to the treatment position via the on-board fan beam CT. This was to mimic BgRT delivery without any in-treatment motion. One or multiple shifts of magnitude of 0.5 cm were then applied to the phantom during treatment delivery as the second scenario. Note that the PTV was still well within the BTZ after applying these shifts. This test aimed to evaluate system performance when abrupt voluntary or involuntary movements of the patient occur during treatment delivery.

The third scenario involved shifting the phantom laterally by 2.5 cm during treatment. This was to mimic relatively larger abrupt in-treatment shifts of the patient. In this case, half of the PTV ended up outside of the BTZ, since the lateral dimension of the BTZ was 5 cm. We wanted to test if the machine could identify this issue and stop delivering dose. If the system did stop treatment, the phantom would be set up again to resume delivery, and the overall dose measured would be compared to the planned dose.

For the fourth scenario, we applied an even larger lateral shift of 4 cm during delivery so that the PTV was moved completely outside of the BTZ. This was to further understand the performance of SCINTIX plan delivery under extreme circumstances of very large shift. Similar to the third scenario, if the treatment system could identify this large motion and stop treatment delivery, we would set the phantom up to the planned treatment position again to resume treatment and compare the overall dose with the planned dose.

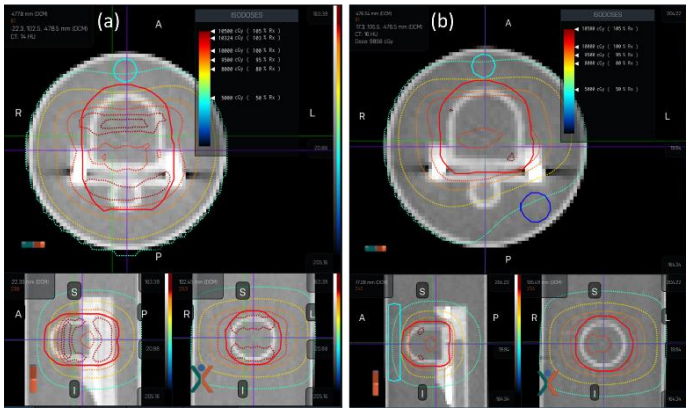


Figure 3. Snapshots of the dose distributions of the optimized (a) IC plan and (b) film plan.

The film plan was then delivered under the aforementioned second and third scenarios to further investigate 2D dose distributions for SCINTIX treatment delivery under discrete motions. We chose these two scenarios as opposed to the fourth one since they are more clinically relevant. Note that the detailed experimental setup, including the directions and magnitudes of the shifts during treatment delivery, was configured to be identical to that used in the IC plan experiments.

3 Results

3.1. Results of IC plan delivery

The point dose measured for the first scenario without in-treatment motion differed by -1.4% compared to the planned dose, illustrating the accuracy of SCINTIX treatment delivery for a static target.

For the second scenario, the phantom was shifted twice at different time points during a single treatment delivery. Specifically, the first shift was lateral, of magnitude 0.5 cm to patient left, when 25% of plan had been delivered. The second shift was superior by 0.5 cm, when 50% of the plan had been delivered. Note that the PTV remained within the BTZ defined in planning stage after these two shifts were executed. The measured point dose was ~0.6% lower than that of the static delivery (i.e. ~2% lower compared to the planned dose). These results indicate the satisfactory tracking capability of SCINTIX in terms of handling abrupt shifts of target within BTZ.

In the third scenario, we shifted the phantom by 2.5cm laterally to patient left when 25% of planned dose had been delivered. The system continued delivering beam until it reached ~35% of plan delivery, at which point an interlock was triggered due to missing target signal in BTZ. After re-setup, we were able to complete the partial delivery and the measured dose was ~4.7% higher than the static delivery (~3.3% higher than planned dose). Based on this test, SCINTIX system was able to detect a large shift and pause the delivery after part of the PTV had moved outside of the BTZ. However, the system response was not instantaneous. Close monitoring for large patient shifts by the treatment team would still be recommended.

We then performed a larger shift of 4cm during delivery to move the PTV completely out of BTZ after 25% of the plan had been delivered. The system suspended treatment after delivering another ~3% of MUs. We then re-set the phantom up and continued the plan delivery till completion. The resulting dose measured was 5% higher than the baseline from the static delivery (~3.5% higher than the planned dose). Compared to the partial-in/partial-out PTV scenario in the previous test, moving the PTV completely out of the BTZ triggered a swifter response of the SCINTIX system. However, the dosimetry results was similar. The experimental setups and the quantitative results of the IC plan are summarized in Table 1.

3.1. Results of film plan delivery

The first film test was carried out using same phantom shifts performed for the second scenario in the IC plan study. The results can be found in Fig. 4 (a). The measured dose generally agreed with the planned dose within 5% of difference according to the line profiles. A relative large decrease in measured dose can be observed in the left side of the horizontal profile. That area is close to the cutting edge of the customized film and hence may introduce a large uncertainty in measurement.

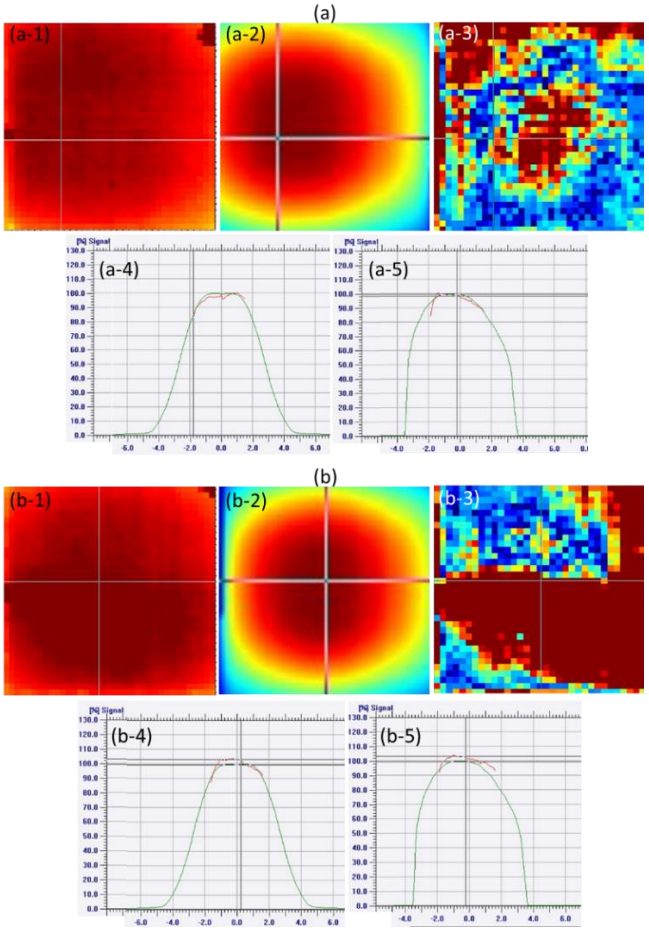


Figure 4. Film analysis results. (a): Two-step shifts with 0.5cm magnitude for each. (b): 2.5cm lateral shift. (a-1) and (b-1): scanned films; (a-2) and (b-2): planned dose; (a-3) and (b-3): gamma analysis (2%/2mm); (a-4) and (b-4): vertical line profile comparisons, (a-5) and (b-5) horizontal line profile comparisons (red: measured, green: planned).

Another film test was conducted with a 2.5 cm lateral shift applied during the treatment, identical to that used in the third scenario of the IC plan tests. The film results have revealed that SCNITIX overdelivered to the target which agrees with the behavior observed in point dose measurement. Similar to the previous film experiment, the measure dose and the planned dose generally match within 5% of difference, although 2D gamma analysis [4, 5] with 2%/2mm shows a relatively large discrepancy. It is also worth mentioning that gamma analysis is not recommended as the evaluation metric for SCINTIX delivery due to the stochastic nature of the PET signals.

4 Discussion

Based on our experimental results, the SCINTIX RT system is capable of recognizing and responding to discrete in-treatment motions. The dose delivery remained accurate for relatively small motions as long as the PTV was still within the BTZ, as evidenced by both point dose (0.6% compared to static delivery and 2% compared to planned dose) and film measurements (well-matched line profiles). When larger motions occurred, moving some or all of the

PTV outside of the BTZ, the SCINTIX system was able to detect the motion and stop the treatment. However, the system did not terminate delivery immediately after the shifts, revealing a continued need for close monitoring of patient movement by the treatment team during SCINTIX treatments. Once delivery had been stopped due to large motions, the system was able to resume treatment after we re-set up the phantom. The point dose measurements differed by 3.2% and 3.5%, respectively, for lateral shifts of 2.5cm (PTV half in/half out of BTZ) and 4cm (PTV completely out of BTZ).

This study only considered relatively simple discrete motions. What could actually happen during treatment could be more complicated than what we have considered here. For instance, in lung cancer treatments continuous periodic motion becomes the major concern. For patients with irregular breathing, abrupt changes in motion magnitude and/or pattern could also occur on top of the periodic respiratory motion. These problems are more challenging and will require further investigations. The study reported here helps us to better understand the behaviors of SCINTIX when discrete movements occur during treatments, and thus lays a solid foundation for future studies on more complex motion patterns.

The gamma analysis results in Fig. 4 depict the difference between measured and planned dose, but this is not recommended for SCINTIX testing. Note that the fluence maps for SCINTIX plan delivery are computed by convolving the back-projected PET signal acquired during treatment with the firing filters optimized on the pre-planning PET. Due to the stochastic nature of the PET signal, the delivered dose will always be different from what was planned. Therefore gamma analysis may not serve the purpose of evaluating plan delivery quality as well as it does for conventional RT. More research is needed to develop a proper metric for 2D dosimetric evaluation of SCINTIX plan delivery quality.

5 Conclusion

We have successfully conducted a discrete motion study to evaluate the performance of SCINTIX RT delivery using an in-house-developed FDG phantom. The results illustrate the efficacy of SCINTIX in recognizing and responding to discrete in-treatment motions to achieve robust delivery.

References

1. Oderinde, O.M., et al., *The technical design and concept of a PET/CT linac for biology-guided radiotherapy.* Clinical and Translational Radiation Oncology, 2021. **29**: p. 106-112.
2. Shirvani, S.M., et al., *Biology-guided radiotherapy: redefining the role of radiotherapy in metastatic cancer.* The British Journal of Radiology, 2020. **94**(1117): p. 20200873.
3. Vitzthum, L.K., et al., *BIOGUIDE-X: A First-in-Human Study of the Performance of Positron Emission Tomography–Guided Radiation Therapy.* International Journal of Radiation Oncology*Biology*Physics, 2023.
4. Low, D.A., et al., *A technique for the quantitative evaluation of dose distributions.* Medical Physics, 1998. **25**(5): p. 656-661.
5. Miften, M., et al., *Tolerance limits and methodologies for IMRT measurement-based verification QA: Recommendations of AAPM Task Group No. 218.* Medical Physics, 2018. **45**(4): p. e53-e83.

	Motion info	Difference to static	Difference to plan
Scenario 1	NA	NA	-1.46%
Scenario 2	0.5cm to left @ 25% of delivery; 0.5cm superiorly @ 50% of delivery	-0.6%	-2.0%
Scenario 3	2.5 cm to left @ 25% of delivery	4.7%	3.3%
Scenario 4	4cm to left @ 25% of delivery	5.0%	3.6%

Table 1. Summary of point dose measurement results

Detection and correction of rigid patient motion in SPECT with exponential data consistency conditions

My Hoang Hoa Bui[1], Antoine Robert[1], Simon Rit[1], and Ane Etxebeste[1]

[1]INSA-Lyon, Université Claude Bernard Lyon 1, CNRS, Inserm, CREATIS UMR 5220, U1294, F-69373, Lyon, France

Abstract Targeted radionuclide therapy has gained an increasingly important role in cancer treatment in recent years, in which molecular imaging techniques such as SPECT can be employed to plan and monitor the treatment. However, rigid patient motion is degrading the quality of SPECT images and may impact the treatment planning. It has been recently proposed to use exponential data consistency conditions — equations describing the redundancy of exponential SPECT measurements — to detect and correct for patient motion. This study aimed at developing a data consistency-based method to estimate and correct for rigid patient motion during SPECT acquisitions. The proposed method was evaluated on Monte Carlo simulated data of a liver patient. The method led to a significant improvement in the image quality. An activity recovery coefficient above 95% with respect to the reconstructed images from motion-free projections was obtained across all tested cases. In particular, for a motion in the middle of the acquisition, the activity recovery coefficient was improved from 66 % (non-corrected projections) to 99% (motion-corrected projections).

1 Introduction

Targeted radionuclide therapy is a favorable option in cancer treatment thanks to its ability to selectively deliver radiation to cancerous cells, thus minimizing toxicity on surrounding healthy tissues [1]. Functional imaging such as single photon emission computed tomography (SPECT) can be used to optimize the treatment plan, i.e. to adjust the amount of injected radionuclides [2–4].

After injecting the patient, the SPECT scanner collects projections of the emission map with one or several rotating gamma cameras. However, data acquisition requires a long time, typically 5 to 40 minutes. During this time, rigid patient motion is likely to occur [5], potentially leading to artifacts in reconstructed images [6].

Exponential data consistency conditions (eDCCs) are mathematical equations that describe redundancies among measurements modeled by a linear operator called the exponential Radon transform. Recently, Wells and Clackdoyle showed the feasibility of using eDCCs to align the attenuation map to the emission map in pinhole cardiac SPECT [7]. In [8], Robert *et al* studied the ability of eDCCs to detect motion on realistic simulated data in parallel SPECT.

To our knowledge, motion compensation in SPECT exploiting eDCCs exclusively has not been studied yet. This study aimed at employing eDCCs for automated estimation and correction of patient motion, requiring solely the acquired projections and the attenuation map of the patient.

2 Materials and method

2.1 Exponential data consistency conditions

The exponential Radon transform is a two-dimensional (2D) operator and eDCCs can be expressed in 2D. Throughout this study, we express them in a three-dimensional (3D) space for deriving the correction of 3D motions but their 2D expression can be obtained by dropping the third coordinate.

Consider a 3D radioactivity distribution f and an attenuating medium μ. The ideal measurement of f in parallel SPECT geometry can be modeled by the attenuated Radon transform:

$$p_\mu(\theta,s,z) = \int_{-\infty}^{+\infty} f(s\mathbf{u}_\theta + t\mathbf{v}_\theta + z\mathbf{e}_z) \exp\left(-\int_t^{+\infty} \mu(s\mathbf{u}_\theta + t'\mathbf{v}_\theta + z\mathbf{e}_z)\,\mathrm{d}t'\right)\mathrm{d}t \quad (1)$$

where $\mathbf{u}_\theta = (\cos\theta, \sin\theta, 0)$, $\mathbf{v}_\theta = (-\sin\theta, \cos\theta, 0)$ and $\mathbf{e}_z = (0,0,1)$.

Assuming that there is a known convex region K where the attenuation is constant, i.e. $\mu(x,y,z) = \mu_0$ if $(x,y,z) \in K$, and all radioactivity is encompassed inside K, the exponential projections $p(\theta,s,z)$ can be derived from $p_\mu(\theta,s,z)$ by a point-wise conversion, as:

$$p(\theta,s,z) = p_\mu(\theta,s,z)C(\theta,s,z). \quad (2)$$

The conversion factor $C(\theta,s,z)$ was introduced in [9]

$$C(\theta,s,z) = \exp\left(\tau_{\theta,s,z}\mu_0 + \int_{\tau_{\theta,s,z}}^{+\infty} \mu(s\mathbf{u}_\theta + t'\mathbf{v}_\theta + z\mathbf{e}_z)\,\mathrm{d}t'\right) \quad (3)$$

with $\tau_{\theta,s,z}$ the point along the integral line towards the detector where the photon leaves the convex region K.

Range conditions of the exponential Radon transform, known as eDCCs, state that two projections at θ_i and θ_j which are not 180° apart ($\theta_i - \theta_j \neq \pi\,[2\pi]$) are consistent if and only if

$$P(\theta_i, \sigma_{ij}, z) = P(\theta_j, \sigma_{ji}, z) \quad (4)$$

with $P(\theta,\sigma,z) = \int_{-\infty}^{+\infty} p(\theta,s,z)e^{\sigma s}\,\mathrm{d}s$ the two-sided Laplace transform of projection line $p(\theta,\cdot,z)$ and $\sigma_{ij} = -\sigma_{ji} = \mu_0 \tan((\theta_i - \theta_j)/2)$.

2.2 Motion simulation

Using a validated GATE model of the Symbia Intevo Bold scanner [10, 11], we simulated a liver radioembolization patient using a CT image and a voxelized sphere source of 20 mm-radius located inside the liver with a ^{99m}Tc activity of 150 MBq. Data were acquired within a primary energy window of 15% centered around the main photopeak of 140.5 keV and a lower scatter window of 15% centered around 120 keV — later used to correct scatter from acquired primary projections. Several sets of projections were generated, each consisting of 60 projections covering 360°, an acquisition time of 15 s per projection with 256 × 256 pixels of 2.39 mm spacing. The SPECT camera followed a circular orbit with a detector-to-center distance of 38 cm.
In this study, we considered a single rigid patient motion between two successive projections of the acquisition. The rigid patient motion was simulated by shifting simultaneously the attenuation and emission maps. One set of projections was simulated without motion and used as a reference. The performance of the motion detection and correction method was evaluated regarding:

1. Motion index: with a motion vector set to $(20,20,20)$ mm, all possible values of motion index (defined as the first frame index affected by motion), were evaluated.
2. Motion magnitude: with a motion index fixed at the middle of the acquisition, several motions were evaluated: 1, 3, 5, 10, 20 and 30 mm in all three axes.

2.3 eDCCs assessment

To replicate a clinical scenario with the CT image acquired only once before the SPECT acquisition, we used the attenuation map placed at the original position to compute eDCCs. We defined the convex K region covering the whole liver (created with auto segmentation of 3D Slicer on the CT image). To evaluate the consistency of the i-th and j-th SPECT projections p_μ with N detector lines in the axial direction, we employed the noise-aware metric [8, 12]

$$\mathrm{NE}_{p_\mu}(i,j) = \frac{\frac{1}{N}\sum_{l=1}^{N} |P(\theta_i,\sigma_{ij},z_l) - P(\theta_j,\sigma_{ji},z_l)|}{\sqrt{N}\sqrt{\mathrm{Var}(\overline{Pij}) + \mathrm{Var}(\overline{Pji})}} \quad (5)$$

with z_l the discretized z coordinate of the l-th line of the detector, the variance

$$\mathrm{Var}(\overline{P_{ij}}) = \frac{(\Delta s)^2}{N^2} \sum_{l=1}^{N}\sum_{k=1}^{M} C_l^2(\theta_i,s_k,z_l)\, p_\mu(\theta_i,s_k,z_l)\, e^{2\sigma_{ij}s_k} \quad (6)$$

Δs the pixel spacing and M the number of pixels per projection line.
According to findings in [8], rigid patient motion is the main contributor to projection inconsistency, i.e. high $\mathrm{NE}_{p_\mu}(i,j)$, whereas the physical effects have almost negligible (Poisson noise and collimator resolution) or moderate (scatter) impact on eDCCs.

To mitigate the effect of scatter on the motion estimation model, we applied scatter correction with the dual energy window (DEW) with a k factor of 1.1. The value of k was chosen to minimize the inconsistency in the reference motion-free acquisition. Then, a mask, defined as the forward projector of the K region, was applied to scatter-corrected frames after DEW to remove the remaining scattering from outside of the convex region.

2.4 Motion correction

If the patient's displacement is known, motion-free projections can be calculated by compensating for the patient's motion in the projection domain. When both the emission and attenuation maps are shifted by a vector in a transverse plane, the corresponding projections are shifted by a translation along the s-axis. Motion in an orthogonal direction, along the rotation axis z, can be directly interpreted as the same shift along the axial coordinate. Together, any 3D motion shift $\mathbf{a}$ can be corrected in the attenuated projection at a given detector angular position θ by a translation of the projection,

$$p_\mu^{\mathbf{a}}(\theta,s,z) = p_\mu(\theta, s+\mathbf{u}_\theta\cdot\mathbf{a}, z+\mathbf{e}_z\cdot\mathbf{a}). \quad (7)$$

2.5 Motion estimation

Let's define motion index m_0 as the first frame index affected by motion. m_0 divides the acquisition into no-motion projections $p_\mu(\theta_i,\cdot)$ for $i<m_0$ and motion-affected projections $p_\mu(\theta_j,\cdot)$ for $j\geq m_0$. Pairing projections in the second set with projections in the first set should result in non-zero eDCCs. In contrast, if these motion-affected projections $p_\mu(\theta_j,\cdot)$ are corrected by the true shift $\mathbf{a}_0$, the eDCCs should be closer to zero. To estimate m_0 and the shift $\mathbf{a_0}$, we therefore minimized the inconsistency between pairs of motion-corrected projections $p_\mu^{\mathbf{a}}$ of the two sets by solving the minimization problem

$$(\hat{m},\hat{\mathbf{a}}) = \underset{m\in\Omega_m,\,\mathbf{a}\in\mathbb{R}^3}{\operatorname{argmin}} F(m,\mathbf{a}) \quad (8)$$

with $\Omega_m = [2,L]$, L the number of projections within the acquisition, the cost function

$$F(m,\mathbf{a}) = \frac{1}{K}\sum_{i<m,j\geq m} \mathrm{NE}_{p_\mu^{\mathbf{a}}}(i,j) \quad (9)$$

and K the number of evaluated projection pairs (fulfilling $i<m\leq j$ and $|\theta_i-\theta_j|\neq 180°$).
The problem was solved by first calculating

$$\tilde{\mathbf{a}}_m = \underset{\mathbf{a}\in\mathbb{R}^3}{\operatorname{argmin}} F(m,\mathbf{a}) \quad (10)$$

for every $m\in\Omega_m$ with the Nelder-Mead downhill simplex algorithm [13] and then taking the value m that minimizes F, i.e., $\hat{\mathbf{a}} = \tilde{\mathbf{a}}_{\hat{m}}$ with

$$\hat{m} = \underset{m\in\Omega_m}{\operatorname{argmin}} F(m,\tilde{\mathbf{a}}_m). \quad (11)$$

2.6 Quantitative assessment of motion correction

The estimated motion shift was compared to the simulated value. For each projection set, the optimization was repeated 20 times with random initializations to estimate the robustness of the estimation.
The effect of motion correction on reconstructed SPECT images was visually assessed with a (20,20,20) mm shift occurring in the middle of the acquisition. The estimated motion was taken as the median of 20 random initializations. The motion-free projections, non-corrected projections and motion-corrected projections were reconstructed by the ordered subset expectation maximization (OSEM) implemented in RTK [14] using 5 iterations and 4 subsets with attenuation correction. The quality of reconstructed images was assessed via the activity recovery coefficient (ARC) in reference of the reconstructed SPECT images from motion-free projections

$$\mathrm{ARC}(\%) = \frac{A}{A_{\text{motion-free}}} \times 100\%. \qquad (12)$$

The values A and $A_{\text{motion-free}}$ are the estimated activities in the source region obtained respectively from evaluated projections (non-corrected or motion-corrected projections) and from motion-free projections taken as reference.

3 Results

The ability to detect a $(20,20,20)$ mm translation appearing at different motion indices of the acquisition is shown in Fig. 1. Plot A indicates the values of the cost function after optimization for every position $m \in [2,60]$, i.e., $F(m,\tilde{\mathbf{a}}_m)$, while B shows the difference between the estimated $\hat{m}$ and the true motion index m_0 for all possible $m_0 \in \Omega_m$ with two selected examples of motion magnitude. For a 20 mm motion per axis (blue curve in Fig. 1B), the motion index was adequately estimated. For 3D motions below 5 mm per axis, for instance the black curve in Fig. 1B, $\hat{m} - m_0$ was not zero, i.e., the estimation of the motion index was incorrect for several values of m_0.

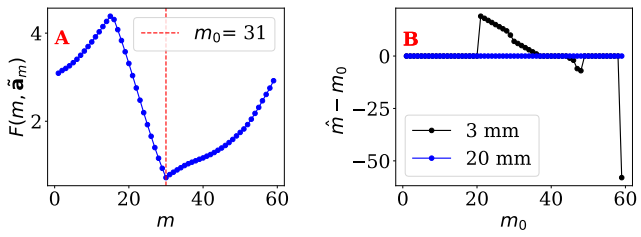


Figure 1: A. Illustration of motion detection with sliding value $m \in [2,60]$ for (20,20,20) mm displacement in the middle of acquisition; B. Performance of motion detection for different motion indices m_0 with respectively (20,20,20) mm shift (blue) and (3,3,3) mm (black).

Fig. 2 illustrates the estimation of (20,20,20) mm translation occurring at different motion index m_0 along the acquisition. The proposed method returned promising accuracy where the absolute differences were close to or below the pixel size of 2.39 mm for each direction. There was also a significant precision at the sub-millimeter level across different repetitions of the optimizer, with only a minor number of outliers.

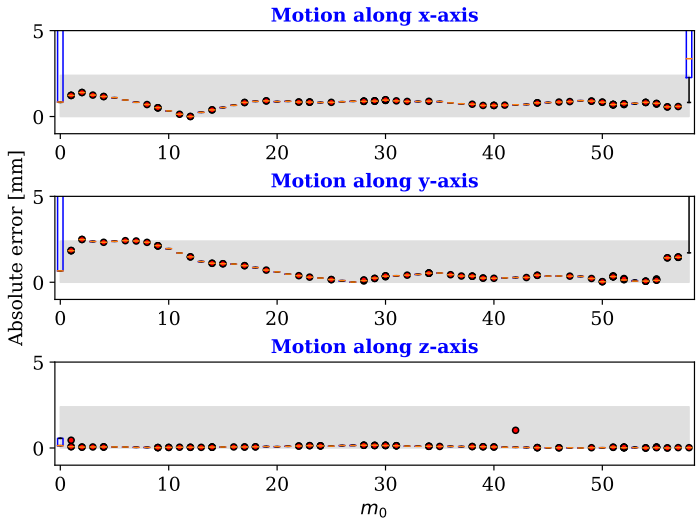


Figure 2: Box and whisker plot (blue) for absolute error of motion estimation at different motion indices m_0 for $\mathbf{a}_0 = (20,20,20)$ mm. Horizontal red lines inside the box represent the median value and circular red dots are outliers with high bias, i.e. values greater than 1.5 inter-quartile range.

In general, motion assessment along z was highly accurate with an error below 1 mm. The mean deviation for all values of motion index m_0 was around 0.05 mm in this axis compared to respectively 1.4 and 2.2 mm in x and y. Also, misestimation in x and y were generally synchronous.
For smaller shifts below 5 mm, the absolute relative errors were larger (data not shown), e.g., up to 10 to 50 % compared to the true values for mid-acquisition motion. When the shift is above 5 mm per axis, these errors are below 10% for all three components. These findings are consistent with the results above.
The ARC before and after motion correction is shown side-by-side in Fig. 3. Without correction, the values of ARC spread in a wide range from 15% to nearly 100%. In contrast, motion correction brought a comparable ARC to motion-free images, i.e. less than a 5% drop for motion occurring early in the acquisition and below 1% for motion occurring after one-third of the acquisition.
The improvement in the reconstructed image is illustrated in Fig. 4. Motion correction greatly mitigated the artifacts, namely blurring and distortion of the sphere.

4 Discussion

In this study, eDCCs were exploited to detect and correct for rigid patient motion in SPECT projections simulated with Monte Carlo simulations using a realistic SPECT scanner model.

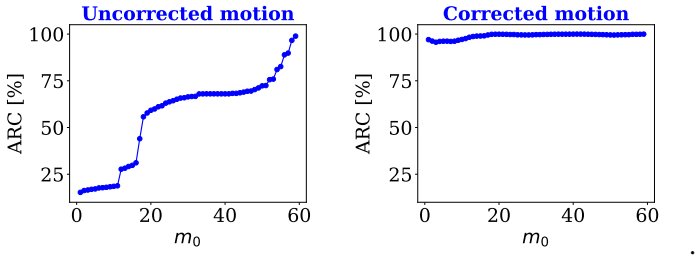


Figure 3: Activity recovery coefficient (ARC) as a function of motion index m_0 for $\mathbf{a}_0 = (20, 20, 20)$ mm.

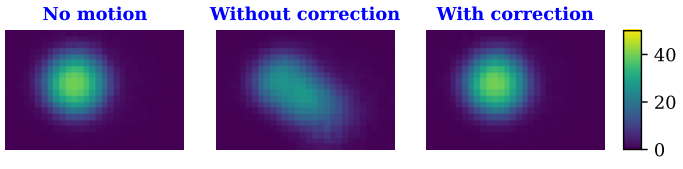


Figure 4: Visualization of reconstruction with motion-free, non-corrected, and motion-corrected projections for $m_0 = 31$ and $\mathbf{a}_0 = (20, 20, 20)$ mm.

The current two-step procedure first showed the reliability of detecting sudden inter-frame translation of (20,20,20) mm happening between any two successive frames of the acquisition, i.e. different possibilities of m_0. For the sliding value of $m \in [2, 60]$, frames after m were corrected simultaneously with the same shift to optimize projection consistency. If $m \neq m_0$, some motion-free ($p(\theta_i, \cdot)$ for $i < m_0$) or non-corrected ($p(\theta_j, \cdot)$ for $j \geq m_0$) projections are wrongly motion-corrected or non-corrected, respectively, and both raised the inconsistency in the cost function (Fig. 1A).
For a known motion index m_0, the motion correction performed on projections from Monte Carlo simulation already achieved quite good agreement with true values. In general, estimation on the z-component, which was independent of motion index m_0, surpassed those in the other two directions in terms of accuracy. Indeed, translation in z is more sensitive to eDCCs than in the other two directions since it directly correlates to the same translation for all angular positions of the detector θ_m.
In good agreement with the findings from Robert *et al.* [8], we observed that the collimator resolution and Poisson noise had a minor impact on the correction of motion with eDCCs using scatter-free simulations obtained by RTK ray tracing modeling these two effects [14] (data not shown). On the contrary, the presence of scatter in the Monte Carlo simulation noticeably degraded the accuracy of our proposed correction method when no scatter correction was performed (data not shown).

5 Conclusions

Rigid patient motion in SPECT has been detected and corrected with an algorithm exploiting exponential data consistency conditions. A two-step process was shown to be practical for SPECT measurements simulating all physical effects which is promising for future application to real scanner data.

References

[1] T. Wheldon and J. O'Donoghue. "The Radiobiology of Targeted Radiotherapy". *International Journal of Radiation Biology* 58.1 (1990). PMID: 1973428, pp. 1–21. DOI: 10.1080/09553009014551401.

[2] A. K. Buck, S. Nekolla, S. Ziegler, et al. "SPECT/CT". *Journal of Nuclear Medicine* 49.8 (2008), pp. 1305–1319. DOI: 10.2967/jnumed.107.050195.

[3] J. Imbrescia, G. Volpi, S. Lucchini, et al. "A Brief Report on the Role of SPECT/TC in the Optimization of Radiotherapy Treatment with Radical Intent for Unresectable Stage III NSCLC". *Applied Sciences* 12.18 (2022). DOI: 10.3390/app12189351.

[4] J. A. Christian, M. Partridge, E. Nioutsikou, et al. "The incorporation of SPECT functional lung imaging into inverse radiotherapy planning for non-small cell lung cancer". *Radiotherapy and Oncology* 77.3 (2005), pp. 271–277. DOI: https://doi.org/10.1016/j.radonc.2005.08.008.

[5] J. Wheat and G. Currie. "Incidence and characterization of patient motion in myocardial perfusion SPECT: part 1". English. *Journal of Nuclear Medicine Technology* 32.2 (2004), pp. 60–65.

[6] A. Z. Kyme and R. R. Fulton. "Motion estimation and correction in SPECT, PET and CT". *Physics in Medicine & Biology* 66.18 (Sept. 2021), 18TR02. DOI: 10.1088/1361-6560/ac093b.

[7] R. G. Wells and R. Clackdoyle. "Feasibility of attenuation map alignment in pinhole cardiac SPECT using exponential data consistency conditions". *Medical Physics* 48.9 (2021), pp. 4955–4965. DOI: https://doi.org/10.1002/mp.15058.

[8] A. Robert, D. Sarrut, A. Etxebeste, et al. "Ability of exponential data consistency conditions to detect motion in SPECT despite other physical effects". *17th International Meeting on Fully 3D Image Reconstruction in Radiology and Nuclear Medicine, Stony Brook, NY, USA* (2023).

[9] F. Natterer. *The Mathematics of Computerized Tomography*. NY: John Wiley & Sons, 1986.

[10] D. Sarrut et al. "Advanced Monte Carlo simulations of emission tomography imaging systems with GATE". *Phys Med Biol* 66.10 (May 2021), 10TR03. DOI: 10.1088/1361-6560/abf276.

[11] M. Bui, A. Robert, J. Badel, et al. "Validation of a model of the Symbia Intevo Bold SPECT scanner for Monte Carlo simulations". *2023 IEEE Nuclear Science Symposium, Medical Imaging Conference and International Symposium on Room-Temperature Semiconductor Detectors (NSS MIC RTSD)*. 2023, pp. 1–1. DOI: 10.1109/NSSMICRTSD49126.2023.10338340.

[12] M. Mouchet, J. M. Létang, J. Lesaint, et al. "Cone-beam pair-wise data consistency conditions in helical CT". *IEEE Transactions on Medical Imaging* (2023), pp. 1–1. DOI: 10.1109/TMI.2023.3265812.

[13] J. A. Nelder and R. Mead. "A Simplex Method for Function Minimization". *The Computer Journal* 7.4 (Jan. 1965), pp. 308–313. DOI: 10.1093/comjnl/7.4.308.

[14] S. Rit, M. Vila Oliva, S. Brousmiche, et al. "The Reconstruction Toolkit (RTK), an open-source cone-beam CT reconstruction toolkit based on the Insight Toolkit (ITK)". *Journal of Physics: Conference Series* 489 (Feb. 2014). DOI: 10.1088/1742-6596/489/1/012079.

Preliminary Study on Residual Motion Evaluation During Breath Hold SBRT

Weihua Mao and Amol Narang

Department of Radiation Oncology and Molecular Radiation Sciences, Johns Hopkins University, Baltimore, Maryland, USA

Abstract Breath hold has been widely used to limit patient respiratory motion for radiation therapy, however residual motions have been reported by many groups. Fiducial markers have been implanted to locate the target for pancreatic cancer treatment at our institution. A new method is presented to locate fiducial positions on every projection image during cone-beam computed tomography (CBCT) scans for patient setup or positioning verification. Data of three pancreatic cancer patients were studied retrospectively. Every patient exhibited residual motions during breath hold maneuvers, with an average drifting range of 6.5 mm observed across 34 CBCT scans. Residual motions within a single CBCT scan ranged up to 10.6 mm, and significant drifting occurred between breath holds. This method does not need any additional equipment or deliver any extra radiation doses to patients. It provides important information on daily residual motions and is invaluable for planning margin adjustments towards online adaptive radiation therapy.

1 Introduction

A key factor in treating pancreatic cancer with stereotactic body radiation therapy (SBRT) is how to manage the motion due to respiration during radiation therapy. At our institution, radio-opaque fiducial markers have been placed endoscopically before radiation therapy. All simulation, patient setup, and beam delivery have been under deep inspiration breath hold (DIBH). However, it has been reported that residual motions occur under breathing holding [1-6]. A novel, cost-effective method is presented to evaluate daily fiducial residual motions during cone-beam computed tomography (CBCT) acquisition.

2 Materials and Methods

Three pancreatic cancer patients treated in the past two months with SBRT were enrolled into this retrospective study under an internal review board protocol (IRB00395200). Two to three fiducial markers were implanted endoscopically in the pancreas for each patient. These cylindrical-shaped fiducials have a diameter of 0.5 mm and a length of 3 mm. Patients were immobilized using Vak-lok (CIVCO Medical Solutions, Coralville, IA, USA) with a wingboard (CIVCO Medical Solutions, Coralville, IA, USA). Active breathing coordination (ABC, Elekta, Stockholm, Sweden) was employed for DIBH. Patients were coached to breath in and hold their breath. Four computed tomography (CT) simulation scans were repeatedly acquired under DIBH using the Brilliance Big Bore CT simulators (Philips Medical Systems, Cleveland, OH, USA). One set of CT images was selected for contouring and treatment planning in RayStation (RaySearch Laboratories AB, Stockholm, Sweden). Slice thickness of CT images was 2 mm while planar pixel size was around 1.4 mm. Fiducials were contoured in the bone window ([450 1600] HU). Volumetric modulated arc therapy (VMAT) with two arcs were planned with prescriptions of 33 to 40 Gy in 5 fractions. Patients were treated by a VersaHD (Elekta, Stockholm, Sweden) linear accelerator. Partial arc (200 degrees) rotation CBCT were used with settings of 120 kV, 40 mA and 40 ms. The source-to-axis distance (SAD) was 1000 mm, the source-to-imager distance (SID) was 1536 mm, the projection image resolution was 512 x 512, and the pixel size was 0.8 mm. Maximum breath hold (BH) durations were 25 seconds for each patient. Each CBCT scan was acquired in two consecutive breath holds (BHs). Reconstructed CBCT images were rigidly registered to the planning CT based on bony structures and finalized using fiducial markers. After multiple reviews, six dimensional corrections were sent to the Hexapod (Elekta, Stockholm, Sweden) to adjust treatment couch. Verification CBCT scans were acquired immediately and/or during the middle of treatment. Treatment beams were delivered under identical DIBH.

Planning CT and structures, CBCT projection data, and ABC results were exported and de-identified from clinical systems. All analyses were performed using Matlab programs (MathWorks, Natick, Massachusetts). Fiducial mass centers were calculated within fiducial structures as fiducial positions in planning CT. Fiducial positions were compared within four sets of simulation CT by registering the planning CT to other three sets of CT focusing on fiducials.

Figure 1 illustrates both linac and imaging coordinate systems. The origin is set at the isocenter. X, Y, and Z axis directions represent the horizontal (lateral), vertical (posterior-anterior), and longitudinal directions, respectively. Projection images are in a two-dimensional (2D) coordinate system (u, v), whose origin is at the center of the imager, which rotates with gantry rotation. Equations (1) and (2) predict the projected location (u, v) of a fiducial at (x, y, z) [7].

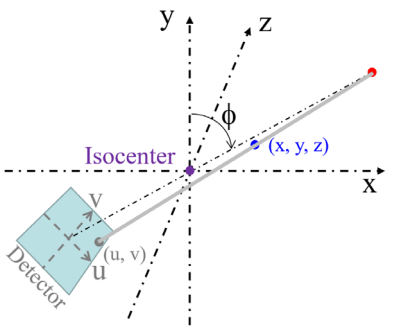


Figure 1: Schematic diagram of the coordinate systems.

$$u = \frac{\mathrm{SID}\cdot(\mathrm{x}\cdot\cos\phi - \mathrm{y}\cdot\sin\phi)}{\mathrm{SAD} - \mathrm{x}\cdot\sin\phi - \mathrm{y}\cdot\cos\phi} \approx \frac{\mathrm{SID}}{\mathrm{SAD}}(\mathrm{x}\cdot\cos\phi - \mathrm{y}\cdot\sin\phi) \quad (1)$$

$$v = \frac{\mathrm{SID}\cdot\mathrm{z}}{\mathrm{SAD} - \mathrm{x}\cdot\sin\phi - \mathrm{y}\cdot\cos\phi} \approx \frac{\mathrm{SID}}{\mathrm{SAD}}\mathrm{z} \quad (2)$$

Projection image analyses focused on regions around fiducial projection locations estimated using Eqs. (1) and (2) for every gantry angle, as shown in Figure 2(a). Relative positions of all fiducials are calculated and used as a pattern (Figure 2(b)) to match and locate the fiducials. Gradients or ridge heights were calculated in four directions (horizontal, vertical, and two oblique directions). Since the thickness of fiducial projection is only one or two pixels, gradients from two sides of a width of one or two pixels were screened by a preset threshold. The square roots of their products were used as heights at every direction (Figures 2(e) – (l)). Final heights were the summation of heights in all directions (Figure 2(c)). They were smeared by a simple average (Figure 2(d)) before matching the pattern (Figure 2(b)).

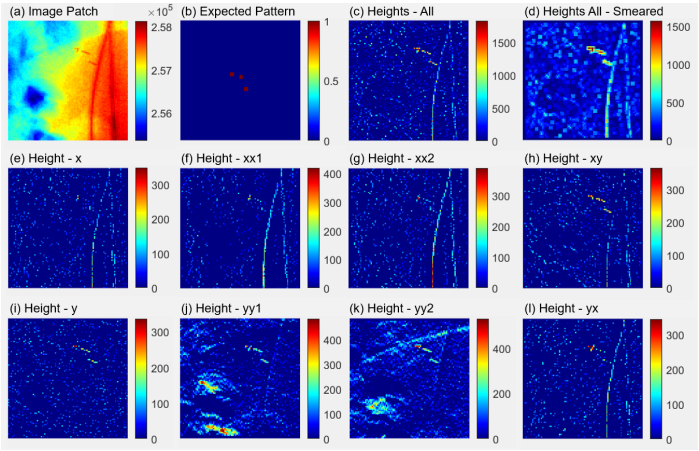


Figure 2. Imaging processing example.

A pre-analysis was performed on 2D fiducial positions to remove outliers based on expected position trendlines in every BH. Three-dimensional (3D) positions were then calculated using projections acquired previous gantry angle 45 degrees apart for two breath holds separately. The fiducial longitudinal positions were calculated in every projection based on Eq. (2). Residual motions in the longitudinal direction were assessed directly.

3 Results

It was found that the fiducials may be off by 3 slices (6 mm) among four sets of simulation CT images acquired under DIBH. CBCT projections from 8 to 16 acquisitions per patient were successfully analyzed. In total, 6697 frames of 2D projection images were acquired and fiducials were successfully detected on 6474 frames of images. 3D positions at two BHs were compared for every scan and their statistic differences on longitudinal positions are listed in Table 1.

Residual motions in the longitudinal direction were observed. Figure 3 illustrates average fiducial longitudinal positions during two BHs in a CBCT scan. Actual fiducial locations are compared in Figure 4 as an example. Significant drift occurred during the same breath hold. Four key positions were defined as the Start and End of each BH.

The drifts in the longitudinal direction were calculated for every scan. The drifting range in every BH and between two BHs was also calculated. Statistic results for every patient are listed in Table 2.

Table 1. List of average fiducial longitudinal position differences between two breath holds from 3D results.

Patients	Scan #	Min	Max	Avg	Stdev
PA01	10	-2.7	0.5	-0.7	1.2
PA02	16	-1.5	2.9	0.3	1.1
PA03	8	-0.2	0.6	0.3	0.3

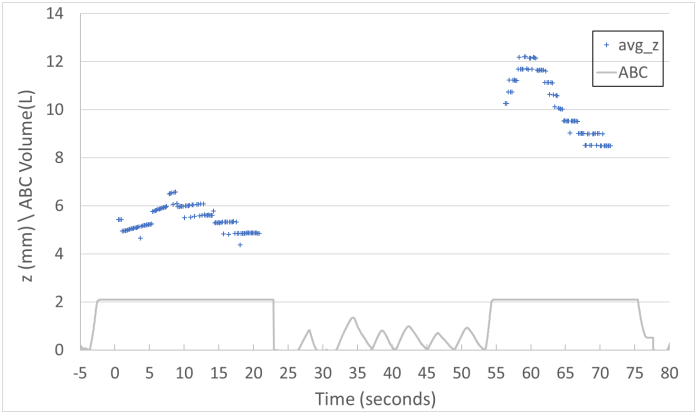


Figure 3. Average fiducial longitudinal positions (z) as functions of time for patient PA01 at scan 4a. ABC signals are illustrated.

Table 2: List of average fiducial longitudinal position differences from 2D results. The E-S of 1st BH and the E-S of 2nd BH results are the statistic differences between the end and start of the 1st BH and 2nd BH, respectively; the Between BHs results are the statistic differences between two BHs (unit: mm).

Patients	Period	Min	Max	Avg
PA01	E-S of 1st BH	-5.5	-0.1	-2.4
	Between BHs	-1.6	7.5	2.6
	E-S of 2nd BH	-4.7	-1.0	-2.7
	Range of 1st BH	2.3	5.4	3.7
	Range of 2nd BH	1.6	5.6	3.4
	Range of 2 BHs	3.1	8.9	5.6
PA02	E-S of 1st BH	-5.2	2.5	-2.9
	Between BHs	0.2	6.8	2.9
	E-S of 2nd BH	-6.2	0.9	-1.8
	Range of 1st BH	3.0	6.1	4.5
	Range of 2nd BH	1.6	6.8	3.0
	Range of 2 BHs	3.1	7.8	5.2
PA03	E-S of 1st BH	-2.5	1.8	-0.6
	Between BHs	-5.9	1.6	-1.6
	E-S of 2nd BH	0.0	5.1	1.9
	Range of 1st BH	4.7	8.9	6.7
	Range of 2nd BH	5.2	8.3	7.2
	Range of 2 BHs	7.3	10.9	8.6

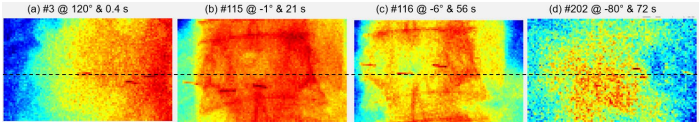


Figure 4. Comparison of fiducial locations at Start and End of each BH for Pt PA01 at scan 4a.

4 Discussion

The projection prediction equations have been simplified. For any fiducials within 20 mm from the isocenter, this may result in an error of 1 pixel on the projection, which is acceptable for this analysis.

Residual motions were observed during multiple simulation CT due to faster CT scan acquisition than CBCT acquisition. CBCT volumetric image reconstruction would smear residual motions. The average drifting during two breath holds were 6.5 mm while the maximum drifts were up to 10.9 mm. Our results also indicate that significant part of the drifts occurred between breath holds.

This is only a preliminary study, with only three patients enrolled. Further studies will continue with more patients enrolled in the near future. This method provides a free assessment of the residual motion during breath hold since it doesn't require any additional equipment or procedures and does not deliver any extra radiation doses to patients. This information will be very helpful for adjusting planning margin for daily adaptive planning.

5 Conclusion

Fiducial residual motions under breath hold could be assessed daily before radiation therapy without any extra cost.

References

[1] D. Sarrut *et al.*, "Nonrigid registration method to assess reproducibility of breath-holding with ABC in lung cancer," (in eng), *Int J Radiat Oncol Biol Phys,* vol. 61, no. 2, pp. 594-607, Feb 01 2005, doi: 10.1016/j.ijrobp.2004.08.007.

[2] C. M. Lutz, P. R. Poulsen, W. Fledelius, B. V. Offersen, and M. S. Thomsen, "Setup error and motion during deep inspiration breath-hold breast radiotherapy measured with continuous portal imaging," (in eng), *Acta Oncol,* vol. 55, no. 2, pp. 193-200, 2016, doi: 10.3109/0284186X.2015.1045625.

[3] M. Blessing *et al.*, "An offline technique to evaluate residual motion of the diaphragm during deep inspiratory breath-hold from cone-beam CT datasets," (in eng), *Strahlenther Onkol,* vol. 194, no. 9, pp. 855-860, Sep 2018, doi: 10.1007/s00066-018-1313-3.

[4] L. Vogel *et al.*, "Intra-breath-hold residual motion of image-guided DIBH liver-SBRT: An estimation by ultrasound-based monitoring correlated with diaphragm position in CBCT," (in eng), *Radiother Oncol,* vol. 129, no. 3, pp. 441-448, Dec 2018, doi: 10.1016/j.radonc.2018.07.007.

[5] J. Boda-Heggemann *et al.*, "Ultrasound-based repositioning and real-time monitoring for abdominal SBRT in DIBH," (in eng), *Phys Med,* vol. 65, pp. 46-52, Sep 2019, doi: 10.1016/j.ejmp.2019.08.003.

[6] C. Zeng *et al.*, "Intrafraction tumor motion during deep inspiration breath hold pancreatic cancer treatment," (in eng), *J Appl Clin Med Phys,* vol. 20, no. 5, pp. 37-43, May 2019, doi: 10.1002/acm2.12577.

[7] W. Mao, R. D. Wiersma, and L. Xing, "Fast internal marker tracking algorithm for onboard MV and kV imaging systems," (in eng), *Med Phys,* vol. 35, no. 5, pp. 1942-9, May 2008, doi: 10.1118/1.2905225.

Super Resolution Dose Calculation Via Implicit Neural Representation Based Blind Deconvolution

Siqi Ye, Sheng Liu, Yu Gao, and Lei Xing

Department of Radiation Oncology, Stanford University, Stanford, CA, USA

Abstract In radiation therapy, achieving high-resolution (HR) dose distributions is crucial but time-intensive, posing a significant bottleneck in the clinical workflow of treatment planning. A common workaround is to first calculate a dose distribution at low resolution and then refine it using super-resolution (SR) techniques. Building upon this approach, we introduce a novel super-resolution (SR) technique for fast and high-quality dose calculation. The proposed SR technique leverages a cutting-edge implicit neural representation (INR) network to learn a continuous function that accurately represents the target HR dose distribution image. In such a way, our model can produce dose maps at virtually any resolution. Unlike previous methods which assume a pre-known blurring kernel, we tackle a blind deconvolution problem where the blurring kernel is unknown and thus is estimated together with the HR image. We tested our method on a pancreas SBRT plan using the MRIdian MRgRT system, demonstrating significant improvements in image quality over an existing INR-based SR method that relies on predefined kernels. Our technique offers rapid SR dose calculation, achieving results in milliseconds.

1 Introduction

In the field of radiation therapy (RT), achieving both accuracy and efficiency in calculating dose distribution is paramount. The Monte Carlo or transport-based method stands as the most accurate dose calculation method [1]. However, the intensive computational demands associated with these methods pose significant challenges to the efficiency of treatment planning. As such, speeding up the dose calculation process without compromising on accuracy is highly demanded. A wealth of research efforts has been dedicated to reducing the time required for dose calculation [2–5], a promising avenue among these efforts involves calculating dose distribution at a lower resolution to save time, followed by the application of super-resolution (SR) reconstruction techniques to enhance the resolution of the calculated dose [6, 7]. However, the task of image SR is notably complex due to its nature as an ill-posed inverse problem, where a single LR input may correspond to multiple HR outputs [8, 9]. Importantly, for medical images, the lack of ground-truth paired training datasets further complicates the issue.

The advent of deep learning (DL) techniques has brought significant advancements to the field of medical image processing, including SR. Traditional DL approaches, which rely on extensive datasets of paired LR and HR images, have shown promising results, particularly with natural images [10–15]. However, the rarity of large-scale medical image datasets, especially those containing clinical-quality HR and LR image pairs, poses a significant challenge to these supervised learning methods.

In this context, implicit neural representation (INR) networks have emerged as a powerful tool for encoding continuous functions that accurately represent signals such as sounds, images, videos, etc. By inputting pixel coordinates, INRs can output the image values for those pixels at any desired resolution [7, 16–22], offering a novel solution for SR by only requiring LR images for training. The network can then generate HR images by querying denser pixel coordinates [18], a method particularly advantageous for medical imaging where HR ground-truth data is scarce.

In this work, we focus on super-resolution for dose distribution images. Our work diverges from previous methodologies [22, 23] that assume LR images are the result of pre-known blurring and downsampling operations. Instead, we address the blind deconvolution problem, aiming to simultaneously estimate both the blurring kernel and the HR image. Unlike approaches [24–26] that utilize image generative models for discrete image and kernel generation, our methodology employs two INR networks to continuously represent both the images and the blurring kernel. This approach is particularly suited to medical imaging, where the blurring kernel often represents the continuous point spread function of the imaging equipment [27, 28]. Our innovative use of INR networks for learning both the blurring kernel and HR images not only caters to the unique challenges of medical image SR but also paves the way for advancements in the calculation of dose distribution in radiation therapy without the need for heavy computations and hardware upgrades.

2 Materials and Methods

2.1 Image observation model in SR imaging

A commonly used image observation model of the LR image from a HR image is

$$\mathbf{y} = \mathbf{DGx} + \mathbf{n}, \tag{1}$$

where $\mathbf{y} \in \mathbb{R}^{N_L}$ is the vectorized LR observation degraded from the HR image $\mathbf{x} \in \mathbb{R}^{N_H}$. The typical image degradation operations—downsampling, blurring, and noise corruption—are denoted by operators $\mathbf{D}$, $\mathbf{G}$, and

$\mathbf{n}$, respectively [29].

2.2 Implicit Neural Representation (INR)

The INR represents an image as a continuous function between its spatial coordinates and the corresponding intensity values:

$$\mathcal{M}(\theta) : \mathbb{R}^{dim} \to \mathbb{R}, \quad \mathbf{c}_i \mapsto \mathcal{M}(\theta; \mathbf{c}_i) = x_i, \qquad (2)$$

where $\mathcal{M}(\theta)$ denotes the INR network with parameters θ, $\mathbf{c}_i \in [0, 1)^{dim}$ is the normalized spatial coordinate of the ith pixel, and x_i is the corresponding intensity value. The notation dim is determined by the dimensionality of an image. In this paper, our investigation focuses on two-dimensional images, i.e., $dim = 2$.

The INR network is constructed based on an MLP network, with each fully-connected layer activated by a periodic function to enhance the representation of image details [30]. When representing images with complex textures containing high-frequency signals, an additional step named Fourier feature embedding (FFE) is often deployed before feeding the spatial coordinates into the INR network [17, 22, 31–33]:

$$\gamma(\mathbf{c}_i) = [\cos(2\pi \mathbf{B}\mathbf{c}_i), \sin(2\pi \mathbf{B}\mathbf{c}_i)]^T. \qquad (3)$$

With FFE, the low-dimensional coordinate of a single pixel is transformed into a set of additional points that cover a broader frequency range determined by the "frequency matrix" $\mathbf{B} \in \mathbb{R}^{N_{embed} \times dim}$. The matrix $\mathbf{B}$ is composed of random samples from a Gaussian distribution $\mathcal{N}(0, \sigma_b^2)$. The embedding size N_{embed} and the standard deviation σ_b are tunable hyperparameters in the network training.

2.3 INR-based blind deconvolution SR imaging

The proposed INR-based blind deconvolution utilizes two INR networks to represent the blurring kernel and the latent HR image, respectively, under the assumption that both the blurring kernel and the image are continuous functions in the spatial domain. Referring to the image observation model in Eq.(1), the blurring kernel $\mathbf{G}$ and the HR image $\mathbf{x}$ are entangled, posing challenges in training two INR networks to represent distinct objects separately.

Inspired by [9, 28], we introduce a physics-driven constraint for the blurring kernel. Specifically, in most optical imaging systems, the blurring kernel models a continuous space point spread function (PSF) of the imaging system, and the PSF corresponding to higher-resolution imaging $\mathbf{b}_H$ is a scaled version of the PSF $\mathbf{b}_H$ at lower-resolution imaging, i.e.,

$$\mathbf{b}_H(\mathbf{c}) = \alpha \mathbf{b}_L(\alpha \mathbf{c}), \qquad (4)$$

where α is the scaling factor from the lower-to-higher resolution image. We also introduce the non-negativity and sum-to-one constraints to the blurring kernel based on the non-negative and energy-limited properties of PSF of most imaging systems [26].

Training the two INR networks only utilizes observations acquired at different resolution scales using the same imaging system. The loss function is formulated as follows:

$$\underset{\theta,\phi}{\operatorname{argmin}} \sum_{l=0}^{L} \sum_{k=1}^{K} \|\mathbf{y}_{l,k} - \alpha_l M_b(\phi; \alpha_l \mathbf{c}_{B_L}) \circledast M_I(\theta; \mathbf{c}_{I_H})\|_2^2. \qquad (P0)$$

In this function, $\mathbf{y}_{l,k} \in \mathbb{R}^{N_l}$ is the kth LR image at the lth scale, α_l is the SR scaling factor from the lth level LR image in the training set to the target HR resolution, $\mathbf{c}_{B_L}$ denotes the coordinate grid of the blurring kernel corresponding to the lowest resolution in the training set, and $\mathbf{c}_{I_H}$ denotes the coordinates of the target HR image. The two INR networks, i.e., the "Blur-INR" for the blurring kernel and the "Image-INR" for the HR image are denoted by $M_b(\phi)$ and $M_I(\theta)$, respectively. Figure 1 shows the architecture of the proposed method.

3 Experiments & Results

We selected one pancreas SBRT plan from the MRIdian MRgRT system. The dose was computed with 1mm, 2mm, and 4mm resolutions. The 1mm dose map served as the HR reference for SR imaging evaluation, while the 2mm and 4mm dose maps were used as LR images for network training. We compared the proposed method with a prior work, MLINR, which does not involve blind deconvolution but uses an empirically pre-determined blurring kernel [22]. In the proposed method, we set the the kernel size to be 27×27. The learning was set to 5e-5 for both Blur-INR and the Image-INR networks. For the compared MLINR method, a Gaussian blur kernel of size 27×27 with a standard deviation of 9 was empirically pre-determined for the image of the lowest resolution in the training set, i.e., the dose map with 4mm resolution. The same scaling operation described in eq. (4) was applied to the higher resolution image (2mm resolution). The learning rate in this case was set to 1e-5 for an optimal result.

Figure 2 displays the SR dose maps along with the dose maps generated from the MRIdian MRgRT system. The proposed method significantly improves the image quality compared with the input LR acquisitions. Compared with the MLINR method, the proposed method achieved 4.7 dB and 0.073 improvements in terms of peak signal-to-noise ratio (PSNR) and structural similarity measure (SSIM). It also provide sharper textures with less artifacts compared with the compared MLINR method.

In terms of computation time, obtaining the dose maps of resolutions 1mm, 2mm, and 4mm from the MRId-

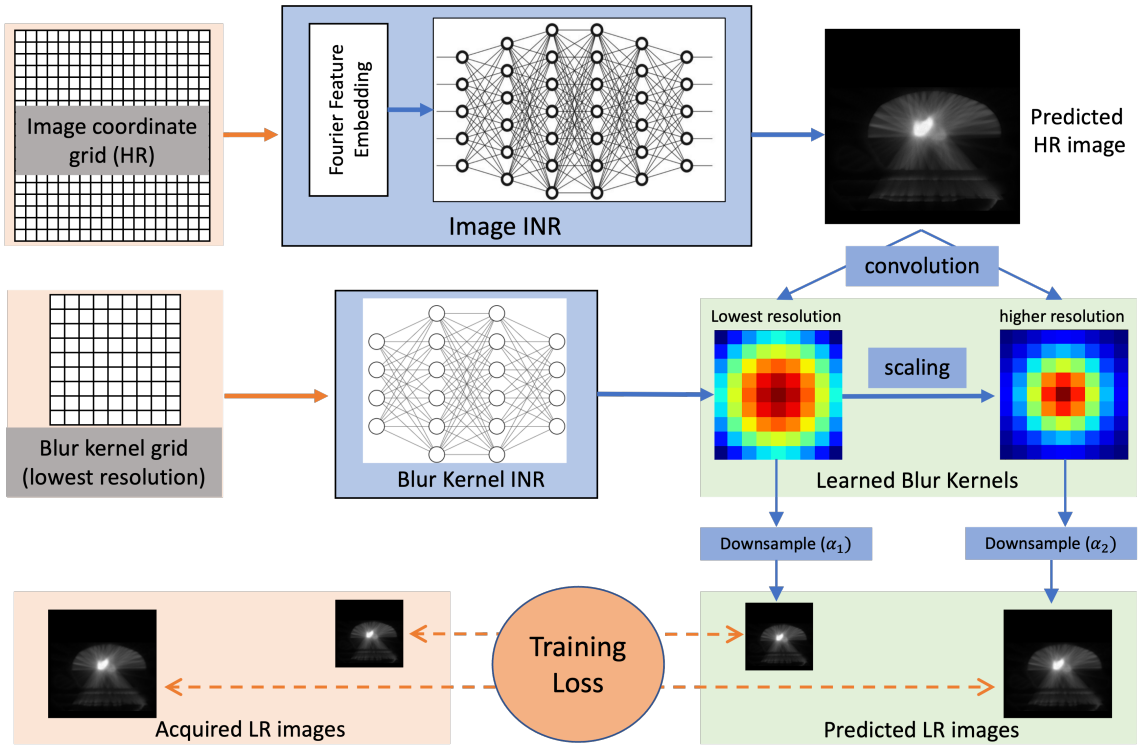


Figure 1: Architecture of the proposed method.

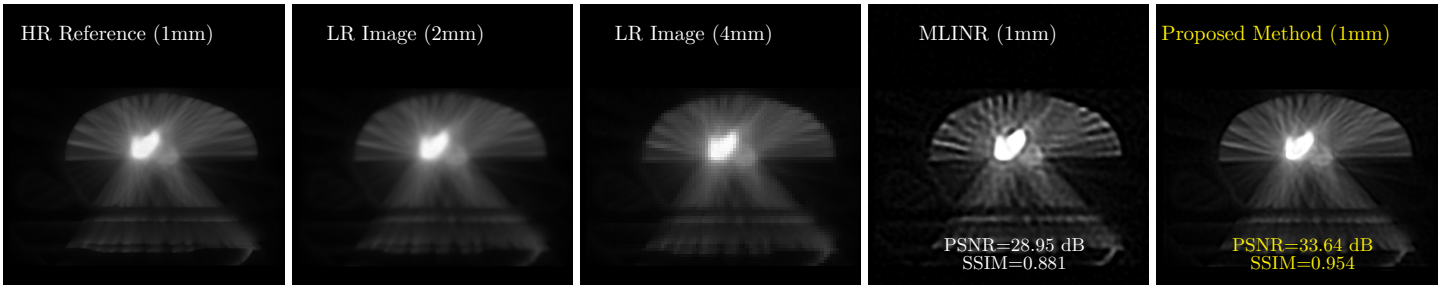


Figure 2: TPS generated dose distributions from a pancreas SBRT plan and the SR dose distribution results of the proposed method and the compared MLINR method [22]. The proposed method outperformed the MLINR in terms of higher PSNR and SSIM values as well as visual qualities.

ian MRgRT system took 2 minutes and 50 seconds, 25 seconds, and 5.5 seconds, respectively. With the proposed method, it took around 3 minutes to train the network and only milliseconds to obtain the HR image at any resolution using a single Nvidia GeForce RTX 3090 graphic card.

4 Conclusion

This paper introduces an innovative blind deconvolution technique based on INR networks for rapid and high-quality SR dose calculation. Leveraging the underlying image observation model and the properties of the PSF of imaging systems, we trained two INR networks to represent the blurring kernel and the latent HR image, respectively. Experiments conducted on dose maps from the MRIdian MRgRT system showcase the effectiveness and efficiency of the proposed method. It's noteworthy that this method doesn't necessitate LR-HR image pairs for model training and is tailored to individual patients. This significantly addresses concerns related to data scarcity and optimizes treatment for patient-specific scenarios in the RT community. Furthermore, while we specifically demonstrate the application of SR dose calculation, the proposed method exhibits versatility and can be readily applied to various biomedical image modalities.

References

[1] D. Rogers. "Fifty years of Monte Carlo simulations for medical physics". *Physics in Medicine & Biology* 51.13 (2006), R287.

[2] A. Boyer and E. Mok. "A photon dose distribution model employing convolution calculations". *Medical physics* 12.2 (1985), pp. 169–177.

[3] R. Mohan, C. Chui, and L. Lidofsky. "Differential pencil beam dose computation model for photons". *Medical Physics* 13.1 (1986), pp. 64–73.

[4] W. Ulmer, J. Pyyry, and W. Kaissl. "A 3D photon superposition/convolution algorithm and its foundation on

results of Monte Carlo calculations". *Physics in Medicine & Biology* 50.8 (2005), pp. 1767–1790.

[5] J. Sievinen, W. Ulmer, W. Kaissl, et al. "AAA photon dose calculation model in Eclipse". *Palo Alto (CA): Varian Medical Systems* 118 (2005), p. 2894.

[6] P. Dong and L. Xing. "Deep DoseNet: a deep neural network for accurate dosimetric transformation between different spatial resolutions and/or different dose calculation algorithms for precision radiation therapy". *Physics in Medicine & Biology* 65.3 (2020), p. 035010.

[7] V. Vasudevan, L. Shen, C. Huang, et al. "Implicit neural representation for radiation therapy dose distribution". *Physics in Medicine & Biology* 67.12 (2022), p. 125014.

[8] G. Rohith and L. S. Kumar. "Paradigm shifts in super-resolution techniques for remote sensing applications". *The Visual Computer* 37.7 (2021), pp. 1965–2008.

[9] B. Bernstein, S. Liu, C. Papadaniil, et al. "Sparse recovery beyond compressed sensing: Separable nonlinear inverse problems". *IEEE transactions on information theory* 66.9 (2020), pp. 5904–5926.

[10] C. Dong, C. C. Loy, K. He, et al. "Image super-resolution using deep convolutional networks". *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38.2 (2015), pp. 295–307.

[11] C. Dong, C. C. Loy, and X. Tang. "Accelerating the super-resolution convolutional neural network". *European conference on computer vision*. Springer. 2016, pp. 391–407.

[12] W. Shi, J. Caballero, F. Huszár, et al. "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network". *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 1874–1883.

[13] Z. Wang, J. Chen, and S. C. Hoi. "Deep learning for image super-resolution: A survey". *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43.10 (2020), pp. 3365–3387.

[14] K. Zhang, L. V. Gool, and R. Timofte. "Deep unfolding network for image super-resolution". *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2020, pp. 3217–3226.

[15] L. Xing, M. L. Giger, and J. K. Min. *Artificial intelligence in medicine: technical basis and clinical applications*. Academic Press, London, UK, 2020.

[16] V. Sitzmann, M. Zollhöfer, and G. Wetzstein. "Scene representation networks: Continuous 3d-structure-aware neural scene representations". *Advances in Neural Information Processing Systems* 32 (2019).

[17] L. Shen, J. Pauly, and L. Xing. "NeRP: implicit neural representation learning with prior embedding for sparsely sampled image reconstruction". *IEEE Transactions on Neural Networks and Learning Systems* (2022).

[18] Q. Wu, Y. Li, L. Xu, et al. "IREM: High-resolution magnetic resonance image reconstruction via implicit neural representation". *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2021, pp. 65–74.

[19] Y. Chen, S. Liu, and X. Wang. "Learning continuous image representation with local implicit image function". *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2021, pp. 8628–8638.

[20] D. Van Veen, R. Van der Sluijs, B. Ozturkler, et al. "Scale-agnostic super-resolution in MRI using feature-based coordinate networks". *arXiv preprint arXiv:2210.08676* (2022).

[21] Q. Wu, Y. Li, Y. Sun, et al. "An arbitrary scale super-resolution approach for 3D MR images via implicit neural representation". *IEEE Journal of Biomedical and Health Informatics* 27.2 (2023), pp. 1004–1015. DOI: 10.1109/JBHI.2022.3223106.

[22] S. Ye, L. Shen, M. T. Islam, et al. "Super-resolution biomedical imaging via reference-free statistical implicit neural representation". *Physics in Medicine & Biology* 68.20 (2023), p. 205020.

[23] Y. Sui, O. Afacan, C. Jaimes, et al. "Scan-specific generative neural network for MRI super-resolution reconstruction". *IEEE Transactions on Medical Imaging* 41.6 (2022), pp. 1383–1399.

[24] D. Ulyanov, A. Vedaldi, and V. Lempitsky. "Deep image prior". *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 9446–9454.

[25] Y. Gandelsman, A. Shocher, and M. Irani. ""Double-DIP": unsupervised image decomposition via coupled deep-image-priors". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 11026–11035.

[26] D. Ren, K. Zhang, Q. Wang, et al. "Neural blind deconvolution using deep priors". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 3341–3350.

[27] T. J. Holmes, S. Bhattacharyya, J. A. Cooper, et al. "Light microscopic images reconstructed by maximum likelihood deconvolution". *Handbook of biological confocal microscopy* (1995), pp. 389–402.

[28] T. Michaeli and M. Irani. "Nonparametric blind super-resolution". *Proceedings of the IEEE International Conference on Computer Vision*. 2013, pp. 945–952.

[29] P. Milanfar. *Super-resolution imaging*. CRC press, 2017.

[30] V. Sitzmann, J. Martel, A. Bergman, et al. "Implicit neural representations with periodic activation functions". *Advances in Neural Information Processing Systems* 33 (2020), pp. 7462–7473.

[31] R. Basri, M. Galun, A. Geifman, et al. "Frequency bias in neural networks for input of non-uniform density". *International Conference on Machine Learning*. PMLR. 2020, pp. 685–694.

[32] N. Rahaman, A. Baratin, D. Arpit, et al. "On the spectral bias of neural networks". *International Conference on Machine Learning*. PMLR. 2019, pp. 5301–5310.

[33] M. Tancik, P. Srinivasan, B. Mildenhall, et al. "Fourier features let networks learn high frequency functions in low dimensional domains". *Advances in Neural Information Processing Systems* 33 (2020), pp. 7537–7547.

Evaluation of organ dose reduction effectiveness in liver 3D/4D computed tomography simulation through radiation shielding: an anthropomorphic phantom study

Jongin Oh[1], Minsik Lee[2], BongHyo Lee[3], DooHyun Lee[1,4], Min-Young Kang[5], Semie Hong[1,5], Jun-Bong Shin[2*], Jeong-Woo Lee[1,5]

[1]Department of Convergent Medical Physics Graduate School on Engineering, Konkuk University, Korea
[2]Department of Radiation Oncology, Kangwon National University Hospital, Chuncheon-si, Korea
[3]Department of Physics, Yeungnam University, Gyeongsan, Korea
[4]Department of Radiation Oncology, Good Morning Hospital, Pyeongtaek-si, Korea
[5]Department of Radiation Oncology, Konkuk University Medical Center, Konkuk University of Medicine, Seoul, Korea

Abstract The purpose of this study was to evaluate the effectiveness of radiation shielding in reducing organ doses was evaluated during liver 3D/4D Simulation Computed tomography (Sim-CT) prior to radiation therapy. A tunnel-shaped acrylic structure was constructed to support the shield, and two lead equivalents of 1 and 3 mm/Pb were used to confirm the dose reduction effect of radiation shielding. Optical stimulated luminescence dosimeters (OSLD) were inserted at 11 locations, including the thyroid, within an anthropomorphic phantom. Under clinical conditions, liver 3D/4D Sim-CT were repeated three times, and radiation shielding was applied to analyze the dose-reducing effects. The use of self-produced tunnel-type shielding tool is more effective in reducing unnecessary effective doses of normal tissue outside the examination site during liver 3D/4D Sim-CT.

1 Introduction

Computed tomography (CT) is a procedure that can easily and quickly produce images of the human body using ionizing radiation. Therefore, it is used in medical diagnosis as well as in nuclear medicine and radiation therapy. CT examinations have become essential worldwide, and the number of national CT examinations belonging to the Organization for Economic Co-operation and Development is increasing annually [1]. However, when performing CT for medical purposes, the As Low as Reasonably Achievable (ALARA) principle does not specify dose limits, but all radiation exposures pose a potential risk for healthcare workers and patients. Therefore, the purpose of radiation defense is to reduce unnecessary radiation and minimize the harmful effects of ionizing radiation [2]. In that sense, efforts are needed to reduce the radiation exposure to patients in all departments using CT. Since radiation dose has a directly impact on image quality [3-5], any attempt to reduce the delivered dose should ensure that the resulting image quality remains suitable for accurate and reliable diagnosis [6, 7]. In recent years, various studies and manufacturers have introduced various methods and technologies to improve image quality while minimizing radiation dose in clinical practice [8-12]. Despite these developments, When the radiation beam is fired towards the target site, the scattering rays scatter to normal organs other than the target site, which may lead to unnecessary exposure of the tissue other than the target site of the patient being tested. This may expose the patient to unnecessary radiation, and the resulting additional radiation exposure may increase the risk of developing cancer in the long run. The patient undergoes multiple CT scans during diagnosis and treatment, which may result in the accumulation of unnecessary radiation in the normal organs. However, CT simulation (Sim-CT) should be performed essentially to determine the treatment site and accurately calculate the treatment dose before radiotherapy [13].

In Sim-CT, the patient's treatment posture is determined, the radiotherapy site is reconfirmed, and the dose is calculated using treatment planning system (TPS) by sketching the outline of normal organs and tumors. In addition, accurate determination of the target and target boundaries in Sim-CT is a prerequisite for successful radiotherapy. Recently, a more accurate radiotherapy plan can be established by selectively using a four-dimensional CT (4D Sim-CT) technique in addition to three-dimensional CT (3D Sim-CT), but previous studies have reported that 4D Sim-CT scan can provide patients with a dose that is 10 times higher than 3D Sim-CT scan [14-16]. This is higher than the CT dose used in the diagnostic field, which can result in increased doses in normal organs outside the therapeutic site. The main difference between 3D Sim-CT and 4D Sim-CT is the technical distinction between CTV (Clinical Target Volume) to ITV (Internal Target Volume) and OAR (Organ at Risk) to PRV (Planning Organs at Risk Volume), which includes the concept of internal motion by tracking the volume change over time of a tumor or a highly mobile part of a normal organ. While 3D Sim-CT provides a detailed view of the patient's anatomy but does not account for organ or tumor motion during the respiratory or cardiac cycle, 4D Sim-CT allows for the assessment of organ and tumor motion due to respiration or cardiac activity. This is particularly important for tumors located in or near organs affected by motion, such as the lungs or liver. In this study, we tried to focus on the dose decrease due to the application of these techniques and the dose reduction effect in terms of radiation protection due to the use of radiation shields, not on the movement of tumors or normal organs themselves.

Therefore, we wanted to evaluate the long-term dose change before and after the implementation of radiation shielding using ATOM® dosimetry phantom (Family Model 701, Sun Nuclear, USA) during liver 3D/4D Sim-CT to evaluate exposure of tissues other than the patient's target site to unnecessary radiation.

2 Materials and Methods

The ATOM® dosimetry phantom imitates the physical characteristics of an adult man and allows for more accurate human absorption dose calculation by considering variations in tissue composition based on age or sex, depending on the model. This facilitates evaluation of the dose weighting affected by radiation exposure specific to different tissues in the human body. In addition, the Phantom family model 701 was precisely divided into 25-mm thick sections, and 271 organs were replicated in these sections. Each dosemeter was designed to be inserted into the corresponding position of the organ, simplifying the measurement of organ-specific doses.

Sim-CT equipment (Aqualion TM LB, Canon Medical Systems, Japan) was utilized in this study, with parameters commonly used for 3D Sim-CT and 4D Sim-CT image acquisition. The parameters used for liver 3D Sim-CT simulation treatment were as follows: tube voltage, 120 kV; tube current, 120 mA; rotation time, 0.5 s; range, 501 mm; fragment thickness, 1.0 mm; total inspection time, 17.8 s; and D-FOV (Digital Field of View), 350. Meanwhile, the parameters for 4D Sim-CT simulation treatment were as follows: tube current, 200 mA; rotation time, 0.5 s; range, 250 mm; slice thickness, 2.0 mm; and total test time D-FOV, 20.2 mm.
The NanoDot® dosemeter (Landauer, France) The NanoDot® dosemeter (Landauer, France) for dose measurements, and calibration was carried out this way [17-18].

The shield was designed as a tunnel-type acrylic structure, divided into upper and lower parts, to provide support and shield the areas outside the CT scan area. The dimensions of the upper and lower shields were 600 × 470 × 250 mm3 and 850 × 470 × 250 mm3, respectively. In addition, to determine the shielding effect based on the lead equivalent, two lead equivalents of 1 mm/Pb and 3 mm/Pb were placed on the acrylic structure to construct an experiment setup (Fig. 1.)



Figure 1: Shielding device (the upper shield was 600 × 470 × 250 mm high, and the lower shield was 850 × 470 × 250 mm high). To determine the shielding effect according to the lead equivalent, lead equivalents of 1 mm/Pb and 3 mm/Pb, respectively, were prepared and placed on an acrylic structure to construct an experiment. (Left) Upper shielding device. (Right) Lower shielding device.

Respiratory gated radiation therapy is used in the treatment of moving tumors, such as lung and liver cancers. These techniques employ a real-time position management respiratory tuning gating system to deliver radiation beams while accounting for organ movement.
Infrared cameras track tumors moving due to the patient's breathing during treatment, linear accelerators increase the accuracy of treatment locations by irradiating tumors with X-rays in a set breathing section. In standard practice, a non-invasive infrared tracking marker box is attached to the patient's body. However, for experimental purpose, an infrared tracking marker box was placed on top of the phantom. The illuminator's infrared rays were reflected from the marker and captured by the camera. This setup enabled precise treatment delivery, while minimizing treatment errors by delivering maximum dose to the tumor and avoiding damage to critical organs. This was achieved by continuously monitoring the chest wall movement during treatment and treating it only within a preset respiration.

In general, the scan range of the Liver 3D Sim-CT was set to the clinical range (from nipple to pubic symphysis), and to measure the dose, an optically stimulated luminescent dosemeter (OSLD) was inserted in 11 locations, including the bladder, testicles, and lens, with 1 cm of bolus attached to the ATOM® dosemetry phantom. The 3D Sim-CT scan was repeated three times for accurate measurements. When applying radiation shielding, it was placed in the upper and lower parts of the hip joint including the heart so that radiation shielding did not directly affect the liver. The dose reduction effect was analyzed for each measurement.
The OSLD analysis involved performing 10 repeated measurements to reduce its inherent errors. The average value of eight measurements, excluding the maximum and minimum values, was rounded to the second decimal place. The reading program used for analysis was InLight microStar Reader Version 4.3.4506.25886, Data Base Version 9.0.0.0, InLight Dose Calc Version 1.13, and Reader Communication Version 1.1.0.3.

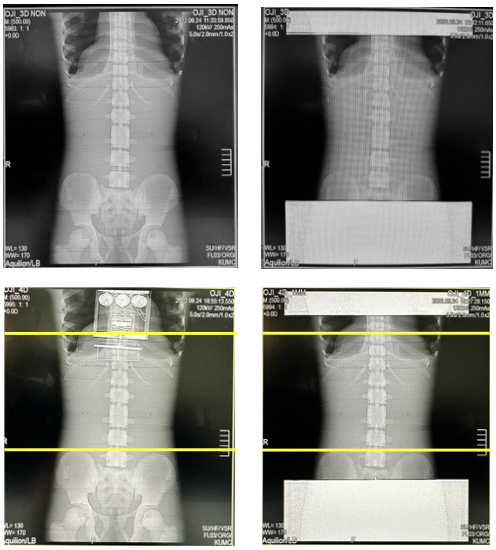


Figure 2: Scout image for the liver during 3D-CT and 4D-CT simulation. (Up,Left) Unshielded scan range for the liver during 3D-CT simulation. (Up, Right) Shielded scan range for the liver during 3D-CT simulation. (Down,Left) Unshielded scan range for the lung during 4D-CT simulation. (Down,Right) Shielded scan range for the lung during 4D-CT simulation.

3 Results

The main finding of the study revealed that the organ dose was significantly reduced when using radiation shielding in both 3D Sim-CT scans and 4D Sim-CT scans. In 3D Sim-CT scans, the organ dose decreased from 17.1 ± 0.2 mGy to 4.8 ± 0.1 mGy when using a 1 mm/Pb shield and to 4.1 ± 0.02 mGy when using a 3 mm/Pb shield. Additionally, in the 4D Sim-CT scan, the organ dose decreased from 108.3 ± 0.3 mGy to 97.9 ± 0.7 mGy when using the 1 mm/Pb shield and to 96.1 ± 0.2 mGy when using the 3 mm/Pb shield.
The organ dose reduction ratio to shield thickness is presented in Tables 4 and 5. In the 3D Sim-CT scans with 1 mm/Pb shielding, the organ average dose of OSLD increased in the right lens by 42.0%, left lens by 42.4%, right thyroid by 18.3%, left thyroid by 33.4%, esophagus by 32.3%, right lung by 48.5%, left lung by 55.3%, right bladder by 81.5%, left bladder by 82.2%, right testicle by 92.4%, and left testicle by 93.1%. In the cumulative analysis, the doses in the lens, thyroid gland, lungs, bladder, and testicles decreased by 43.85%, 66.30%, 52.07%, 81.84%, 92.74%, respectively. With the 3 mm/Pb shield, the doses decreased by 26.1% in the right lens, 21.9% in the left lens, 21.1% in the right thyroid, 21% in the left thyroid, 16.4% in the esophagus, 14.7% in the right lung, 7.8% in the left lung, 6.6% in the right bladder, 3.2% in the left bladder, 24.9% in the right testicle, and in the 24.1% left testicle. In the cumulative analysis, the doses in the lens, thyroid gland, lungs, bladder, and testicles decreased by 59.0%, 73.91%, 58.53%, 83.84%, and 94.39%, respectively. The total dose decreased by 72.0% and 4.1% with the 1 mm/Pb shield and 3 mm/Pb shields, respectively (Figs. 2–5).

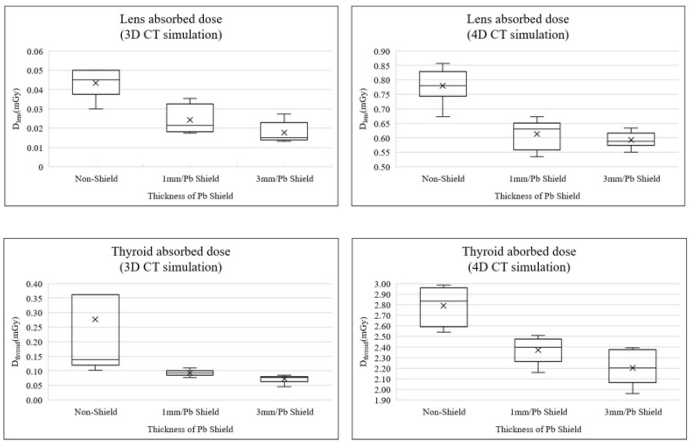


Figure 3. Comparison of the absorbed dose in the lens (up) and thyroid (down) with non-shielding and 1 mm/3 mm shielding in liver 3D Sim-CT and 4D Sim-CT simulation.

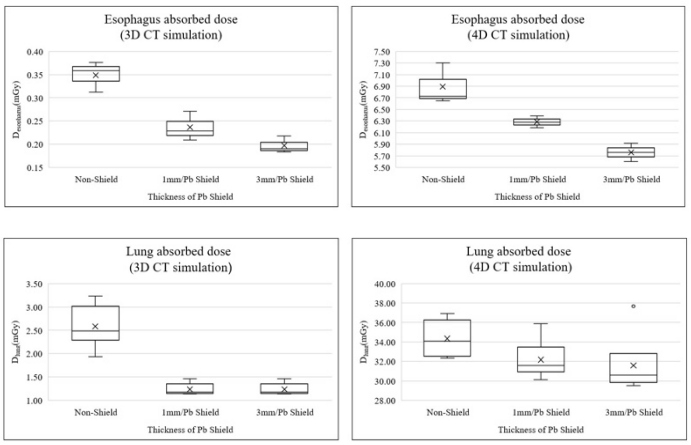


Figure 4. Comparison of the absorbed dose in the esophagus (up) and lung (down) with non-shielding and 1 mm/3 mm shielding in the liver 3D Sim-CT and 4D Sim-CT simulation.

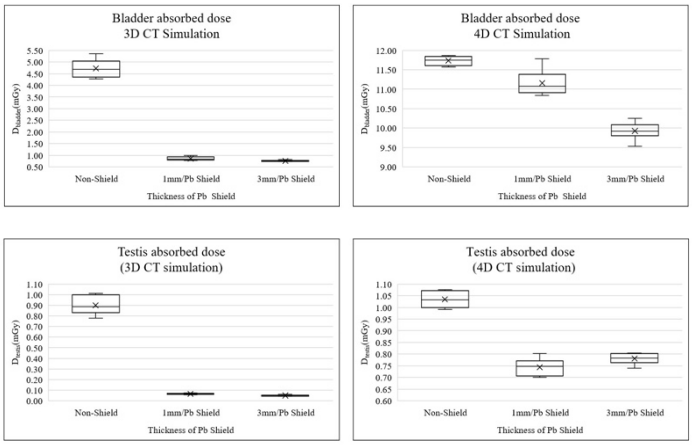


Figure 5. Comparison of the absorbed dose in the bladder (up) and testis (down) with non-shielding and 1 mm/3 mm shielding in the liver 3D Sim-CT and 4D Sim-CT simulation.

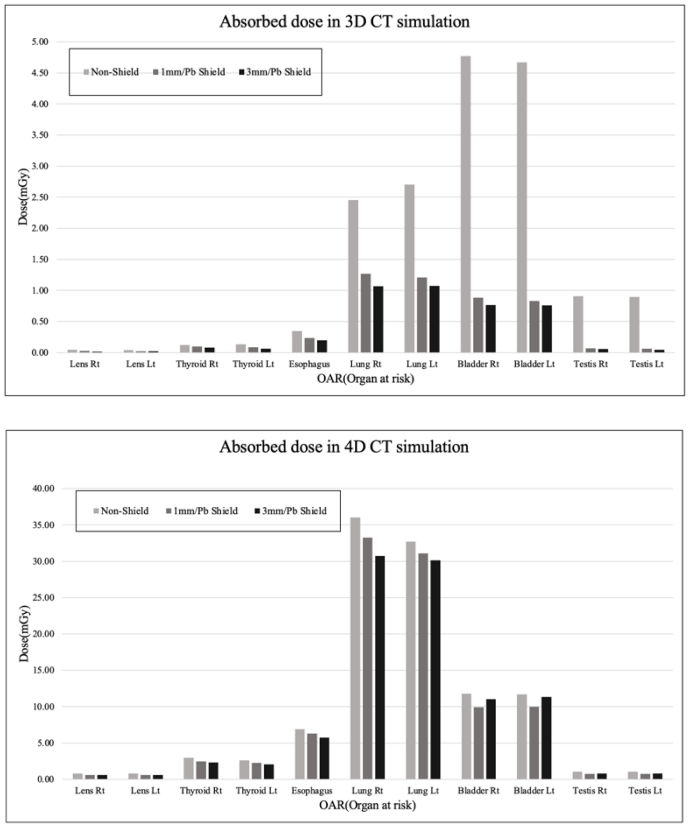


Figure 6. Comparison of total organ absorbed doses between non-shielding and 1 mm/3 mm shielding in liver 3D Sim-CT (up) and 4D Sim-CT simulation (down).

In 4D Sim-CT scans with 1 mm/Pb shielding, the organ average doses of OSLD increased in the right lens by 21.1%, left lens by 21.5%, right thyroid by 16.4%, left thyroid by 13.6%, esophagus by 8.8%, right lung by 7.6%, left lung by 4.9%, right bladder by 16%, left bladder by 14.9%, right testicle by 28.8%, and left testicle by 27.2%. In the cumulative analysis, the doses in the lens, thyroid gland, lungs, bladder, and testicles decreased by 21.3%, 15.07%, 6.32%, 4.9%, and 28.01%, respectively. In the 3 mm/Pb shield, the doses decreased by 26.1% in the right lens, 21.9% in the left lens, 21.1% in the right thyroid, 21.0% in the left thyroid, 16.4% in the esophagus, 14.7% in the right lung, 7.8% in the left lung, 6.6% in the right bladder, 3.2% in the left bladder, 24.9% in the right testicle, and 24.1% in the left testicle. In the cumulative analysis, the doses in the lens, thyroid gland, lungs, bladder, and testicles decreased by 24.0%, 21.06%, 8.04%, 15.41%, and 24.50%, respectively. The total dose was decreased by 9.5% and 11.3% with the 1 mm/Pb 3 mm/Pb shields, respectively (Table 1 and Figs. 2–5).

Region	Non-shield (mGy)	1-mm shield (mGy)	3-mm shield (mGy)	P-value
Lens	0.78 ± 0.05	0.61 ± 0.06	0.59 ± 0.02	<0.001
Thyroid	2.79 ± 0.06	2.37 ± 0.07	2.20 ± 0.10	<0.001
Esophagus	6.89 ± 0.36	6.29 ± 0.10	5.79 ± 0.16	<0.001
Lung	34.34 ± 0.64	32.18 ± 1.62	31.58 ±2.60	<0.001
Bladder	11.73 ± 0.12	9.93 ± 0.21	11.16 ±0.31	<0.001
Testis	1.03 ± 0.04	0.74 ± 0.04	0.78 ± 0.02	<0.001

** Kruskal–Wallis test.*

Table 1. Organ doses. P-values represent the statistical difference between the non-shield and shield groups in4D-CT simulation.

4 Discussion

Previous studies examining the organs at risk and associated cancer risk in patients undergoing 3D Sim-CT and 4D Sim-CT examinations have reported an average dose of 7.82–11.84 cGy and 0.64–0.85 cGy per 4D Sim-CT scan and 3D Sim-CT scan, respectively. The average dose per 4D Sim-CT scan was 12.8 times higher than that of 3D Sim-CT scans. Organ dose decreased linearly with the effective diameter function of the patient, and 4D Sim-CT scans were expected to increase the relative risk of cancer compared to that with the 3D Sim-CT scans [16]. The results of the previous studies have demonstrated that 4D Sim-CT scans during Sim-CT tests result in higher dose increases compared to that of the 3D Sim-CT scans. This is particularly concerning for children and young patients who are more sensitive to radiation. In this study, tunnel-type shields were designed to minimize the potential harmful effects of ionizing radiation by reducing unnecessary radiation exposure in 3D Sim-CT and 4D Sim-CT tests. The aim was to evaluate the benefits of these shields by comparing the difference in absorbed doses with or without their application using the ATOM® dosemetry phantom. The effectiveness of dose reduction using self-produced tunnel-type radiation shielding tools during 3D/4D Sim-CT simulation treatment was evaluated. Previous studies explored the application of shielding bodies on patients. These studies have employed various shapes of shielding bodies [19], including those that reduce the dose delivery to the gonads during hip imaging [20]. Additionally, researchers have investigated shielding methods involving the use of bismuth to reduce the dose delivered to the thyroid, eyeball, and breast [21].

In this study, a tunnel-type shield was designed to provide simultaneous front and rear shielding for the patient during the 3D/4D Sim-CT test prior to radiation therapy, aimed at confirming the unnecessary dose reduction effect for major organs outside the test site, such as the rear scattering line. The shield's shape was designed to be comfortable and does not interfere with the test when applied to patients in clinical practice, and the content of lead equivalent was changed to 1 mm/Pb and 3 mm/Pb. The shielding effect according to lead equivalent was also studied.

However, because the posture of CT patients is the same as that of radiation therapy patients, there is a limitation that radiation shields cannot be used together if they are large, such as immobilization used for breast therapy. Therefore, in the next study, we plan to find a solution for the clinical application of radiation shield.

The results of this study also showed that the mean dose per 4D Sim-CT scan was higher than that of the 3D Sim-CT scan [16], although it varied from site to site when no shield was used, as in previous studies. In addition, it was confirmed that the long-term dose decreased linearly with the effective diameter of the patient, and the 4D Sim-CT scan is expected to increase the relative risk of cancer compared to the 3D Sim-CT scan, which is thought to require active shield use.
Furthermore, when tunnel-type shields were used, the shielded organ dose significantly reduced with the use of radiation shielding in both 3D Sim-CT scans and 4D Sim-CT scans. This finding suggests that increasing the lead thickness from 1 mm to 3 mm in the 3D/4D Sim-CT test can effectively reduce radiation dose. However, using 3-mm shields in 4D Sim-CT simulations led to higher organ doses in the bladder and testis compared to that of 1-mm shield. This may be attributed to scattered lines generated within the thicker shield (Fig. 4). To reduce the scattering line in this area, applying additional shields or using shields of appropriate thickness was considered necessary. Further, the image quality of the 3D/4D Sim-CT inspection area before and after the shield application as well as the organ dose of the scanned area needs to be evaluated. Thus, further research is needed in the future.

5 Conclusion

This study confirmed the use of self-produced tunnel-type shielding tools in effectively reducing unnecessary radiation exposure during CT examination, which is crucial for medical diagnosis and treatment. The effectiveness of a shielding device is particularly noticeable in 3D Sim-CT examinations, where it significantly reduces the radiation dose outside the patient's examination area. Moreover, more active shielding may be required when applying it to CT for general diagnosis purpose.

References

[1] OECD, *Health at a Glance 2021*. (2021).

[2] E. L. Mitchell and P. Furey, "Prevention of radiation injury from medical imaging," (in eng), J Vasc Surg, vol. 53, no. 1 Suppl, pp. 22s-27s, Jan 2011, DOI: 10.1016/j.jvs.2010.05.139.

[3] K. J. Strauss et al., "Image gently: Ten steps you can take to optimize image quality and lower CT dose for pediatric patients," (in eng), AJR Am J Roentgenol, vol. 194, no. 4, pp. 868-73, Apr 2010, DOI: 10.2214/ajr.09.4091.

[4] A. S. Brody, D. P. Frush, W. Huda, and R. L. Brent, "Radiation risk to children from computed tomography," (in eng), Pediatrics, vol. 120, no. 3, pp. 677-82, Sep 2007, DOI: 10.1542/peds.2007-1910.

[5] L. F. Donnelly, "Reducing radiation dose associated with pediatric CT by decreasing unnecessary examinations," (in eng), AJR Am J Roentgenol, vol. 184, no. 2, pp. 655-7, Feb 2005, DOI: 10.2214/ajr.184.2.01840655.

[6] W. A. Kalender et al., "Technical approaches to the optimisation of CT," (in eng), Phys Med, vol. 24, no. 2, pp. 71-9, Jun 2008, DOI: 10.1016/j.ejmp.2008.01.012.

[7] M. K. Kalra et al., "Strategies for CT radiation dose optimization," (in eng), Radiology, vol. 230, no. 3, pp. 619-28, Mar 2004, DOI: 10.1148/radiol.2303021726.

[8] M. Alshamari et al., "Impact of iterative reconstruction on image quality of low-dose CT of the lumbar spine," (in eng), Acta Radiol, vol. 58, no. 6, pp. 702-709, Jun 2017, DOI: 10.1177/0284185116669870.

[9] M. Kim et al., "Adaptive iterative dose reduction algorithm in CT: effect on image quality compared with filtered back projection in body phantoms of different sizes," (in eng), Korean J Radiol, vol. 15, no. 2, pp. 195-204, Mar-Apr 2014, DOI: 10.3348/kjr.2014.15.2.195.

[10] F. Becce et al., "Computed tomography of the cervical spine: comparison of image quality between a standard-dose and a low-dose protocol using filtered back-projection and iterative reconstruction," (in eng), Skeletal Radiol, vol. 42, no. 7, pp. 937-45, Jul 2013, DOI: 10.1007/s00256-013-1576-9.

[11] M. S. May et al., "Dose reduction in computed tomography of the chest: image quality of iterative reconstructions at a 50% radiation dose compared to filtered back projection at a 100% radiation dose," (in eng), Rofo, vol. 186, no. 6, pp. 576-84, Jun 2014, DOI: 10.1055/s-0033-1356254.

[12] M. S. May et al., "Dose reduction in abdominal computed tomography: intraindividual comparison of image quality of full-dose standard and half-dose iterative reconstructions with dual-source computed tomography," (in eng), Invest Radiol, vol. 46, no. 7, pp. 465-70, Jul 2011, DOI: 10.1097/RLI.0b013e31821690a1.

[13] A. Brahme, "Treatment Optimization Using Physical and Radiobiological Objective Functions," in Radiation Therapy Physics, A. R. Smith Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 1995, pp. 209-246.

[14] G. Wu, Q. Wang, J. Lian, and D. Shen, "Reconstruction of 4D-CT from a single free-breathing 3D-CT by spatial-temporal image registration," (in eng), Inf Process Med Imaging, vol. 22, pp. 686-98, 2011, DOI: 10.1007/978-3-642-22092-0_56.

[15] D. P. Shang, C. X. Liu, and Y. Yin, "A comparison of the different 3D CT scanning modes on the GTV delineation for the solitary pulmonary lesion," (in eng), Radiat Oncol, vol. 9, p. 211, Nov 12 2014, DOI: 10.1186/1748-717x-9-211.

[16] C. Yang, R. Liu, X. Ming, N. Liu, Y. Guan, and Y. Feng, "Thoracic Organ Doses and Cancer Risk from Low Pitch Helical 4-Dimensional Computed Tomography Scans," (in eng), Biomed Res Int, vol. 2018, p. 8927290, 2018, DOI: 10.1155/2018/8927290.

[17] P. A. Jursinic, "Characterization of optically stimulated luminescent dosimeters, OSLDs, for clinical dosimetric measurements," (in eng), Med Phys, vol. 34, no. 12, pp. 4594-604, Dec 2007, DOI: 10.1118/1.2804555.

[18] R. M. Al-Senan and M. R. Hatab, "Characteristics of an OSLD in the diagnostic energy range," (in eng), Med Phys, vol. 38, no. 7, pp. 4396-405, Jul 2011, DOI: 10.1118/1.3602456.

[19] L. T. Dauer, K. A. Casciotta, Y. E. Erdi, and L. N. Rothenberg, "Radiation dose reduction at a price: the effectiveness of a male gonadal shield during helical CT scans," (in eng), BMC Med Imaging, vol. 7, p. 5, Mar 16 2007, DOI: 10.1186/1471-2342-7-5.

[20] Y. S. Tsai et al., "Shielding during x-ray examination of pediatric female patients with developmental dysplasia of the hip," (in eng), J Radiol Prot, vol. 34, no. 4, pp. 801-9, Dec 2014, DOI: 10.1088/0952-4746/34/4/801.

[21] 21. Seoung Y. H. Evaluation of surface radiation dose reduction and radiograph artifact images in computed tomography on the radiation convergence shield by using sea-shells. J Korea Converg Soc. 8, 113–120 (2017).

Artificial Intelligence Dose Calculation Engine Based Pencil-beam for Intensity Modulated Radiation Therapy

Zirong Li[+1], Yaoying Liu[+2,3], Xuying Shang[2,3], Huashan Sheng[1], Wei Zhao[2], Gaolong Zhang[2], Qichao Zhou[*1], and Shouping Xu[*3,2]

[1]Department of Research Algorithms, Manteia Technologies Co., Ltd, Xiamen, China
[2]School of Physics, Beihang University, Beijing 102206, People's Republic of China
[3]National Cancer Center/ National Clinical Research Center for Cancer/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, 100021, China

Abstract The study proposed a data-generating method for training a deep learning (DL) model to the boost pencil beam (PB) algorithm accuracy of the dose, and comparable to Monte Carlo (MC). The study only involved 43 clinical cases but generated 11278 pairs of beam dose data for training, validating, and testing. The DL model showed significant ability in converting PB dose to MC dose, which improved the GPRs (1mm/1%) from 83.73±6.55%, 84.28±6.55% and 82.72±6.75%, to 96.07±2.22%, 96.17±1.90% and 95.31±2.80%, for training, validating and testing data. Moreover, the DL model has an ultra-speed with 180ms calculation time cost. The study successfully realized fast-accurate photon dose calculation by DL model.

1 Introduction

Cancer remains the leading cause of death globally[1]. Treatment modalities comprise radiation therapy, surgery, chemotherapy, immunotherapy, and hormonal therapy. Radiation therapy remains an essential component of cancer treatment with approximately 50% of all cancer patients receiving radiation therapy during their course of illness[2]. The most commonly used technology is intensity-modulated radiotherapy (IMRT)[3, 4].

Treatment planning is based on inverse planning algorithms to generate treatment beams with varying intensities across the beam section. Inverse optimization searches for the most suitable target through predefined dose limits and many iterations. In this process, an accurate dose calculation engine is required to make the final optimization result reliable. The Monte Carlo(MC) algorithm can accurately perform dose calculations, but the excessive calculation makes it challenging to reduce the time consumption. Although much work has been proposed based on GPU Monte Carlo[5, 6], and the time consumption of a single calculation has been reduced from hours to minutes, but still cannot meet the speed requirements of reverse optimization. In addition to the MC algorithm, dose calculations based on analytical algorithms are usually high-speed, but these algorithms sacrifice high accuracy, for example, the pencil beam(PB) algorithm. Many studies have optimized PB to improve the accuracy of dose calculation, but there are still differences compared with MC, especially in areas with uneven density changes[7, 8].

Neural networks have developed rapidly in recent years, and many studies have used deep learning(DL) to accelerate dose calculations[9–11]. However, the effect of neural networks is generally based on the amount of data during training. A dataset that is too small will lead to poor generalization ability of the model, making it perform well on the same dataset but not be usable in practice. However, clinical data are very scarce, and collecting massive datasets for training like natural images is challenging.

Therefore, we proposed a training method based on virtual planning, which can quickly generate a dose distribution equivalent to MC dose calculation by inputting the less accurate PB dose calculation results.

2 Materials and Methods

We proposed a training method based on the virtual plan to quickly complete the MC dose calculation by inputting the initial dose of PB, thus accelerating the plan optimization process, and, validating the clinical plan. Many single-beam PB and MC doses were generated through virtual planning and then trained using 3D Dense-U-Net. The training process takes the PB single-beam dose as the input, the MC single-beam dose as the label, and MSE as the loss function. Finally, a clinical plan was used for verification, and the Gamma passing rate(GPR) of AI and MC doses were compared. The Gamma criteria was 1mm/1%.

2.1 Clinical Data

A cohort of lung cases was collected for the study. The CT images of the cases were taken by Philips Brilliance Big Bore CT Scanner, with a pixel spacing of 1mm × 1mm × 5mm and slice size of 512 × 512. All cases had a seven-beam IMRT plan made by the Eclipse treatment planning system; the D_{95} of the planning target volume (PTV) for the cases was 60Gy. All plans were clinically implemented, and the RTSTRUCTURE DICOM files were collected.

[+] The authors contributed equally to this work.
[*] Authors to whom correspondence should be addressed.

2.2 Virtual Plan

Virtual plans are automatically generated using CT images of actual patients and a set of optimization rules. Specifically, the optimization rules consist of three parts. The first is to generate a random plan. Compared with clinical plans, random plans bring more possibilities, thereby improving the robustness and generalization of network training. The second is the automatic optimization of random plans. Traditional plan optimization needs to meet specific clinical goals, which is difficult to achieve in a fully automatic way. However, only differential dose distribution intervals need to be generated for random plans without considering the dose in organs at risk. Finally, the dose of each single beam is calculated through the traditional PB and MC dose calculation engine.

The automatic optimization process is based on fluence map optimization(FMO). The optimization operations are preset to generate random plans with various dose distributions. The optional gantry angle settings were [0,40,80,120,160,200,240,280,320] or [20,60,100,140,180,220,260,300,240]; the target area was sequence set at PTV, Lung-L, Lung-R, and esophagus, with isocenter at the geometry center of the target area; the optional wight of optimization for normal tissue were 100, 500, 1000, and the optimization target was 0. Meanwhile, the optimization target of PTV D_{95} was 60Gy, and the weights was 2000. Finally, perform a fixed number of rounds for FMO. It should be noted that such optimization conditions must not be achieved, and due to different beam angles between different plans, the dose obtained will be very different. Even if the two groups of beam angles are similar, because of the weight ratio of normal tissue and target interval, the differences lead to widely varying optimization results.

2.3 Network Structure

The 3D Dense-U-Net was chosen for this research[12, 13], with additional information in figure 1. The CNN utilizes a 5-stage U-Net-like backbone to capture global and local features. It entails four instances of down-sampling through max pooling (stride = $2 \times 2 \times 2$) and four instances of upsampling through deconvolution (kernel shape: $2 \times 2 \times 2$). Each stage of the CNN comprises three 'Denseblocks' with dense connections. The computation flow of a Denseblock entails using Relu, convolution, and batch norm. The convolution kernel size is $3 \times 3 \times 3$, except for the final convolution layer, which uses a $1 \times 1 \times 1$ kernel. Zero padding is applied in each convolution operation. The convolution layers have eight channels that serve as dense connections. As the features exit a Denseblock, the number of feature channels increases by 8.

3 Results

A total of 43 clinical patients' images were used, and 24 virtual plans were generated for each patient. The number of beams in each plan was randomly between 9 and 18 (due to the beam split optimization algorithm). In the end, we obtained 11,278 pairs of PB and MC single-beam doses.
The GPRs (1mm/1%) of training, validating, and testing for the PB and AI-boosting dose, with MC dose as reference. The corresponding mean GPRs for the PB dose were 83.73±6.55%, 84.28±6.55%, and 82.72±6.75%; in contrast, the mean GPRs for the NN-boosting dose were 96.07±2.22%, 96.17±1.90% and 95.31±2.80%. The average passing rate in the test set increased by 15%, the calculation times of an AI-boosting was 0.18 s. The results showed that the CNN significantly improved the accuracy of PB dose, with an ultra-fast speed. The GPRs were summarized in figure 2.

4 Discussion

In the age of large-scale models, the big-size dataset is significant for AI in medicine. In the study, we proposed an auto-generating method for the dataset, which can get nearly infinite training data, based only on a tiny primary dataset with 43 cases. The valuable data generation method can be applied in further research, not only in photon dose calculations. The study only involved lung cases, but it showed that applying the method in building an AI dose calculation model for cases with another cancer site is feasible. In future studies, we look forward to adding cases with other cancer site, such as the head and abdomen. Moreover, the corresponding treatment tech can also be broadened from IMRT to VMAT, SBRT, and even proton therapy. Another significant research in the future is to explore the performance of DL models in TPS, such as an FDA-passed commercial TPS MOZI (Manteia, China), which uses PB in D_{ij} calculations and MC to calculate the final dose. We are supposed to insert our DL model after PB to eliminate the difference between the optimized and final dose, which is toward stable and accurate planning.

5 Conclusion

The study established an auto-generating method for building a training dataset, of dose-boosting DL models in radiation therapy. The method significantly reduced the demand for many clinic-implemented cases but achieved outstanding accuracy.

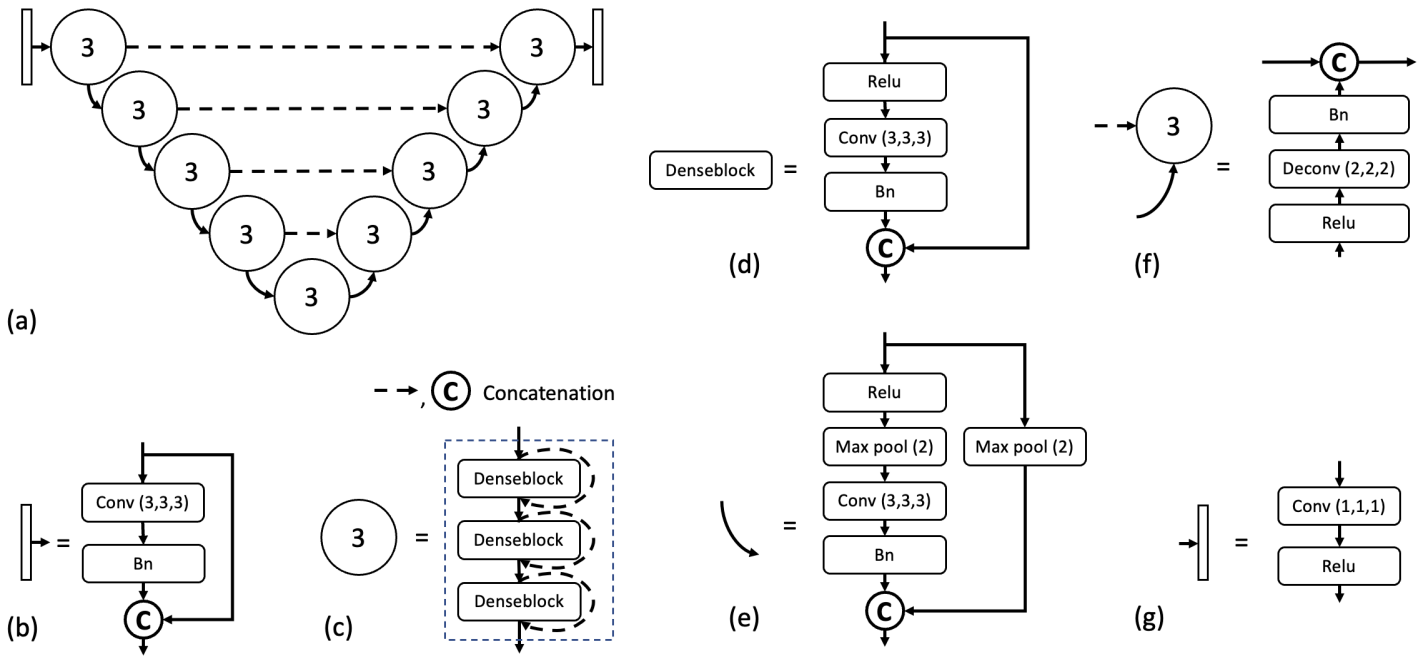


Figure 1: The architecture of the 3D Dense-U-NET. (a) Overview of the 3D Dense-U-NET. (b) The input layers of the CNN are composed of convolution and batch norm operations, along with the concatenation of the input array. (c) Each stage of the CNN consists of three Denseblocks interconnected by dense connections. (d) The flow of a Denseblock includes Relu activation, convolution, and batch norm operations, in addition to the concatenation of input features. (e) The down-sampling process. (f) The up-sampling process entails applying deconvolution to increase feature resolution, followed by concatenation between the up-sampled features and the features from the previous stage. This combined output then flows into the next Denseblock. (g) A $1 \times 1 \times 1$ convolution followed by a Relu activation is performed to generate the single-channel output array.

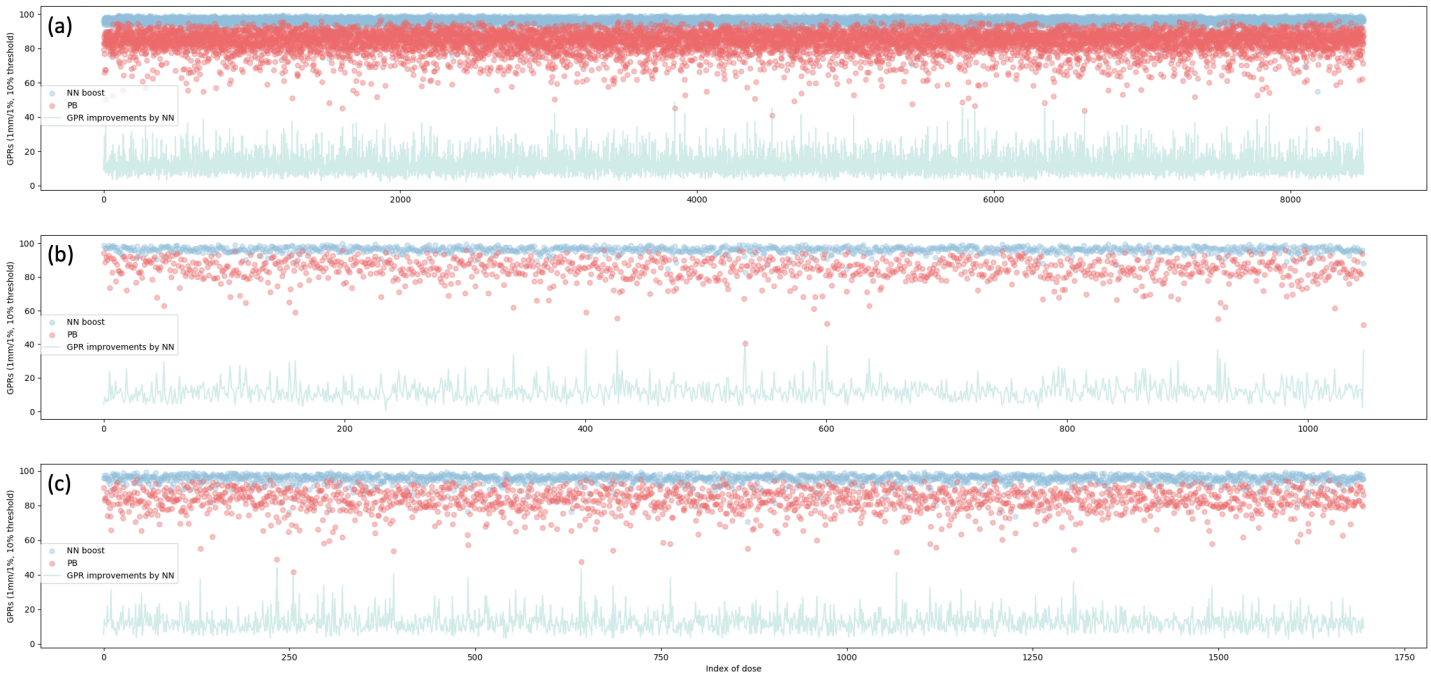


Figure 2: The GPRs of PB and NN boosting dose, with MC dose as reference. (a) training dataset, (b) validating dataset, (c)testing dataset.

References

[1] R. L. Siegel, K. D. Miller, N. S. Wagle, et al. "Cancer statistics, 2023". *Ca Cancer J Clin* 73.1 (2023), pp. 17–48.

[2] R. Baskar, K. A. Lee, R. Yeo, et al. "Cancer and radiation therapy: current advances and future directions". *International journal of medical sciences* 9.3 (2012), p. 193.

[3] T. Hong, M. Ritter, W. Tome, et al. "Intensity-modulated radiation therapy: emerging cancer treatment technology". *British journal of cancer* 92.10 (2005), pp. 1819–1824.

[4] S. A. Leibel, Z. Fuks, M. J. Zelefsky, et al. "Intensity-modulated radiotherapy". *The Cancer Journal* 8.2 (2002), pp. 164–176.

[5] J. Sempau, S. J. Wilderman, and A. F. Bielajew. "DPM, a fast, accurate Monte Carlo code optimized for photon and electron radiotherapy treatment planning dose calculations". *Physics in Medicine & Biology* 45.8 (2000), p. 2263.

[6] M. Karbalaee, D. Shahbazi-Gahrouei, and M. B. Tavakoli. "A novel GPU-based fast Monte Carlo photon dose calculating method for accurate radiotherapy treatment planning". *Journal of Biomedical Physics & Engineering* 10.3 (2020), p. 329.

[7] U. Jeleń, M. Söhn, and M. Alber. "A finite size pencil beam for IMRT dose optimization". *Physics in Medicine & Biology* 50.8 (2005), p. 1747.

[8] X. Gu, U. Jelen, J. Li, et al. "A GPU-based finite-size pencil beam algorithm with 3D-density correction for radiotherapy dose calculation". *Physics in Medicine & Biology* 56.11 (2011), p. 3337.

[9] Y. Xing, D. Nguyen, W. Lu, et al. "A feasibility study on deep learning-based radiotherapy dose calculation". *Medical physics* 47.2 (2020), pp. 753–758.

[10] B. Yang, Y. Liu, Z. Chen, et al. "Tissues margin-based analytical anisotropic algorithm boosting method via deep learning attention mechanism with cervical cancer". *International Journal of Computer Assisted Radiology and Surgery* 18.5 (2023), pp. 953–959.

[11] Y. Wang, Y. Liu, Y. Bai, et al. "A generalization performance study on the boosting radiotherapy dose calculation engine based on super-resolution". *Zeitschrift für Medizinische Physik* (2023). DOI: 10.1016/j.zemedi.2022.10.006.

[12] Y. Liu, Z. Chen, J. Wang, et al. "Dose prediction using a three-dimensional convolutional neural network for nasopharyngeal carcinoma with tomotherapy". *Frontiers in Oncology* 11 (2021), p. 752007.

[13] Y. Peng, Y. Liu, Z. Chen, et al. "Accuracy Improvement Method Based on Characteristic Database Classification for IMRT Dose Prediction in Cervical Cancer: Scientifically Training Data Selection". *Frontiers in Oncology* 12 (2022), p. 808580.

Development and evaluation of novel parameters for describing anatomical changes and predicting radiotherapy replanning for head and neck cancer patients

Odette Rios-Ibacache[1], James Manalad[1], Aixa X. Andrade Hernandez[2], Kayla O'Sullivan-Steben[1], Emily Poon[1], Luc Galarneau[1,3], and John Kildea[1,3]

[1]Medical Physics Unit, McGill University, Montreal, QC, Canada
[2]Department of Bioinformatics, University of Texas Southwestern Medical Center, Dallas, TX, USA
[3]Research Institute of the McGill University Health Centre, McGill University, Montreal, QC, Canada

Abstract Head and neck cancer patients under radiotherapy treatment are prone to anatomical changes due to the effects of weight loss on non-rigid structures, such as inflammation and tumor shrinkage. Across several treatment fractions, these changes may become significant enough to render the initial treatment plan invalid and require treatment replanning. However, replanning often demands a high amount of resources and may be poorly timed in radiotherapy workflows with competing priorities. To address this, we aim to identify and generate parameters extracted from three-dimensional radiotherapy structures, that can describe anatomical alterations and can affect the decision for replanning. These parameters are intended to be used in the development of a machine learning classification model to predict if and when patients should undergo replanning, taking into account the progression of the information over time. In our study, we have considered a cohort of 105 head and neck cancer patients, and more than 20 parameters have been created. Currently, we can begin classifying patients as requiring replanning or not from the first week of treatment.

1 Introduction

Head and neck cancer (H&N) is a common disease in many regions of the world, accounting for 900,000 cases and over 400,000 deaths annually [1]. Primary risk factors include tobacco use, alcohol consumption, the human papillomavirus infection, and the Epstein-Barr virus infection [2].

When H&N cancer patients go under radiotherapy (RT), they may experience significant anatomical changes, a consequence of associated side effects, that may invalidate the original treatment plan due to unacceptable variations in dose coverage in normal structures and defined target volume. Image-guided radiotherapy techniques have made possible the identification of the changes with considerable reliability and treatment replanning is regularly done in the clinic.

H&N RT replanning is relevant as it may maintain tumor coverage leading to tumor control, and it may improve patient quality of life by helping to avoid toxicity [3]. However, there has been no definitive method established to define the amount of weight loss or anatomical change in the H&N region that necessitates a replan, even though the replanning process requires a significant investment of effort and is usually delayed, which can impact many patients [4]. Therefore, it is important to identify as early and as reliably as possible which patients will most likely need replanning.

This research aims to develop and evaluate parameters related to anatomical structures, that are used and known in the clinic or related to them, to predict if and when patients will need replanning ultimately.

2 Materials and Methods

2.1 Patient and clinical data characteristics

Our retrospective study incorporates 105 H&N patients treated at McGill University Health Centre within the RT department from 2017 to 2023, who were diagnosed and biopsy-proven with Squamous Cell Carcinoma (HNSCC) or Adenocarcinoma, and with different sites of diagnosis, such as Oropharynx (52.4%) and Nasopharynx (8.57%). This cohort consists of female (28.6%) and male (71.4%) patients who were treated with a total prescribed dose of 60 Gy (10.5%), 66 Gy (3.8%), and 70 Gy (85.7%), with curative or palliative treatment intent, and delivered with a conventional dose of 2 Gy per fraction. Treatment regimes included radiotherapy alone as well as combined with systemic therapies, such as chemotherapy and targeted therapy.

Computed Tomography simulation (CT sim) images, taken at the beginning of the treatment planning, and Cone Beam CT (CBCT) images, taken on a daily or almost daily basis, were collected from the treatment planning system at our institution. Each patient was required to have had a CT sim image and at least 2 CBCT images taken on two distinct fractions. CBCT images were used during treatment to track patients' anatomical changes and to identify if a new plan was needed. Our initial sample size counted 61 replanned patients and 44 non-replanned patients. The latest fraction at which patients were replanned in this cohort was 25, while the most frequent fraction for replanning was 15.

At present, additional patient data are being assessed to reach an intended sample size of around 300 patients.

2.2 Anatomic 3D structure and parameters definition

For each patient, the CT sim image along with its structures was registered with each CBCT image. Each patient's

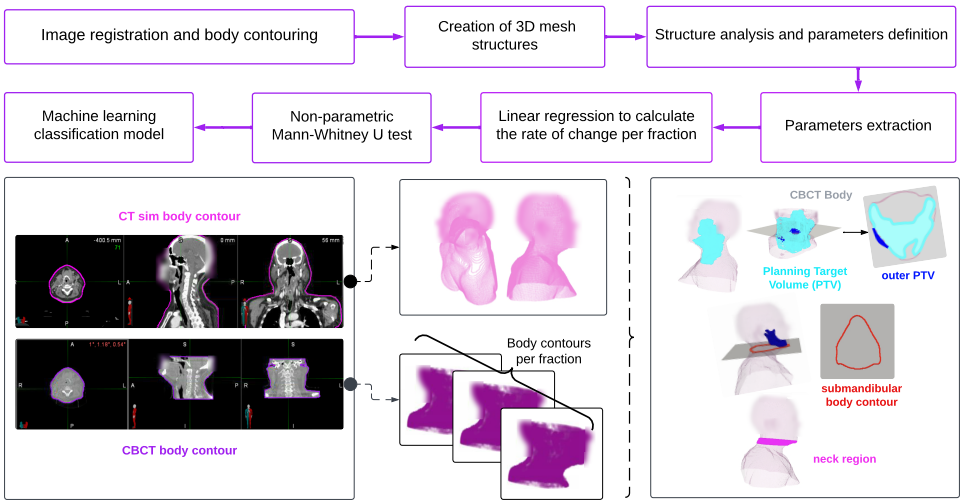


Figure 1: Pipeline diagram of the presented work. Images have been defaced to preserve patient anonymity.

outer body structure was generated by the commercial image registration software MIM MAESTRO (MIM Software Inc. Cleveland OH), using its automatic body contour and thresholding tool. Each slice was examined to ensure that the structure was correctly encompassed. The files were saved in a DICOM-RT Structure Set format.

To obtain geometrical information from the RT structures, we used `Python v3.7` and the open-source library `Pyvista v0.43.2`. Certain structures were rendered in a 3D format by a mesh creation of the corresponding RT structure that allowed us to analyze them interactively, and the creation of surfaces by using a surface marching cubes algorithm.

Based on the 3D shape and 2D contours of the body and Planning Target Volume (PTV) structures, we defined a series of parameters to describe anatomical changes throughout treatment delivery. These parameters were intended to describe body shape features associated with weight loss, such as skin separation, and the body shrinkage between treatment fractions, by calculating the Chamfer distance. We also included parameters related to the PTV, to detect when the target was not being covered correctly for treatment. For instance, by examining the minimum distance between the PTV and the body contour. A complete list of the parameters that were created and studied is presented in Table 1.

2.3 Univariate statistical analysis

Incorporating the data of all patients, a statistical analysis was performed for each extracted parameter, involving the calculation of the rate of change of the parameter to see how it changed across fractions. As it was discovered in previous studies [5, 6], weight loss and decrease in neck area are more prominent for replanned patients than non-replanned patients. Thus we hypothesized that the rate at which the parameters change may provide useful information. A linear regression model was applied by using the open-source library `skcit-learn` to obtain the slope value that the parameter had obtained up to each fraction. Thus every parameter p presented in Table 1, had a fraction-specific slope version of it, denoted as $\dot{p}$.

Graphs were generated to visually represent the longitudinal data across treatment fractions and to show the progression of the parameters over time. Each parameter was analyzed using the non-parametric Mann-Whitney U test to determine if the distributions corresponding to the classes replanned vs non-replanned were statistically different. A Bonferroni correction was applied due to the application of several hypothesis tests (p-value$<$ 0.002, 24-25 tests) [7]. Preliminary classification models were built using the Logistic Regression classifier: at the 1st fraction that successfully passed the test and at fraction 15. A Repeated (n = 100) Stratified 5-fold Cross-Validation technique was used, with the Area Under the ROC Curve (AUC) as a training metric. The 95% Confidence Interval (C.I) was calculated.

3 Results

By applying the pipeline presented in Figure 1, a series of statistical results was obtained. The most important results are presented in Figure 2. In general, we observe that the averages of the longitudinal data of the two classes behave differently, indicating that the parameters can be used to characterize the classes; replanned versus non-replanned. We noticed that the parameters $\dot{R}_{mean}$ and $\dot{VO}_{PTV}$- Body ratio (a particular case of VO_{PTV} normalized by the volume of the body) presented significant differences between replanned and non-replanned from an early fraction with p-value$<$ 0.002 (after Bonferroni correction), starting from as early as fraction 3 in the case of $\dot{VO}_{PTV}$- Body ratio and fraction 5 for $\dot{R}_{mean}$, corresponding to the 1st week of treatment. The statistical difference tends to increase at higher fractions. Thus, we can state that it should be possible to start predicting whether some patients may need a replanning from fraction

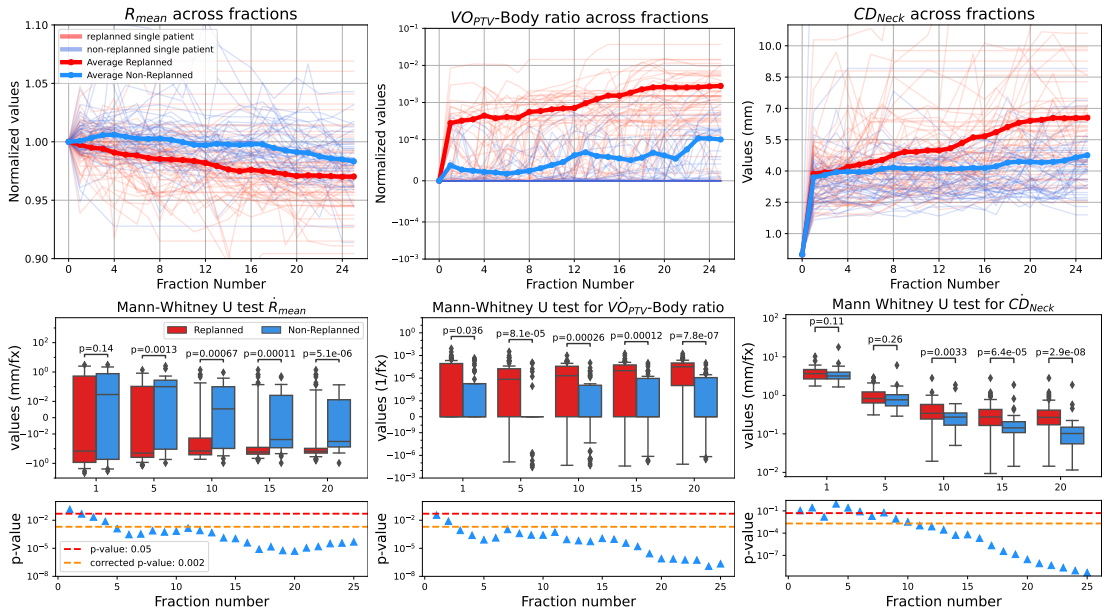


Figure 2: Results obtained by our analysis. In the top row, the longitudinal data pertaining to three parameters are presented, and in the middle and bottom rows, the results from the Mann-Whitney U test for the rate of change of the parameters at five specific treatment fractions and p-values for all treatment fractions are shown respectively.

3. The estimated AUC values for $\dot{VO}_{PTV}$-Body ratio were 0.67 [0.52-0.84] for fraction 3. While $\dot{R}_{mean}$ showed an AUC value of 0.68 [0.52-0.88] for fraction 5.

We also noticed that the $\dot{CD}_{Neck}$ showed statistically significant differences between the replanned and non-replanned populations from fraction 11 (3rd week), with an estimated AUC of 0.69 [0.52-0.88]. A preliminary multivariate model using the three parameters mentioned above at fraction 15, resulted in an AUC of 0.82 [0.62-0.97].

4 Discussion

Head and neck cancer patients are subject to significant anatomical changes during treatment due to tumor shrinkage and weight loss. Owing to clinical workflow and economic factors, it is important to identify as early as possible which patients are likely to be replanned and when. Our research aims to find standardizable metrics that may distinguish replanned and non-replanned patients and combine them in multivariate machine learning models. These models should help clinical personnel to determine from as early as possible which patients will need treatment replanning and when.

As part of our ongoing and upcoming work, we are examining additional parameters that may be associated with weight loss. This includes the contour of the radiotherapy head mask relative to the body contour, which has the potential to not only track weight loss but also detect when the device is not fitting properly for treatment. Moreover, we are considering the movement of the parotid glands as a potential parameter in our study, as it has been shown that their spatial orientation appears to shift during treatment [4].

5 Conclusion

The combination of 3D and 2D geometrical information from radiotherapy structures, related to anatomical changes, has the potential to serve as a predictor for radiotherapy replanning.

Acknowledgements

This research was supported by the Rossy Cancer Network and by the CREATE Responsible Health and Healthcare Data Science (SDRDS) grant from the Natural Sciences and Engineering Research Council. The authors would like to acknowledge M.D. Julia Khriguian for her contribution to this research.

References

[1] The International Agency for Research on Cancer (IARC). *Global Cancer Observatory*. en. https://gco.iarc.fr/. Accessed: 2024-1-20.

[2] A. Barsouk, J. S. Aluru, P. Rawla, et al. “Epidemiology, risk factors, and prevention of head and neck squamous cell carcinoma”. en. *Med. Sci. (Basel)* 11.2 (June 2023).

[3] L. A. Dawson and M. B. Sharpe. “Image-guided radiotherapy: rationale, benefits, and limitations”. en. *Lancet Oncol.* 7.10 (Oct. 2006), pp. 848–858.

[4] H. E. Morgan and D. J. Sher. “Adaptive radiotherapy for head and neck cancer”. en. *Cancers Head Neck* 5.1 (Jan. 2020), p. 1.

[5] A. Andrade-Hernandez and J. Kildea. *Exploration of Potential Parameters that Influence when to Replan Head and Neck Cancer Patients during Radiotherapy*. 2020.

[6] L. Capelle, M. Mackenzie, C. Field, et al. “Adaptive radiotherapy using helical tomotherapy for head and neck cancer in definitive and postoperative settings: initial results”. en. *Clin. Oncol. (R Coll. Radiol.)* 24.3 (Apr. 2012), pp. 208–215.

[7] R. A. Armstrong. “When to use the Bonferroni correction”. en. *Ophthalmic Physiol. Opt.* 34.5 (Sept. 2014), pp. 502–508.

Table 1: Parameter definitions that were created to analyze the anatomical changes of replanned and non-replanned patients. The neck region was based on the common region between z = 0 mm and z = -18 mm, referenced to the plane of the submandibular contour.

Parameter	Symbol	Definition
PTV related		
Minimum distance from PTV to Body	x_{min}	Minimum value of the distances measured from the planning target volume drawn in the CT sim to a body contour
Average distance from PTV to Body	x_{ave}	Average value of the distances measured from the planning target volume drawn in the CT sim to a body contour
Median distance from PTV to Body	x_{med}	Median value of the distances measured from the planning target volume drawn in the CT sim to a body contour
Volume PTV outer	VO_{PTV}	Volume of the PTV section outside a body contour
Volume PTV inner	VI_{PTV}	Volume of the PTV section inside a body contour
Body related		
Hausdorff distance	HD_{Body}	Maximum distance between the body contour of the CT sim and a body contour
Chamfer distance	CD_{Body}	Average distance between pair nearest neighbor points from the body contour of the CT sim to a body contour
Average distance (2D)	$\bar{D}_{Body}$	Average distance between a body contour and the CT sim body contour, across each plane z
Maximum distance (2D)	D_{Body}	Maximum distance between a body contour and the CT sim body contour, calculated across each plane z
Body volume	V_{Body}	Volume enclosed by a body contour
Neck related		
Neck volume	V_{Neck}	Volume of a defined neck region
Minimum radius (3D)	R_{min3D}	Minimum 3D radius between the center of CT sim's neck region and a neck region
Maximum radius (3D)	R_{max3D}	Maximum 3D radius between the center of CT sim's neck region and a neck region
Mean radius (3D)	R_{mean3D}	Average 3D radius from the radius distribution between the center of the CT sim's neck region and a neck region
Chamfer neck distance	CD_{Neck}	Chamfer distance between CT sim's neck region and a neck region
Hausdorff neck distance	HD_{Neck}	Maximum distance between CT sim's neck region and a neck region
Surface area	SA_{Neck}	The amount of space enclosing the outside of a neck region
Compactness	C_{Neck}	Ratio between the surface area and the volume of a neck region
Submandibular plane related		
Area	$Area$	Total space covered by the submandibular body contour plane
Minimum radius (2D)	R_{min}	Minimum radius between the center of the CT sim submandibular body contour plane and a submandibular body contour plane
Maximum radius (2D)	R_{max}	Maximum radius between the center of the CT sim submandibular body contour plane and a submandibular body contour plane
Mean radius (2D)	R_{mean}	Average radius between the center of the CT sim submandibular body contour plane and a submandibular body contour plane
Maximum longitudinal chord	L_y	Maximum distance that can be drawn from the outermost points of a submandibular body contour plane in a longitudinal direction
Maximum lateral chord	L_x	Maximum distance that can be drawn from the outermost points of a submandibular body contour plane in a lateral direction

Software-Assisted Optimisation of Linear Accelerator Energy Consumption

Genotan Reggian[1] and Catherine Stanford-Edwards[1]

[1]South West Wales Cancer Centre, Swansea Bay University Health Board, Swansea, United Kingdom

Abstract

In the pursuit of environmental sustainability within healthcare, this study introduces a pioneering approach to optimise the energy consumption of linear accelerators using a software-based approach. Through a comprehensive analysis at the South West Wales Cancer Centre, the energy utilisation of various linac models was scrutinised, as well as stereotactic ablative body radiotherapy (SABR) and hypofractionated radiotherapy (HFRT) treatment methods. The findings reveal that SABR and HFRT treatments exhibit significantly lower energy consumption compared to conventional methods, indicating the potential for substantial energy savings. Furthermore, the newer Agility and Versa HD Elekta linacs had significantly decreased energy consumption compared to the older Precise linac demonstrating that replacement of linac technology can also lead to substantial energy savings. Further research and collaboration targeted at sustainability in radiotherapy can facilitate the optimisation of treatment delivery while maintaining a commitment to sustainability within radiotherapy.

1 Introduction

The UK's National Health Service (NHS) is at the forefront of integrating sustainability into healthcare delivery, aligning with global objectives. This aim is particularly pertinent in the context of linear accelerators (linacs), which, although being a fundamental method of treatment of cancer through radiotherapy, is a significant contributor to the NHS's carbon footprint.

The need to reduce the NHS's environmental impact is set by legislative requirements to deliver a healthcare service with a 'net-zero' carbon footprint by 2040 [1]. The impetus of this legally binding agreement arises from the NHS's carbon footprint being 4-5% of the UK's carbon footprint and 40% of the UK's public sector emissions and therefore a 'net-zero' NHS would significantly decrease national emissions [2].

Linacs are a pivotal technology in cancer treatment, not only being an advanced medical treatment but also as a focal point for energy consumption optimization. The operation of these machines is energy-intensive, with varying power demands depending on their operational states. As 2040 approaches, by which the NHS aims to operationalise its 'net-zero' commitment, the need to interrogate the energy usage of linacs has become increasingly pressing.

Recognising the dual necessity of maintaining clinical efficiency while enhancing energy efficiency, this study introduces an innovative software-based approach to monitor and optimize the power usage of linacs. Doing so addresses the critical gap in the management of healthcare technologies. This initiative is not only a technological challenge but instead represents a larger paradigm shift towards a greener, more sustainable healthcare model that other healthcare providers globally can implement.

2 Materials and Methods

The core of this research is the Linac Energy and Power Analysis Software (LEPAS), an innovative desktop application designed to utilise data analytics in pursuit of optimising the energy efficiency of linacs. LEPAS was engineered with a Python-based graphical user interface using a cross-platform Qt toolkit that provides a user-friendly point of interaction. Through the system, energy consumption data was captured across various operational linac states such as standby and treating states – each having unique energy profiles and patterns.

Data acquisition was facilitated by the Fluke 1738 3-Power Logger, an instrument selected for its precision and reliability. The Fluke Logger was configured to capture energy usage at one-second intervals, providing a high-resolution data stream, critical for assessing the dynamics of linac energy consumption as time passed. This granular data capture allowed for highly detailed analysis, enabling the identification of peak energy consumption periods as well as opportunities for energy-saving interventions [3].

Treatment data was extracted from the oncology information system Mosaiq using Crystal Reports exports. The data retrieved from Mosaiq consisted of multiple variables, including but not limited to, treatment techniques, beam energies and numbers of fractions. Additionally, linac data was procured from linac log files which were decrypted by Elekta. The linac-specific data yielded insights into variables such as operational states, gantry angles, and monitor units among others.

The LEPAS system facilitated the synchronisation of data streams from energy consumption, treatment details and linac operations. A rigorous time calibration process to ensure the integrity and accuracy of the synchronisation process. Subsequent to synchronisation, LEPAS provided functionalities for data filtration predicated on various treatment and linac parameters which enabled the generation of a filtered dataset. LEPAS' analytical capabilities were leveraged to interrogate the filtered dataset. This process using LEPAS allowed the simple

analysis of different treatment options, machines and other variables.

The data utilised within the scope of this paper was sourced from the South West Wales Cancer Centre (SWWCC). At the time of data collection, the centre was equipped with three Elekta linac models: Agility, Versa HD and a Precise model fitted with an Agility head. An analytical comparison of energy consumption across these models was conducted to determine their relative energy efficiencies.

In addition to conventional external beam radiotherapy, the centre also regularly performed stereotactic ablative body Radiotherapy (SABR) lung treatments and hypofractionated radiotherapy (HFRT) prostate treatments. An analytical comparison of energy consumption was performed between SABR and conventional lung treatments, as well as between HFRT and conventional prostate treatments. The energy consumption of both SABR lung treatments and HFRT prostate treatments were compared to their conventional counterparts.

For the purposes of the comparisons, only Volumetric Modulated Arc Therapy (VMAT) treatments with an energy of 6 MV were considered. Active energy consumption was sampled at one-second intervals to ensure high-resolution data capture.

3 Results

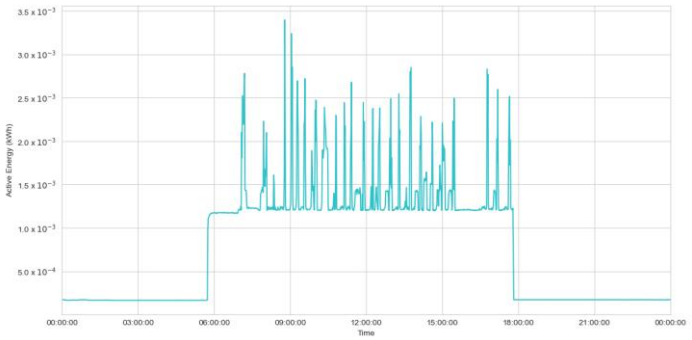


Fig. 1: Line graph displaying the active energy during a treatment day at the SWWCC. Times of interest include linac engineers' run-up at 07:00, radiographers' QC at 07:50, DIBH treatment at 10:15 and SGRT staff training at 15:30.

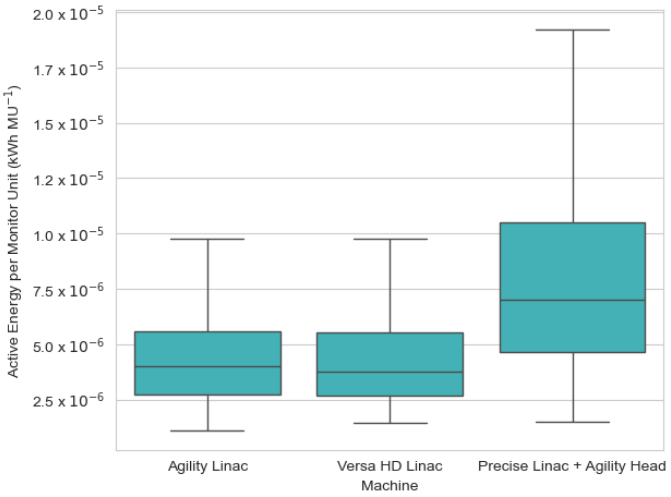


Fig. 2: Box-and-whisker plot displaying the distribution of active energy per monitor unit for three different types of Elekta linacs used at SWWCC.

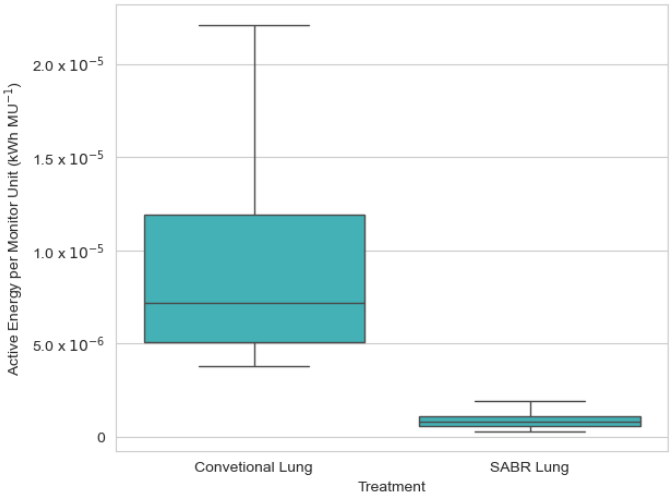


Fig. 3: Box-and-whisker plot displaying the distribution of active energy per monitor unit for conventional lung and SABR lung treatments.

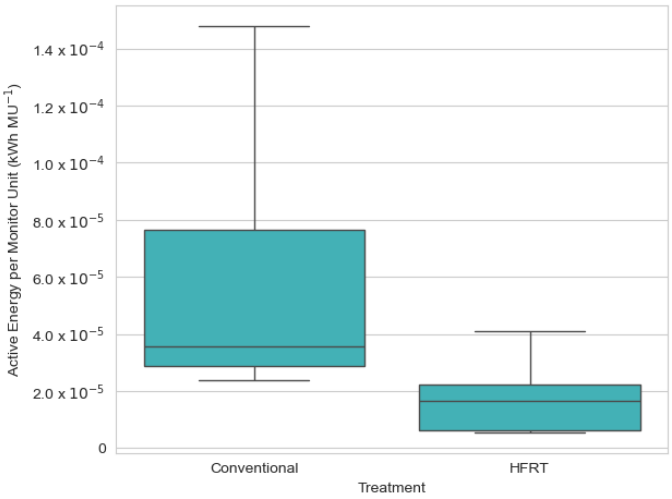


Fig. 4: Box-and-whisker plot displaying the distribution of active energy per monitor unit for conventional prostate and HFRT prostate treatments.

4 Discussion

The analysis of energy consumption during a typical treatment day (Fig. 1) illustrated energy demands during

deep inspiration breath-hold (DIBH) treatment techniques was larger than conventional treatments, suggesting that patient-specific treatment adaptations could yield energy savings assuming that clinical effectiveness is not affected. This is likely since between breaths, the surface-guided radiotherapy (SGRT) system forces the beam to pause until the next breath-hold which consumes more energy than if the beam was simply initiated and terminated [4].

Examining the energy consumption patterns among linac models reveals that the newer iterations, such as the Agility and Versa HD, exhibit much more favourable energy consumption profiles than the Precise. The improvements in energy consumption in newer models are likely due to technological improvements, and newer components.

SABR treatments for lungs demonstrated a substantial reduction in active energy consumption per monitor unit in comparison to conventional lung treatments. This observation could be attributed to the omission of a flattening filter in SABR treatments. The inclusion of a flattening filter in conventional treatments necessitates increased beam energy to achieve the desired therapeutic dose, as the flattening filter's presence attenuates the beam [5].

Similarly, HFRT treatments for prostates also demonstrated a reduction in active energy consumption per monitor unit per treatment course in comparison to conventional prostate treatments. The increased efficiency could be due to the reduced frequency of initiation and termination sequences when fewer fractions are used [6]. HFRT treatments not only align with sustainable healthcare practices by reducing the energy expenditure for treatments but also reduce emissions from patient travel. This, in turn, significantly enhances patient convenience and can make radiotherapy treatments more available to patients by increasing machine availability [7].

Despite the clear advantages that HFRT treatments demonstrate, careful implementation is necessary to ensure treatment efficacy is not compromised by the reduction in the number of fractions [8]. Brunt et al. (2020) found that HFRT for breast patients had an equal efficacy as conventional treatments [9]. The guiding principle should be to administer a minimal number of fractions while maintaining treatment effectiveness and patient safety.

These outcomes highlight the potential for software tools like LEPAS to facilitate more nuanced areas concerning medical technologies. Such tools can provide healthcare administrators and policymakers with actionable insights, enabling the development and innovation of departments to reduce the total energy consumption within departments.

The aggregation of datasets from a multitude of treatment centres presents a unique opportunity to broaden the comparative framework to include an array of treatment methodologies. It would also facilitate the evaluation of emergent advanced technologies such as magnetic resonance linacs (MR-linacs) and proton linacs. Upon full development of LEPAS, access to the private repository can be extended to other departments. This collaboration would lead to a robust database that is reflective of the diverse array of treatments that centres offer so the radiotherapy community can optimise treatment delivery while maintaining a commitment to environmental sustainability within radiotherapy.

5 Conclusion

In conclusion, LEPAS represents an instrumental advancement for the energy optimisation of linacs. This study has provided empirical evidence of its potential to significantly inform energy management strategies in radiotherapy thereby supporting the NHS in its net-zero aim. As LEPAS and other energy consumption-focused software evolve, they will become essential components of the sustainable healthcare infrastructure, facilitating not only immediate operational efficiencies but also long-term strategic planning for environmental sustainability in healthcare both within the UK and internationally.

6 Acknowledgements

Stephen Herbert for setting up the Fluke Logger and retrieving energy data

Becky Bowen for creating the Crystal Report to retrieve treatment data

7 Conflicts of Interest

This project has been funded by Elekta.

8 References

[1] NHS England. (2022). Delivering a net zero NHS. Retrieved from: https://www.england.nhs.uk/greenernhs/a-net-zero-nhs

[2] British Medical Association. (2023). More support needed to help the NHS reach net zero. Retrieved from: https://www.bma.org.uk/news-and-opinion/more-support-needed-to-help-the-nhs-reach-net-zero

[3] Chuter, R., Stanford-Edwards, C., Cummings, J., Taylor, C., Lowe, G., Holden, E., Razak, R., Glassborow, E., Herbert, S., Reggian, G., Mee, T., Lichter, K., & Aznar, M. (2023). Towards estimating the carbon footprint of external beam radiotherapy. Physica Medica, 112(102652). DOI: 10.1016/j.ejmp.2023.102652

[4] Chen, L., Bai, S., Li, G., Li, Z., Xiao, Q., Bai, L., Li, C., Xian, L, Hu, Z., Dai, G., & Wang, G. (2020). Accuracy of real-time respiratory motion tracking and time delay of gating radiotherapy based on optical surface

imaging technique. Radiation Oncology, 15(170). DOI: 10.1186/s13014-020-01611-6

[5] Kragl, G., Baier, F., Lutz, S., Albrich, D., Dalaryd, M., Kroupa, B., Wiezorek, T., Knöös, T., & Georg, D. (2011). Flattening filter free beams in SBRT and IMRT: dosimetric assessment of peripheral doses. Zeitschrift für Medizinische Physik, 21(2), pp. 91–101. https://doi.org/10.1016/j.zemedi.2010.07.003

[6] Salas-Salas, B. G., Domínguez-Nuez, D. J., Cabrera, R., Ferrera-Alayón, L., Lloret, M., & Lara, P. C. (2020). Hypofractionation vs. conventional radiotherapy fractionation in the conservative treatment of T1 glottic cancer: a prospective cohort study. Clinical & Translational Oncology, 22(1), pp. 151–157. DOI: 10.1007/s12094-019-02118-7

[7] Swanson, W., Samba, R., Lavelle, M., Elzawawy, A., Sajo, E., Ngwa, W., & Incrocci, L. (2021). Practical Guidelines on Implementing Hypofractionated Radiotherapy for Prostate Cancer in Africa. Frontiers in Oncology, 11(725103). DOI: 10.3389/fonc.2021.725103

[8] Wong, K., Delaney, G., & Barton, M. (2016). Evidence-based optimal number of radiotherapy fractions for cancer: A useful tool to estimate radiotherapy demand. Radiotherapy and Oncology, 119(1), pp. 145–149. DOI: 10.1016/j.radonc.2015.12.001

[9] Brunt, A., Haviland, J., Wheatley, D., Sydenham, M., Alhasso, A., Bloomfield, D., Chan, C., Churn, M., Cleator, S., Coles, C., Goodman, A., Harnett, A., Hopwood, P., Kirby, A., Kirwan, C., Morris, C., Nabi, Z., Sawyer, E., Somaiah, N., Stones, L., Syndikus, I., Bliss, J., Yarnnold, J. & FAST-Forward Trial Management Group (2020). Hypofractionated breast radiotherapy for 1 week versus 3 weeks (FAST-Forward): 5-year efficacy and late normal tissue effects results from a multicentre, non-inferiority, randomised, phase 3 trial. The Lancet, 395(10237), pp. 1613–1626. DOI: 10.1016/S0140-6736(20)30932-6

Deep reinforcement learning for custom applicator design in personalised high dose rate brachytherapy

Cédric LOISEAU[1], Philippe BEREJNY[1], Marlon SILVA[2], Juliette THARIAT[2], and Alain BATALLA[1]

[1]Department of Medical Physics, Comprehensive Cancer Centre François Baclesse, Caen, France
[2]Department of Radiotherapy, Comprehensive Cancer Centre François Baclesse, Caen, France

Abstract Brachytherapy is a radiotherapy technique that uses radiation from a radioactive source to treat a tumour. In personalised brachytherapy, the source is guided by an applicator adapted to the patient's anatomy and the positions of the source (PoS). The dwell time (DT) associated with the PoS are the spatial and temporal parameters to be optimised, which govern the dose distribution. This distribution must comply with dosimetric constraints: coverage of the tumour and minimum dose to organs at risk. Usually, to meet the dosimetric and geometric constraints of applicator manufacture, the definition of PoS and DT is long and laborious. Our project exploits deep reinforcement learning models and 3D printing technology to design custom applicators. For a simplified case of a virtual patient, we have demonstrated the feasibility of our model, trained on virtual data, to offer PoS by dealing with spatial and temporal optimisation to meet constraints.

1 Introduction

Brachytherapy is a radiotherapy technique in which a radioactive source is placed within a body cavity or inside the tissue to be close to the tumour. It allows steep dose gradients and protects healthy tissues very well due to accurate source positioning and the inverse square law [1]. To improve brachytherapy treatments, personalised applicators can be generated using 3D printing. Moreover, their design phases can be optimised using artificial intelligence (AI) based methods. The applicator is a device that allows the source to be positioned according to the patient's anatomy. The degree of customisation varies from standardised modular applicators to customised personalized applicators made to conform to each patient's analtomy with a high fidelity. The Comprehensive Cancer Centre François Baclesse (CFB), located in Normandy (Caen-France), has a 30-year long expertise in customised brachytherapy and has been involved in innovations including AI to improve patient care.

The current state of the art for customised brachytherapy treatments is the use of an applicator tailored to the patient's anatomy [2]. The radioactive sources can be guided along multiple source paths within applicators. The sources stop at defined positions (PoS position of source) along each path. The problem with personalised brachytherapy is to produce an applicator for a given patient that satisfies several constraints: geometric constraints, dosimetric constraints and constraints on the fabrication of the final applicator, which has to comply with medical device regulations. At present, the manual production of an applicator that complies with all constraints involves two optimization stages (yellow box in figure 1).

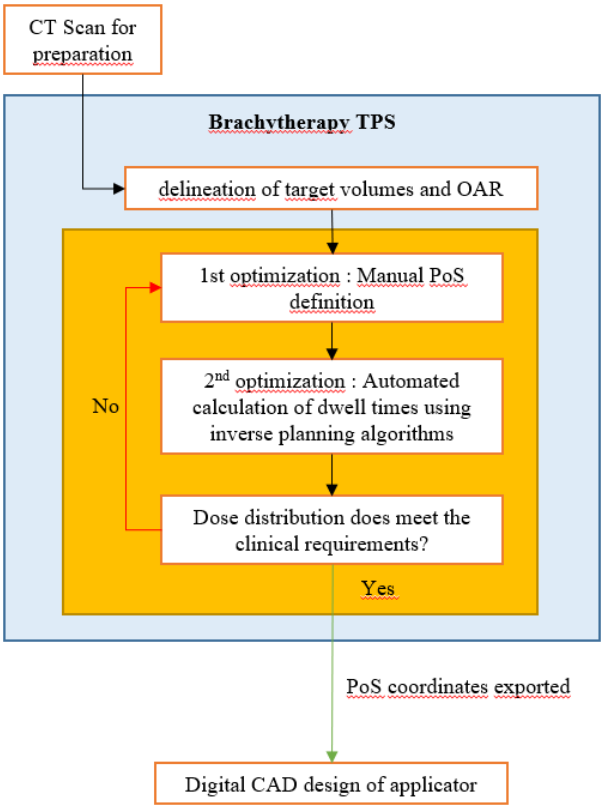


Figure 1: The calculation of PoS and DT is carried out in CE-certified dose calculation software (treatment planning system TPS). The steps are shown in a yellow box. When done manually, they are time consuming and laborious. Our AI model aims to automate these two steps.

The geometric constraints are the distances between the PoS and the topology of the treatment area (tumour and healthy organs/tissues). The first step is therefore to propose a solution. This solution defines PoS that respect the geometric/anatomic and dose-distance and manufacturing constraints (mechanical properties of materials and 3D printers). For example, the source may not circulate if one of the paths is too curved. Due to the lack of dedicated tools, this phase is currently done manually but is time-consuming and laborious. It is necessary to work empirically and iteratively on the applicators before clinical application. There is no guarantee that the result will be optimal. At this stage, there is no informa-

tion about the distribution of the dose, so the expert works by experience and deduction, assuming that the applicator will fulfill the dosimetric needs. This is the first stage of the optimisation process.

A second step is necessary to find the source stop times that respect the dosimetric constraints. Current brachytherapy planning involves constraints on both the tumor and surrounding healthy tissues. For example in the case of treatment of a skin tumor, international recommendations are to deliver at least 36 Gy to the tumor but less than 140% of 36GY to avoid necrosis and as little as possible to healthy skin [3, 4]. This means that the dose distribution has to be shaped by adjusting DT of each PoS. This means that the dose distribution must be shaped by adjusting the DT of each PoS. Two elements are involved in calculating the dose, the source-point of interest distance and the stopping time. The variation with the stopping time is linear, while the dose varies inversely squared with the distance [1]. Solving this requires complex mathematical methods[5]. Clinicians and physicists are therefore assisted by certified CE-marked dose calculation softwares relying on mathematical algorithms chosen by software vendors. This is the second step in the optimisation process as does today.

The third step is to evaluate the final result and check that all the constraints have been met: If they are met, the manufacturing process is initiated to build the applicator. If they are not, the current process requires that physicists and clinicians go back to the first step and test a different PoS solution. The full process is therefore time-consuming and depends on the experience of the operators.

The H3DMED project stems from the need to streamline the previously described processes by leveraging AI.

2 Materials and Methods

Up to now, no research has successfully amalgamated the two optimization steps into a single process, primarily due to their inherent complexity. The multitude of constraints, both temporal and spatial, poses a significant challenge, with the risk of computational explosion being a major concern. Addressing this challenge constitutes a central point of the H3DMED project.

Conventional machine learning methods require substantial datasets for effective learning. However, in the realm of medical data, stringent privacy regulations pose formidable legal hurdles. Implementation necessitates meticulous procedures, including impact assessments to mitigate data risks, obtaining patient consent, and potentially securing regulatory approvals from commissions such as "CNIL" and "CESREES" in France. Moreover, even with adherence to these protocols, access to data remains limited to information managed by medical institutions, often insufficient for robust model training.

Consequently, we opted for a specialized branch of machine learning: reinforcement learning. Reinforcement learning circumvents the data bottleneck by creating a virtual environment to simulate physiological organs and the dose distribution resulting from various positions and dwell times of the radioactive source. Here, an agent is trained to perform actions such as catheter movements, deformations, or adjustments of dwell times, iteratively improving its performance through interactions with the environment. A catheter consists of a sequence of multiple source positions that must adhere to geometric constraints.

This approach draws upon statistical principles described in the seminal work of Sutton and Barto [6].

During the training phase, the simulated environment mirrors a diverse patient database, providing a rich substrate for agent learning. In the simulation, the environment dynamically adjusts the parameters of the target tumor volume (TV) and organs at risk (OAR) for each episode. Initially, the study is limited to a 2D model, focusing on catheter movement within one plane while the TV and OAR are situated in another plane. This step validates our approach before transitioning to a 3D model.

The development of the 2D model involves addressing various complexities such as calculation points (TV and OAR) and dwell points, while adhering to dose constraints and tissue protection criteria. Inspired by [7] and, to start, utilizing the DQN reinforcement learning algorithm, the 2D model employs simplified catheters.

Benchmarking using test cases leads to the selection of algorithm as the most efficient agent, offering continuous actions for better control. A simplified spherically symmetrical dose calculation model is adopted for its speed and directness, compared to a slower, more accurate directional model extracted from the TG43 standard [8].

Each catheter, representing a basic unit, consists of a set of dwell points spaced 5 mm apart and movable within the same plane. The learning process involves simplified catheters with up to 3 dwell points, enabling rotation orders and deformation akin to the movement of a snake. The 2D model comprises two parallel planes separated by 5 mm: the first plane contains the OAR and TV, while the second plane contains the dwell points, each color-coded to indicate its affiliation with a catheter.

Transitioning to 3D involves introducing a thick TV and optimizing for multiple catheters while managing inter-catheter distances and adjusting time thresholds. Expansion to 3D requires adjustments to initial conditions and training steps, emphasizing control of curvature radius and additional actions for modifying dwell times.

3 Results

The study developed an environment that allows agents to activate catheters to irradiate the clinical target volume (CTV) while minimizing damage to organs at risk (OAR). The train-

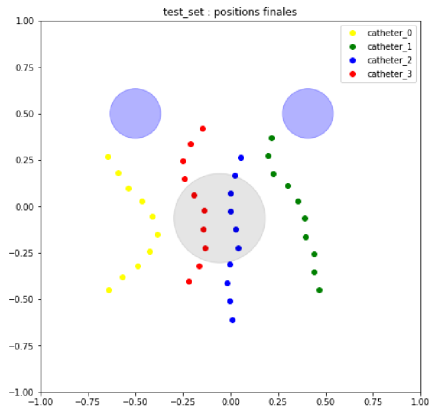


Figure 2: elements of the 2D demonstrator. They are distributed on two parallel planes separated by 5 mm. The first plane contains the OARs (blue) and TVs (gray), and the second plane contains the dwell points, whose color indicates the catheter to which they belong.

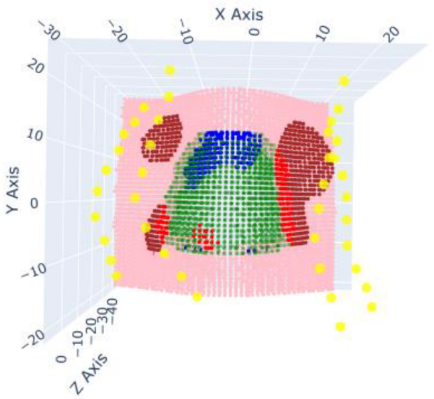


Figure 3: 3D representation of the catheters around the phantom. The catheters are shown in yellow. On the TV: in green, dose between 100 and 140%, in blue less than 100% and in red greater than 140%. On healthy skin: in pink, dose < 140%, in brown > 140%. The skin must be protected.

ing sessions simulated 100,000 patient cases, incorporating constraints to reflect real-world conditions: preventing the catheter from bending below a bend radius of 1.3 cm, imposing a minimum distance of 1 cm between catheters and varying downtime requirements (figure 2).

In the most complete configuration, the training calculation time was several days, the model inference was fast (of the order of a minute).

The performance of the model was evaluated using a 3D digital phantom (figure 3), showing that the model was capable of irradiating the TV between 100% and 140% of the prescribed dose and best preserving the OARs.

The transition to 3D planning has increased complexity and computational burden. However, integrating skin protection into the modele effectively reduced skin exposure to radiation. Further research is needed to optimize the accuracy and complexity of the model.

4 Discussion

We have developed a 3D simulation environment for radiotherapy, enabling an AI agent to manipulate catheters by controlling position, orientation, and dwell time. The agent's objective is to deliver the prescribed dose (between 100% and 140%) to the target volume while minimizing dose exposure to surrounding organs at risk. Realistic constraints have been integrated to enhance clinical validity, including minimum catheter curvature, inter-catheter distance, and dwell time thresholds. Protecting patient data is paramount, governed by stringent regulations such as the Data Protection Act, GDPR, and Public Health Code. Obtaining authorization to use such data involves navigating complex procedures with regulatory bodies. Alternatively, generating virtual data circumvents these challenges, allowing for an expansive dataset with customizable parameters. In our study, each training session simulated 100,000 patient cases, demonstrating the potential for our AI model to generalize across diverse practical scenarios, essential for comparison against professional expertise. The underlying algorithm leverages reinforcement learning, where the AI agent learns optimal dose distribution by receiving rewards for its actions, guided by geometric and dosimetric constraints. Initially applied in a simplified 2D context, the technique was subsequently extended to a more realistic 3D scenario. This proof of concept paves the way for AI to autonomously propose PoS to healthcare professionals, streamlining preparation time. However, it's essential to acknowledge that AI remains a decision-making tool and cannot replace CE-marked software. Final validation involves importing PoS into certified dose calculation software for patient treatment. This Proof of Concept showcases an AI model capable of predicting PoS for a simplified virtual patient case. The model successfully proposes PoS that adhere to anatomical and dosimetric constraints for brachytherapy treatment without relying on patient-specific data. By 2023, 60 patients had undergone personalized brachytherapy treatment at the CFB, with at least one applicator manufactured per week by dedicated healthcare professionals. The use of AI, derived from this proof of concept and generalized to realistic cases, coupled with 3D printing technology, promises a substantial reduction—between 50% and 80%—in staff time dedicated to applicator manufacturing. The utilization of additive manufacturing technology by i3D ensures precise reproduction of AI-proposed dose paths, surpassing the accuracy achievable through manual manufacturing methods. This integration underscores the potential synergy between AI-driven treatment planning and advanced manufacturing techniques, promising improved treatment outcomes and efficiency in clinical practice.

5 Conclusion

The H3DMED project represents a significant advancement in the field of personalized brachytherapy, offering a proof of

concept grounded in innovative technologies tailored to meet the genuine needs of healthcare professionals. Leveraging AI to optimize the design process of 3d print personalized brachytherapy applicators marks a pioneering approach, potentially groundbreaking in its application. Our proof of concept aims to refine the design process by utilizing AI to predict the positioning of radioactive sources in alignment with geometric and dosimetric constraints. While previous literature explored AI's role in determining stopping times for a fixed distribution of Points of Service (PoS), our project expands this scope by introducing spatial degrees of freedom, augmenting the AI model's complexity. Despite the current simplicity of the cases examined, our project illustrates the feasibility of obtaining clinically relevant results in reasonable computational times. Training times, although spanning several days per 100,000 cases, align favorably with clinical applicability, with inference requiring only minutes. Nonetheless, further improvements are imperative to produce a solution suitable for clinical deployment, which requires exploring methods to accelerate training calculations as data volumes increase with case complexity. These achievements surpass the current state of the art in the field, laying the groundwork for future research and development efforts aimed at clinical implementation. However, realizing these prospects will require substantial investment in both technical and regulatory domains.

References

[1] S. Marcié, J. Gerard, C. Dejean, et al. "The inverse square law: A basic principle in brachytherapy". *Cancer Radiother* 26.8 (2022), pp. 1075–1077. DOI: 10.1016/j.canrad.2022.04.002.

[2] M. Lecornu, M. Silva, V. Barraux, et al. "Applicateur numérique par impression tridimensionnelle en curiethérapie de contact [Digital applicator by 3D printing in contact brachytherapy]". *Cancer Radiother* 23.4 (2019), pp. 328–333. DOI: 10.1016/j.canrad.2019.03.008.

[3] J. Guinot, A. Rembielak, J. Perez-Calatayud, et al. "GEC-ESTRO ACROP recommendations in skin brachytherapy". *Radiother Oncol* 126.3 (2018), pp. 377–385. DOI: 10.1016/j.radonc.2018.01.013.

[4] C. Shah, Z. Ouhib, M. Kamrava, et al. "The American Brachytherapy society consensus statement for skin brachytherapy." *Brachytherapy* 19.4 (2020), pp. 415–426. DOI: 10.1016/j.brachy.2020.04.004.

[5] T. Major and C. Polgár. "Treatment planning for multicatheter interstitial brachytherapy of breast cancer – from Paris system to anatomy-based inverse planning". *Journal of Contemporary Brachytherapy* 9.1 (2017), pp. 89–98. DOI: 10.5114/jcb.2017.66111.

[6] R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[7] G. Pu, S. Jiang, Z. Yang, et al. "Deep reinforcement learning for treatment planning in high-dose-rate cervical brachytherapy". *Phys Med* 94 (2022), pp. 1–7. DOI: 10.1016/j.ejmp.2021.12.009.

[8] M. Ferré, H. Mailleux, N. Pierrat, et al. "Calcul de la distribution de dose en curiethérapie". *Cancer/Radiothérapie* 17.2 (2013). Curiethérapie Coordonné par Jean-Michel Hannoun-Lévi, pp. 89–92. DOI: https://doi.org/10.1016/j.canrad.2013.01.005.

Angle Dependence on Positional Radioisotope Source using Monte Carlo Simulation in Brachytherapy

Moo-Jae Han[1], Sang-Won Kang[1], Woong Cho[2], Dong-San Kang[3], and Jin-Beom Chung[1,*]

[1]Department of Radiation Oncology, Seoul National University Bundang Hospital, Seongnam, Korea
[2]Department of Radiation Oncology, Seoul National University Borame Medical Center, Seoul, Korea
[3]Department of Biomedical Engineering and of Biomedicine & Health Sciences, College of Medicine, The Catholic University, Seoul, Korea

Abstract In this study, Monte Carlo Simulation was employed to investigate the angle dependence for tracking the location of a radioisotope (RI). The analysis revealed margin of error values for the top 10% of each slope, with measurements of 5.90% at 0°, 8.08% at 30°, and 20.90% at 60°. The total absorbed dose ratio was observed at 83.77% (30°) and 53.36% (60°) relative to 0° (100%), exhibiting a tendency to decrease with increasing slope. Notably, the maximum value for all gradients was consistently observed at 30° No. 9 pixels, with a corresponding 7.24% decrease at No. 10 pixels. The study highlights a substantial degree of angle dependence, leading to the recommendation that maintaining a proper distance of at least 1 cm between the RI and dosimeters is essential to minimize these effects.

1 Introduction

Brachytherapy is a vital treatment modality involving the introduction of a radioisotope (RI) source into the body to optimize radiation delivery to target lesions [1]. However, the precision of source positioning is crucial to avoid excessive radiation to normal tissues, necessitating rigorous evaluation of dwell position and dwell time as part of quality assurance (QA) measures [2,3]. Traditional methods relying on visual inspection, such as ruler and film analysis using a camera, lack quantitative error data and are prone to inaccuracies [4,5].

This study investigates the potential of digital dosimeters, focusing on reproducibility, linearity, distance dependence, and, notably, angle dependence [6]. Angle dependence is known for its sensitivity to variations in RI position, necessitating an accurate measurement setup despite the inherent challenges in achieving precision.

To address this, Monte Carlo simulation was employed in the analysis, utilizing the FLUKA (FLUktuierende KAskade) tool developed by CERN. The dosimeter material selected for the study is Mercury (Ⅱ) Iodide (HgI_2), chosen for its relatively high signal generation efficiency. The research aims to provide insights into the trend of angle dependence for unit-cell dosimeters in the context of brachytherapy QA, shedding light on the potential advantages of digital dosimeters and advancing the understanding of measurement accuracy in this critical medical application.

2 Materials and Methods

Radiation Monte Carlo simulation codes, such as FLUKA, involve a series of crucial steps, including geometry modeling, source term definition, scoring, and data analysis. This study focuses on the foundational step of geometry modeling to realistically represent radiation exposure and the irradiation environment. Using the FLUKA code, we executed the geometry modeling of the unit-cell dosimeter, as illustrated in Figure 1, to assess angle dependence.

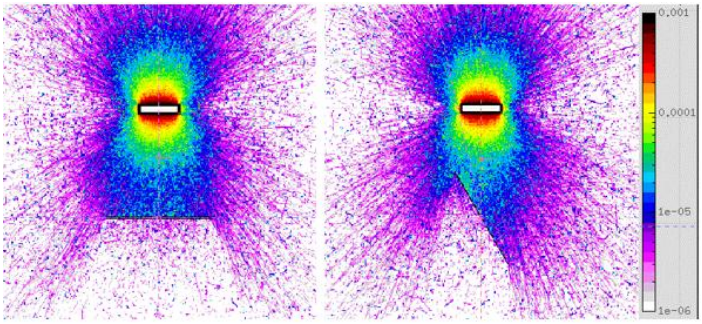


Figure 1: Geometry and dose distribution of 0° and 60° angle

The unit-cell dosimeter, with dimensions of 1 cm × 1 cm, was subdivided into 100 pixels of size 0.1 cm × 0.1 cm for comprehensive evaluation of dose distribution. Meeting the minimum thickness requirement for measuring energy above 6 MV, the dosimeter had a thickness of 200 μm.

In terms of RI sources, the unit-cell dosimeters were strategically positioned 1 cm apart from the center of the designed unit-cell dosimeter, with measurement angles set at 0°, 30°, and 60°. Source term information for the Ir-192 source, as implemented in the FLUKA code, is presented in Table 1.

RI Information	Z (Atomic Number)	77
	A (Mass Number)	192
	Energy	Approximately 300 KeV
RI Size	Type	Cylinder
	Diameter	0.65 mm
	Length	3.60 mm
RI Shield	Material	AISI316L steel
	Diameter	0.90 mm
	Length	4.50 mm

Table 1. RI Source term

This research contributes to advancing the understanding of radiation exposure analysis through precise geometry modeling and simulation methodologies, providing valuable insights for applications such as brachytherapy quality assurance.

In the context of radiation exposure analysis using Monte Carlo simulation, the scoring methodology is a critical aspect. In this study, a unit-cell dosimeter was utilized in conjunction with a USRBIN card for scoring purposes. To comprehensively capture the absorbed dose distribution, the scoring type was configured as dose (GeV/mass), employing a 10 × 10 binning strategy designed to yield 100 absorbed doses in 1 mm × 1 mm size.

For the subsequent distributional analysis, each slope result was normalized to the maximum value, providing a comparative framework for understanding dose distribution patterns. Additionally, an error rate is presented for the top 10% of the distribution, highlighting the magnitude change in the region of high dose distribution. To facilitate a comprehensive view of the total absorbed dose values for all angle, normalization was performed relative to the total absorbed dose at 0°.

Numerical comparisons are provided for each slope across various angles, considering the x-axis located at the center of the unit-cell dosimeter. The comparisons are further normalized to the No. 9 pixel value at 30°, representing the highest absorbed dose value. This methodology enhances our understanding of absorbed dose distribution trends and provides a quantitative framework for evaluating the impact of different measurement angles in radiation exposure simulations.

3 Results

This study delves into the examination of the angle dependence trend by analyzing dose distribution plots at tilts of 0°, 30°, and 60° relative to the central axis of the unit-cell dosimeter, utilizing a fixed Ir-192 source. Figure 2 provides an overview of the distribution plots corresponding to each measurement angle.

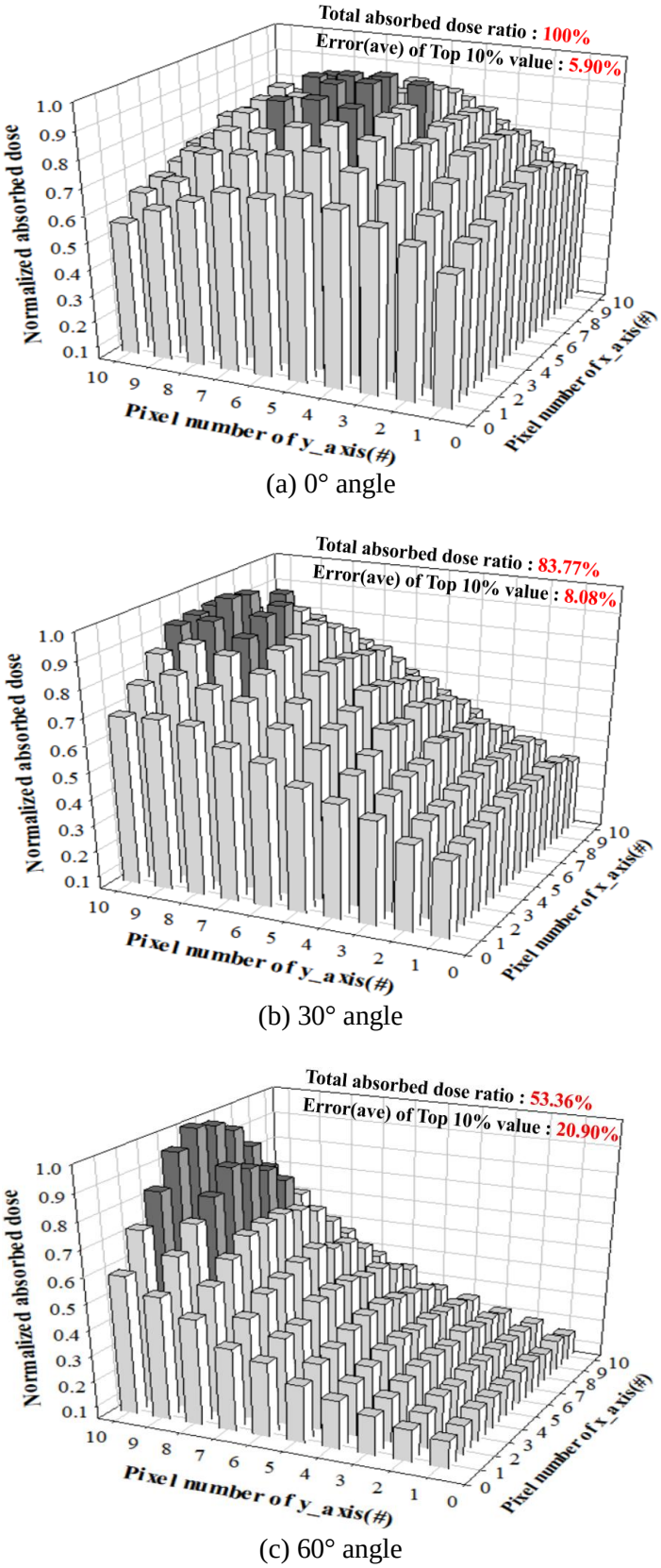


Figure 2: Absorbed dose distribution according to each angle.

Within Figure 2, (a) corresponds to 0°, (b) to 30°, and (c) to 60° measurement angles. The observed error rates for the top 10% of readings were 5.90% at 0°, 8.08% at 30°, and 20.90% at 60°, indicating a substantial increase in distribution towards regions proximal to the RI. The

percentage of the total absorbed dose exhibited a trend of 83.77% at 30° and 53.36% at 60° relative to 0° (considered 100%), revealing a diminishing absorbed dose trend with increasing angles of the unit-cell dosimeter.

For a quantitative comparison at each angle, the study focused on pixel values along the y-axis, specifically the center point (No.5) of the x-axis. Figure 3 depicts normalized values for absorbed dose at 0°, 30°, and 60°. At 0°, the maximum dose value was localized at No. 5, representing the center point. At a 30° angle, absorbed dose displayed a linear increase, while at 60°, it exhibited a rapid quadratic rise.

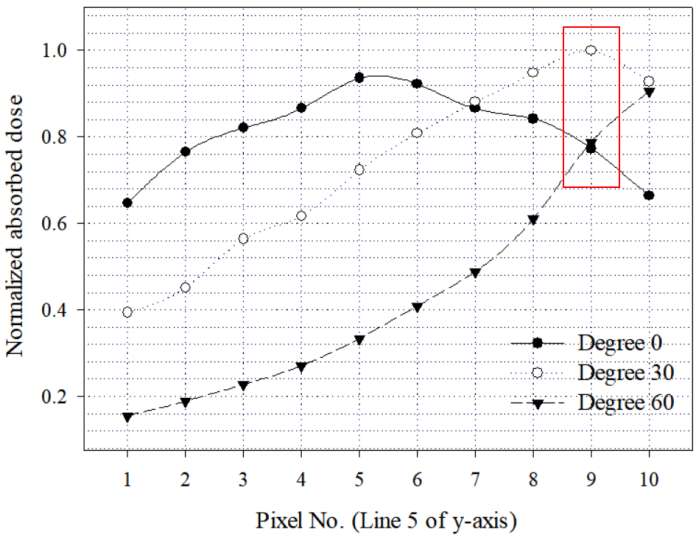


Figure 3: Absorbed dose about line 5 of y-axis.

The peak absorbed dose occurred at No. 9 pixel at 30°, with a subsequent 7.24% decrease at No. 10 pixel. This suggests that the maximum measurement point is at the No. 9 pixel at 30°. Remarkably, the No. 9 pixel at 0° and the No. 9 pixel at 60°, despite being at different distances from the source, demonstrated similar absorbed doses, indicating comparable measurement outcomes. The findings contribute valuable insights into the nuanced angle dependence in brachytherapy dosimetry.

4 Discussion

This study systematically assessed the inclination of angle dependence in brachytherapy QA by employing Monte Carlo simulation techniques facilitated through the FLUKA code. The results illuminate a notable trend: as the dosimeter measurement angle increases, the distance from the RI becomes more substantial in the nearby region. However, as illustrated in Figure 1, the dose is effectively transmitted in the immediate area, leading to an overall lower absorbed dose.

In Figure 2, a discernible reduction in the top 10% area is observed. The diminishing number of pixels in the top 10% of the dose distribution as the angle increases suggests a significant variation in absorbed dose between pixels, potentially influenced by disparities in the size and shape of the irradiation surface.

Based on these findings, the study recommends maintaining the unit-cell dosimeter at an appropriate distance, aligning with the reference distance of 1 cm for High Dose Rate (HDR) brachytherapy treatment planning systems, to mitigate angle dependence [7]. Further exploration into thickness variations to enhance the absorbed dose of the dosimeter is suggested, offering potential reinforcement for the current research. It is emphasized that the dosimeter's construction material plays a crucial role, influencing the transmitted dose based on its physical properties and thickness. These insights contribute to a more nuanced understanding of angle dependence considerations in brachytherapy QA.

5 Conclusion

This study unveils that variations in dosimeter absorbed dose stem from disparities in the size and distance of the irradiation field, influenced by the angle of RI measurement. Depending on the dosimeter's shape, these variations can yield diverse measurement outcomes owing to absorbed dose discrepancies at distances less than 1 cm. Consequently, the angle dependence of the dosimeter becomes a crucial consideration for HDR brachytherapy QA, particularly for materials with high low-intensity and optimal dosimeter thickness. This research contributes to a deeper understanding of the nuanced factors affecting dosimeter measurements in HDR brachytherapy, emphasizing the significance of accounting for angle dependence in QA protocols.

Acknowledgment

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (NRF-2022R1F1A107219912 and NRF-RS-2023-00212193).

References

[1] Nag, Subir. "High dose rate brachytherapy: its clinical applications and treatment guidelines". Technology in cancer research & treatment 3.3 (2004), pp. 269-287. DOI: https://doi.org/10.1177/153303460400300305.

[2] Lee, KyungJa, SoHyun An, and Rena Lee. "Monte carlo evaluation for the effect of positional inaccuracy of source on patient's dose in high-dose-rate brachytherapy". Journal of the Korean Physical Society 55.6 (2009), pp. 2361-2365. DOI: https://doi.org/10.3938/jkps.55.2361.

[3] K. Yogo, A. Matsushita, Y. Tatsuno, et al. "Imaging Cherenkov emission for quality assurance of high-dose-rate brachytherapy". Scientific reports 10.1 (2020), P. 3572. DOI: https://doi.org/10.1038/s41598-020-60519-z.

[4] Nath. R., Anderson. L. L., Meli. J. A, et al. "Code of practice for brachytherapy physics: report of the AAPM Radiation Therapy

Committee Task Group No. 56." Medical physics 24.10 (1997), pp. 1557-1598. DOI: https://doi.org/10.1118/1.597966.

[5] J. Venselaar and J. Perez-Calatayud. “Apractical guide to quality control of brachytherapy equipment. (2004).

[6] Heo. Y. J., Kim. K. T., Han. M. J, et al. "Development of a stable and sensitive semiconductor detector by using a mixture of lead (II) iodide and lead monoxide for NDT radiation dose detection." Journal of Instrumentation 13.03 (2018), p. C03023. DOI: https://doi.org/10.1088/1748-0221/13/03/C03023.

[7] Sawicki. M. "Treatment Planning in Brachytherapy HDR Based on Three- Dimensional Image." Computed Tomography Advanced Applications. (2017).

State-of-the-art of CT enhancement methods for RT planning: a review

Özgür Özer[1,2,3], Alexandre Huat[2], David Gibon[2], and Juliette Thariat[1,3]

[1]GORTEC, Tours, France
[2]AQUILAB by Coexya, Lille, France
[3]Laboratoire de Physique Corpusculaire (UMR 6534), University of Caen Normandie, Caen, France

Abstract Computerized tomography (CT) scan uses X-rays and is the primary material for defining ROIs delineated by radiation oncologists for radiotherapy (RT) planning and radiotherapy quality assurance (RTQA). Additionally, doses are calculated from the CT electronic densities of tissues, but CT quality varies. CTs often exhibit blur, noise, or artefacts stemming from metals, which can be remedied by super resolution (SR) and metal artifact reduction (MAR) models, respectively. A literature search on AI models that can exploit the latest computer vision algorithms and combine both tasks is presented. An interesting approach is based on a content consistent super-resolution (CCSR) model, which utilises Stable Diffusion. The two tasks can be conducted by this model by simulating both SR and MAR distortions in the training data.

1. Introduction

Computerized tomography (CT) scan uses X-rays and stands as one of the preeminent medical imaging techniques for screening and diagnosis in many diseases including cancer, as well as radiotherapy (RT) planning and radiotherapy quality assurance (RTQA). Low quality CT scans may obscure the borders of structures considered as regions of interest (ROI, such as organs at risk (OAR) and tumour), hindering the delineation of ROIs. CTs with high resolution and high signal-to-noise ratio hold the potential to augment the precision of radiological features. Imaging quality contributes to more accurate definition of ROI and their delineation by radiation oncologists in their part of the RT planning. Super resolution (SR) reconstruction from deep learning approaches is an auspicious method of SR to increase the visual quality of images.

Metals in the area imaged by the CT scanner can degrade the quality by introducing artefacts. Linear inverse problem algorithms, such as filtered back projection, induce artefacts in CT images [1]. Deep learning based methods have achieved respectable performances for metal artefact reduction (MAR) tasks. These are based on sinogram domain enhancement, image domain enhancement, or dual domain (joint sinogram and CT image) enhancement.

SR and MAR are addressed by different models. A model that removes noise, blur or metal artefacts, or improves the visuals based on the deficiency of the CT has not been studied. Such a model may improve the quality of ROIs so that radiation oncologists can better delineate them during RT planning, and better assess delineations during RTQA processes as part of clinical research.

2. Materials and Methods

We reviewed the literature on the academic research website Scinaps [2] using the search keywords *CT super resolution* and *metal artifact reduction CT*. Only the articles that used AI on CT scans that surpassed previous models across the scores they presented were selected. Also, the latest algorithms on computer vision were investigated on the following websites: Stability AI Research Blog [3], Metal AI Research on computer vision [4], Microsoft research on computer vision [5], NVIDIA publications [6], Google Publications on machine intelligence [7], and Papers With Code on super resolution [8] and metal artefact reduction [9]. In all cases, the results earlier than 2023 were filtered. The most suitable models elected based on their potential to be adapted to the geometry and the unit of CT scans (3D and HU).

3. Results

Among AI methods, KerSRGAN [10] introduced a feature extractor in Generative Adversarial Networks (GAN) [11] in an unsupervised manner. Residual feature attentional fusion network [12] incorporated a feature extraction block to elevate visual acuity. Multi-scale Attention Mechanism [13] uses convolutional layers with various kernel sizes to capture information with different sizes. These papers used bicubic resizing to reduce the resolution rather than simulating long slice thickness. A more recent method named Linking In-Plane and Through-Plane Transformers [14] simulates longer slice thickness to reduce the quality.

For MAR, most methods published in 2023 were deep learning based. Standard vision transformer was tested for MAR [15] in image domain. A generative AI was attempted using Res-U-Net GAN [16] in the image

domain [17]. Polyner [18] took an unsupervised approach using implicit neural representation [19] on image domain. Plugging Diffusion Priors in dual domains [20] used patch diffusion [21] for unsupervised training.

Outside medical AI, there have been major developments in computer vision and SR. Stable Diffusion [22] generates realistic images based on text. It incrementally refines noise from the vector representation from which the images are decoded. Its derivatives prevailed in many applications such as segmentation [23], anomaly detection [24] etc. Other models based on Stable Diffusion, such as Content Consistent Super-Resolution (CCSR) [25], Pixel-Aware Stable Diffusion (PASD) [26] outperformed other metods for SR [25]. PASD surpasses CCSR on image quality that aligns better with human perception [25]. Still, it could be unsuitable for RT. Generative AI models give a different output at every run. PASD score varied more than CCSR between runs for the same image. Frechet inception distance (FID), the score inversely related to the fidelity of generated output to real data, was high [25] for PASD i.e. it created non-existent details to have higher quality. On the other hand, CCSR made use of ControlNet [27], to confine the outputs to a narrower distribution for score consistency and lower FID score [25] compared to PASD on RealSR [28]: 105.55 vs 123.08 [25].

4. Discussion

A network that can manage MAR and SR simultaneously should be complex enough to model both but reasonably fast for training and inference. PASD was did not only fail to give consistent results and low FID score but also its very complex network may have a very costly training and inference load. In addition, for medical context it would be necessary to replace CLIP [29] with medCLIP [30], CLIP for medical context. Nevertheless, medCLIP is not mature enough for practical use because it has not been trained with enough data [30]. A modified version of CCSR for SR and MAR on CTs could overperform previous models in modelling both tasks with high fidelity to actual radiological features and consistency in quality throughout executions. For both SR and MAR, training data creation start with simulation of noise, blur, or metal artefact. By creating instances of CT scans that can have any combination of these in the training data, a model may gain the ability to conduct SR or/and MAR depending on the issue. Nonetheless, in real practice, there may be distortions that were not taken into account during training data creation.

5. Conclusion

One of the latest algorithms in computer vision, Stable Diffusion, has not been implemented in RT planning. Adapting Stable Diffusion-based CCSR model to 3D CT data and training with SR and MAR distortions may conceive a model that implements both simultaneously and consistently contingent upon CT deficiency with superior performance. A satisfactory model would improve the visual distinction between tissues and remove the metal artefacts before a radiation oncologist works defining ROIs during RT or RTQA in clinical trials.

Bibliography

[1] A. C. Kak and M. Slaney, *Principles of Computerized Tomographic Imaging*. Society for Industrial, Applied Mathematics, 2001. doi: 10.1137/1.9780898719277.

[2] “Scinapse.” [Online]. Available: https://www.scinapse.io/

[3] “Research Blog.” Accessed: Feb. 02, 2024. [Online]. Available: https://stability.ai/research

[4] “Meta AI Publications.” Accessed: Feb. 02, 2024. [Online]. Available: https://ai.meta.com/results/?page=1&content_types%5B0%5D=publication

[5] “Computer Vision.” Accessed: Feb. 02, 2024. [Online]. Available: https://www.microsoft.com/en-us/research/research-area/computer-vision/

[6] “Publications | Research.” Accessed: Feb. 02, 2024. [Online]. Available: https://research.nvidia.com/publications

[7] “Publications – Google Research.” Accessed: Feb. 02, 2024. [Online]. Available: http://research.google/pubs/

[8] “Papers with Code - Super-Resolution.” Accessed: Feb. 02, 2024. [Online]. Available: https://paperswithcode.com/task/super-resolution

[9] “Papers with Code - Metal Artifact Reduction.” Accessed: Feb. 02, 2024. [Online]. Available: https://paperswithcode.com/task/metal-artifact-reduction

[10] Y. Li, L. Chen, B. Li, and H. Zhao, “4× Super-resolution of unsupervised CT images based on GAN,” *IET Image Processing*, vol. 17, no. 8, pp. 2362–2374, 2023, doi: 10.1049/ipr2.12797.

[11] R. Gao, B. Lou, Z. Xu, D. Comaniciu, and A. Kamen, “Flexible-Cm GAN: Towards Precise 3D Dose Prediction in Radiotherapy,” in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 715–725. doi: 10.1109/CVPR52729.2023.00076.

[12] K. Yang *et al.*, “Residual Feature Attentional Fusion Network for Lightweight Chest CT Image Super-Resolution,” *Computers, Materials & Continua*, vol. 75, no. 3, pp. 5159–5176, 2023, doi: 10.32604/cmc.2023.036401.

[13] J. Yin, S.-H. Xu, Y.-B. Du, and R.-S. Jia, "Super resolution reconstruction of CT images based on multi-scale attention mechanism," *Multimedia Tools and Applications*, vol. 82, no. 15, pp. 22651–22667, 2023, doi: 10.1007/s11042-023-14436-8.

[14] Z. Chen, C. Niu, Q. Gao, G. Wang, and H. Shan, "LIT-Former: Linking In-plane and Through-plane Transformers for Simultaneous CT Image Denoising and Deblurring," *IEEE Transactions on Medical Imaging*, p. 1–2, 2024, doi: 10.1109/TMI.2024.3351723.

[15] S. Zheng *et al.*, "Vision Transformer with Progressive Tokenization for CT Metal Artifact Reduction," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5. doi: 10.1109/ICASSP49357.2023.10095579.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition." 2015.

[17] Z. Zhang, M. Yang, L. Xu, J. Yang, H. Guo, and J. Wang, "An Innovative Metal Artifact Reduction Algorithm based on Res-U-Net GANs," *Current Medical Imaging*, vol. 19, no. 13, pp. 1549–1560, 2023, doi: 10.2174/1573405619666230217102534.

[18] Q. Wu *et al.*, "Unsupervised Polychromatic Neural Representation for CT Metal Artifact Reduction." 2023.

[19] Y. Xie *et al.*, "Neural Fields in Visual Computing and Beyond." 2022.

[20] X. Liu, Y. Xie, S. Diao, S. Tan, and X. Liang, "Unsupervised CT Metal Artifact Reduction by Plugging Diffusion Priors in Dual Domains." 2024.

[21] T. Luhman and E. Luhman, "Improving Diffusion Model Efficiency Through Patching." 2022.

[22] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models." 2022.

[23] J. Tian, L. Aggarwal, A. Colaco, Z. Kira, and M. Gonzalez-Franco, "Diffuse, Attend, and Segment: Unsupervised Zero-Shot Segmentation using Stable Diffusion." 2023.

[24] H. He *et al.*, "DiAD: A Diffusion-based Framework for Multi-class Anomaly Detection." 2023.

[25] L. Sun, R. Wu, Z. Zhang, H. Yong, and L. Zhang, "Improving the Stability of Diffusion Models for Content Consistent Super-Resolution." 2023.

[26] T. Yang, P. Ren, X. Xie, and L. Zhang, "Pixel-Aware Stable Diffusion for Realistic Image Super-resolution and Personalized Stylization." 2023.

[27] L. Zhang, A. Rao, and M. Agrawala, "Adding Conditional Control to Text-to-Image Diffusion Models." 2023.

[28] J. Cai, H. Zeng, H. Yong, Z. Cao, and L. Zhang, "Toward Real-World Single Image Super-Resolution: A New Benchmark and A New Model." 2019.

[29] A. Radford *et al.*, "Learning Transferable Visual Models From Natural Language Supervision." 2021.

[30] Z. Wang, Z. Wu, D. Agarwal, and J. Sun, "MedCLIP: Contrastive Learning from Unpaired Medical Images and Text." 2022.

A Tool for Individualised Dose Mapping Uncertainty Quantification in the Reirradiation Setting: Head and Neck Cancer

Chelmis M. Thiong'o[1], Ane Appelt[3], Marcel van Herk[1,2], Matthew Lowe[1,4], David Thomson[1,5], and Eliana Vasquez Osorio[1,2]

[1]Division of Cancer Sciences, The University of Manchester, Manchester, United Kingdom
[2]Radiotherapy Related Research, The Christie NHS Foundation Trust, Manchester, United Kingdom
[3]Leeds Institute of Medical Research at St. James's, University of Leeds, Leeds, United Kingdom
[4]Christie Medical Physics and Engineering, The Christie NHS Foundation Trust, Manchester, United Kingdom
[5]Clinical Oncology, The Christie NHS Foundation Trust, Manchester, United Kingdom

Abstract Deformable image registration (DIR) is an important tool to enable the assessment of prior radiation doses in the reirradiation (reRT) setting. However, the effect of geometric uncertainties on dose mapping from previous treatments onto the reRT planning scan remains a key challenge, which limits its clinical application. We propose a novel tool for individualised dose mapping uncertainty quantification in the reirradiation setting for head and neck cancer. The tool includes a Graphical User Interface (GUI) and creates a report on the uncertainties using a commercially available treatment planning system, RayStation (Version 11B-R, RaySearch Lab, Stockholm, Sweden). We demonstrate the use of this tool in a dataset of eight patients, evaluating both a serial and a parallel organ, i.e., brainstem and parotid glands. 87.6% of 1524 registrations for all eight patients' organs at risk were deemed plausible (met AAPM TG 132 threshold of 0.3cm) and used in dose mapping. However, this was very patient and organ-specific. For example, 100% of registrations for the brainstem were plausible, while 91.7% were plausible for each of the parotids. Overall, this tool will allow for the assessment of mapped radiation doses following repeat registrations with differing parameters.

1 Introduction

Deformable image registration (DIR) in the reirradiation setting is used to enable the assessment of prior radiation doses. The use of DIR in the clinic setting is, however, challenged by the lack of assessment of the effect of geometric uncertainties on dose mapping from previous treatments [1]. This reduces the confidence in the mapped doses by the treating physicians, often leading to more conservative treatment prescriptions which may affect treatment outcomes.

We propose a novel tool for individualised dose mapping uncertainty quantification in the reirradiation setting. The tool includes a Graphical User Interface (GUI) and creates a report on the dosimetric impact of geometric registration uncertainties using a commercially available treatment planning system, RayStation (Version 11B-R, RaySearch Lab, Stockholm, Sweden). We demonstrated the use of this tool in a set of patients who underwent head and neck reirradiation.

2 Materials and Methods

The GUI is shown in Figure 1. A user selects the initial and reirradiation CT scans based on the examination dates. The GUI then gives the user an option to select relevant organs at risk (OARs) from drop-down menus populated from all structures under the same name in both CT scans. These are populated by pressing the button, "Find Common Structures". The threshold value to determine locally plausible registration (based on the mean distance to agreement between mapped and delineated structures) is then set in the text box, with a default of 0.3cm as suggested by AAPM TG-132 [2]. The GUI gives the user an option to add relevant OARs using the "Add row" button. On each row, an OAR is selected. This OAR can be included as a controlling structure in the registration if the checkbox is selected. The report can then be customised to include: 1) DVH estimate uncertainty, 2) DVH uncertainty, plotting DVH curves of all plausible registrations, and 3) voxel-wise spatial uncertainty, computing the standard deviation (SD) of the mapped doses in all voxels within the OAR.

The user then clicks the "Run Registration" button. This allows the script to run six registrations as defined in the methods. Once this is complete, the user clicks on "Run Evaluation". This performs the geometric and dosimetric evaluation of the OARs added to the table in the GUI and generates a report. The report shows a summary of the patient's details (such as the patient ID, disease sites, initial planning CT date and reirradiation CT date), results from the geometric and dosimetric assessment carried out on the selected OARs in the form of tables, and the DVH plots of the mapped dose distribution which were selected to be included in the report. Other details such as the plausible registration threshold, the type of registration performed, and the date and time the report was generated are also included.

Data from eight patients with head and neck cancer who were reirradiated (six Type I and two Type II according to ESTRO-EORTC consensus on reirradiation [3]) were used. Both initial and reRT planning CTs were automatically contoured using Limbus Contour (Limbus.AI v 1.7.0-B3) to ensure consistency. Planning CTs, dose distributions and

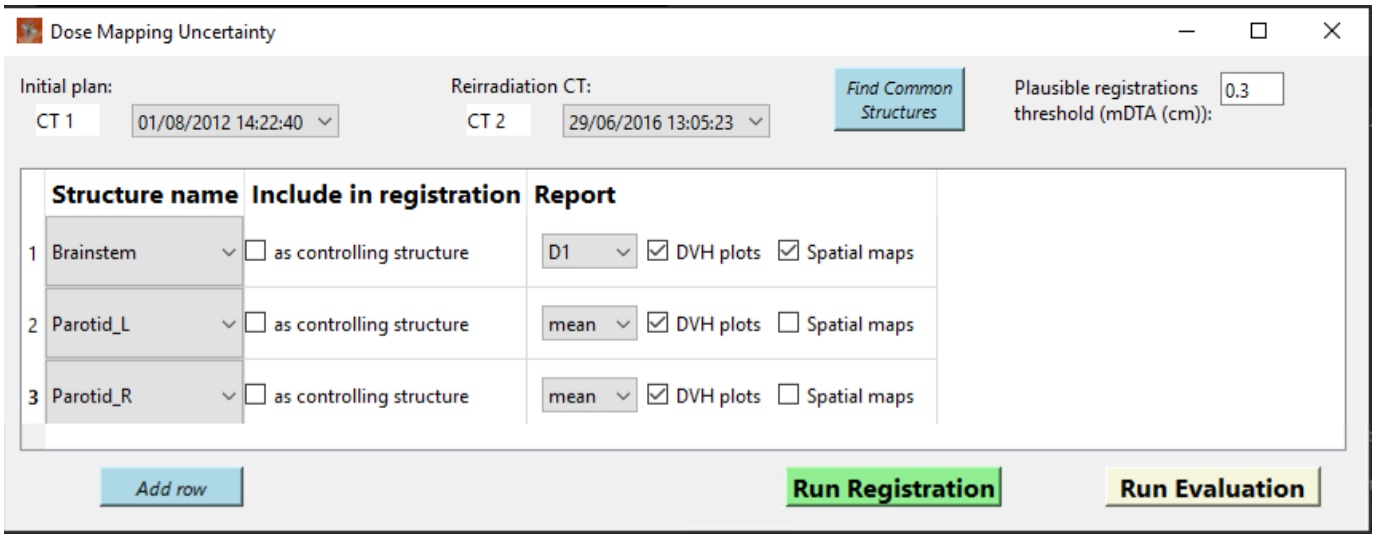


Figure 1: A screenshot of the GUI showing the interface the user interacts with. A Python script runs in the background to perform DIR, dose mapping and an assessment of the registrations performed and the resulting mapped doses.

contours for the initial and reRT course were then imported into RayStation.

To demonstrate the use of this tool, we used it to map the dose of a previous radiotherapy course into a reirradiation planning CT for the eight patients, exploring the impact of geometric registration uncertainties on the mean dose of the parotid glands and the near maximum (D1%) for the brainstem. We also recorded the time it took to run the entire pipeline on a system with NVidia Quadro RTX 6000 GPU, 6.4GHz Intel(R) Xeon(R) CPU and 64GB RAM.

Patient	Brainstem	Left Parotid	Right Parotid
Patient 1	6	6	6
Patient 2	6	6	6
Patient 3	6	6	6
Patient 4	6	6	6
Patient 5	6	6	6
Patient 6	6	2	2
Patient 7	6	6	6
Patient 8	6	6	6

Table 1: Number of registrations passing for the brainstem, left and right parotid glands. For each organ, a total of six registrations were carried out. The number of registrations passing with mDTA of $\leq$ 0.3cm are shown.

3 Results

Figure 2 on Page 4 shows the report generated for patient 6 using the configuration presented in Figure 1. Table 1 shows the number of registrations that met the AAPM TG 132 suggested threshold of 0.3cm. Table 2 shows the dose mapping uncertainty results for the 8 patients in the cohort for the brainstem and parotid glands. On average, for all patients, the script runs in 11.9$\pm$4.5 minutes.

For our selected use case, an example of a sample report for Patient 6 is presented in Figure 2. The report includes patient details, registration uncertainty details and dose mapping uncertainty details for the brainstem and the parotid glands. The mapped doses for all three OARs varied greatly from the initial treatment plan's dose to these organs. The DVH maps also showed the variation in the mapped doses for each of the plausible registrations.

4 Discussion/Conclusion

We demonstrated a novel tool for individualised dose mapping uncertainty quantification in the reirradiation setting for head and neck cancer. The tool is developed on RayStation's scripting interface and consists of a GUI and generates a report (see Figure 2). The tool performs six DIRs followed by the quantification of the mapped dose uncertainty through the assessment of the registrations performance and the impact these had on the resulting mapped dose for OARs selected by the user in the GUI.

Reviews by Murr et al 2023 and Nenoff et al 2023 [1, 4] recommend researchers repeat registrations with differing parameters for the assessment of uncertainties. However, this process can be time-consuming and may introduce errors when carried out manually. Here, we ran six registrations for each patient.

Our tool quantifies the impact of plausible registration uncertainties on mapped dose distributions in an automatic and consistent way that can be extended to patients in the clinic.

Patient	ReRT type	Time (min)	mean mDTA of passing registrations ± standard deviation			mean mapped dose ± standard deviation		
			Brainstem	L. Parotid	R. Parotid	Brainstem	L. Parotid	R. Parotid
Patient 1	I	22	0.06 ± 0.03	0.08± 0.03	0.16 ± 0.01	34.31 ± 0.82	8.64 ± 0.07	45.90 ± 1.32
Patient 2	I	12	0.05 ± 0.01	0.13 ± 0.02	0.16 ± 0.03	44.31 ± 0.92	40.78 ± 0.90	26.26 ± 0.90
Patient 3	I	9	0.15 ± 0.01	0.13 ± 0.02	0.25 ± 0.02	13.38 ± 0.59	3.87 ± 0.06	23.89 ± 0.06
Patient 4	II	9	0.09 ± 0.02	0.13 ± 0.01	0.10 ± 0.01	24.61 ± 0.14	5.35 ± 0.05	18.42 ± 0.30
Patient 5	I	9	0.13 ± 0.02	0.07 ± 0.01	0.28 ± 0.01	24.40 ± 0.34	6.61 ± 0.39	44.40 ± 0.41
Patient 6	I	14	0.11 ± 0.01	0.39 ± 0.08	0.36 ± 0.09	41.77 ± 0.58	41.18 ± 1.61	41.43 ± 2.06
Patient 7	I	9	0.11 ± 0.02	0.12 ± 0.08	0.12 ± 0.08	0.52 ± 0.01	0.75 ± 0.02	0.89 ± 0.03
Patient 8	II	11	0.12 ± 0.02	0.12 ± 0.03	0.21 ± 0.03	3.21 ± 0.13	31.87 ± 0.93	5.61 ± 0.22

Table 2: A summary of dose mapping uncertainty for all patients in our cohort.
ReRT - reirradiation; min - minutes; mDTA - mean distance to agreement

It can be argued that running only six registrations may not generate enough spread to reveal trends, therefore we are currently working on creating registration templates to allow users to run more registrations for a more granular analysis (for example, cases with extreme dose gradients in the region of interest). Further work is also being carried out to incorporate the assessment of cumulative doses with a consideration of biological factors and perform external validation of this tool.

In conclusion, the development of this tool would allow us to systematically and consistently assess dose mapping uncertainty for patients at the reirradiation planning stage.

References

[1] M. Murr, K. K. Brock, M. Fusella, et al. "Applicability and usage of dose mapping/accumulation in radiotherapy". *Radiotherapy and Oncology* 182 (May 2023), p. 109527. DOI: 10.1016/J.RADONC.2023.109527.

[2] K. K. Brock, S. Mutic, T. R. McNutt, et al. "Use of image registration and fusion algorithms and techniques in radiotherapy: Report of the AAPM Radiation Therapy Committee Task Group No. 132: Report". *Medical Physics* 44.7 (2017). DOI: 10.1002/mp.12256.

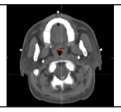
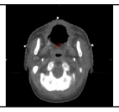
[3] N. Andratschke, J. Willmann, A. L. Appelt, et al. "European Society for Radiotherapy and Oncology and European Organisation for Research and Treatment of Cancer consensus on re-irradiation: definition, reporting, and clinical decision making". *The Lancet Oncology* 23.10 (2022), e469–e478. DOI: https://doi.org/10.1016/S1470-2045(22)00447-8.

[4] L. Nenoff, F. Amstutz, M. Murr, et al. "Review and recommendations on deformable image registration uncertainties for radiotherapy applications". *Physics in Medicine & Biology* 68.24 (Dec. 2023), 24TR01. DOI: 10.1088/1361-6560/AD0D8A.

Dose mapping uncertainty report for re-irradiation

Patient ID: Patient 6 DOB: 15/11/1941

Patient details

	First treatment CT	Second treatment CT
		
Plan	70 / 33 VMAT (NASOPHARYNGEAL)	N/A
Disease site	Nasopharynx	Oral Cavity
Plan date	01/08/2012	N/A
Image acquisition date	01/08/2012	29/06/2016

Registration uncertainty details

6 intensity-based registrations performed on 32 OARs; Plausible registrations threshold: mDTA ≤ 0.3cm.

Registration performance details

mDTA (mean DTA) metrics in cm, reported as mean +/- stdev.

OAR	Number passing thresholds	All	Passing	Not passing
		mDTA	mDTA	mDTA
Brainstem	6/6	0.11 ± 0.01 cm	0.08 ± 0.03 cm	NA
Left Parotid	2/6	0.39 ± 0.08 cm	0.29 ± 0.003 cm	0.44 ± 0.02 cm
Right Parotid	2/6	0.36 ± 0.09 cm	0.25 ± 0.03 cm	0.42 ± 0.04 cm

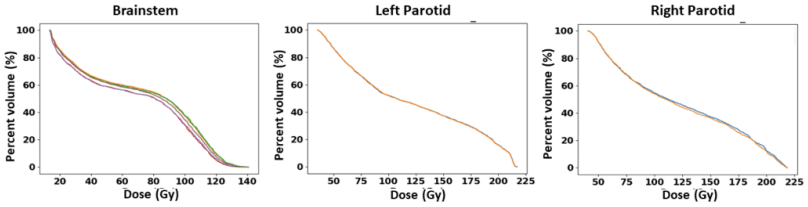
NA: not applicable: for < 2 values for the Stdev or < 1 value for the mean.

Dose mapping uncertainty details

DVH metric uncertainty

OAR	**DVHmetric**	**Value in original planning CT**	**Mean Mapped value (+- sd)**
Brainstem	Dmax	42.38 Gy	41.77 ± 0.58 Gy
Left Parotid	Dmean	35.88 Gy	41.18 ± 1.61 Gy
Right Parotid	Dmean	38.34 Gy	41.43 ± 2.06 Gy

DVH uncertainty maps



Spatial uncertainty map, Brainstem

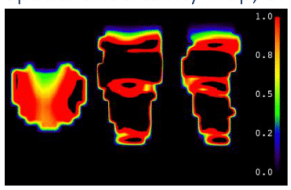


Figure 2: A copy of the generated report for Patient 6 adapted for this abstract. The report follows Figure 1 and is generated when the script completes running. It includes a summary of the patient's details, and results from the geometric evaluation as well as dosimetric assessment, based on the options selected by the user from the GUI. All the registrations for the brainstem were plausible, but only two registrations for both parotids were plausible. The script took 14 minutes to run for this patient, including interacting with the user interface.

Robustness of Radiomics features to reconstruction kernel batch effect: How bad is the situation?

Elfried Salanon[1], Aditya Apte,PhD[1], Anqi Fu,PhD[1], Usman Mahmoud,PhD[1], Zehor Belkhatir,PhD[1],Amita Shukla-Dave,PhD[1], Joseph O.Deasy,PhD[1]

[1]Department of Medical Physics, Memorial Sloan Kettering, New York, NY, USA

Abstract

The aim of this study was to evaluate the impact of kernel batch effect on radiomic features extracted from CT scans. Radiomic features extracted from CT scans of a 3D printed phantom using CERR, comprised 282 features. Batch effects were assessed via principal component analysis (PCA) and correlation measures. Robustness was measured using the concordance correlation coefficient (CCC) and Pearson correlation coefficient, with a CCC threshold of ≥ 0.9. Statistical analysis was performed using R software. PCA identified two clusters: Standard, ASIRs, ASIRV, and Soft kernels in one, and Lung and Bone Kernels in the other. A gradient from ASIR10 to ASIR50 and ASIRV1 to ASIRV5 was observed in the distance to the Standard Kernel. Correlation matrices revealed little change in ASIRs, ASIRVs, and the Standard Kernel but significant changes in Bone and Lung Kernels. Combat-based correction improved robustness, particularly in first-order statistics, mitigating batch effects in ASIRs and the Standard Kernel. 40 robust features were identified. However, challenges persist in harmonizing Bone and Lung kernels. The robustness of central tendance parameters in contrast to the lack of robustness in dispersion parameters, suggest a non-constant and a nonlinear batch effect influencing specific image regions.

1 Introduction

Repeatability and reproducibility challenges persist in omics sciences, limiting the generalizability and translation of biomarkers and signatures [1–3]. In omics, where data acquisition occurs in sequences or batches, the non-biological variation between sequences, known as batch effects, are well-documented and affect the statistical power and introduce strong bias in cross validation [2, 4]. The field of radiomics, which focuses on the extraction of features from radiographic imaging, faces similar challenges due to non-biological variations occurring throughout the acquisition process [5–9]. This paper considers a batch as a group of images sharing an acquisition parameter; - the reconstruction kernel in the context of this paper - with batch effect defined as the resulting non-biological variation from that parameter. The complexity of batch effect correction in radiomics is emphasized, given the multistep analysis process, numerous changing factors, and the retrospective nature of data [10]. Studies addressing this issue fall into three categories: identifying sources of batch effects, determining robust radiomics features, and developing correction methods for data harmonization such as *Combat* [6, 11–15]. Despite these multiple studies, consensus is lacking, leading to unresolved questions such as the mechanism behind different batch effects, the impact on various classes of radiomics features, and the effectiveness of correction methods in preserving relevant biological information. To contribute to resolving these gaps, this study evaluates the impact of kernel batch effects on various classes of radiomics features. Additionally, based on identified robust and non-robust features and statistical distribution assumptions, we propose a hypothesis on the mechanism of the studied batch effects. Addressing these questions is crucial for the development of better harmonization tools and refining data acquisition and analysis workflows to facilitate, clinical use.

2 Materials and Methods

The phantom data used in this study, was published by Usman et al[16]. To build this phantom, a multimaterial 3D printer (PolyJet Objet 260 Connex 3, Stratasys, Eden Prairie, Minnesota) equipped with Voxel Print software was employed to deposit droplets of UV-curable photopolymer resins in a layer-by-layer manner. Two base photopolymer resins, TangoPlus and VeroWhite, were used, with attenuation ranging from approximately 65 Hounsfield units (HU) to 125 HU at 120 kVp. The printer's resolution was on the order of 48×84×30 μm, finer than that of typical CT scanners. To replicate contrast differences observed in tumor-specific patterns on patient CT scans, multiple resin droplets were mixed in a single voxel using the Floyd Steinberg dithering algorithm. Four pancreatic ductal adenocarcinoma tumors were printed that captured the different morphological appearances and intensity patterns seen on CT scans. Prior to dithering, the HU values of each tumor and background were separately converted to double precision intensities within a range of 0 to 1. After 3D

printing, the phantom underwent 30 repeat scans using a 64-slice CT scanner (HD750, General Electric, Madison, Wisconsin) with the following parameters: 120 kVp, 0.7 s, 0.984; and the standard, bone, soft, ASIR and ASIRV kernels was applied to the data and two different tube currents were used (*100 mAs and 280mAs*). [16]

Radiomic Feature Extraction

The features were extracted from the original image using the computational environment for radiobiological research (CERR)[17]. 282 radiomics features were extracted on CT images, including first order statistics (FOS), gray-level co-occurrence matrix (GLCM), gray-level run-length matrix (GLRLM), gray-level size-zone matrix (GLSZM), neighborhood gray-tone difference matrix (NGTDM), and neighboring gray-level dependence matrix (NGLDM). CT images were discretized using a fixed bin width of 5. The minimum, maximum, mean, mean absolute deviation, and standard deviation for each of the texture feature were computed over all 13 directions to obtain rotational invariance. A patch-wise volume of 2 × 2 × 2 mm voxels was used for neighborhood gray level dependence features.**Statistical analysis**

- **Assessment of batch effect**

Two approaches were used to evaluate the batch effects. A PCA was conducted on a matrix of all the features and on the separated by class. The correlations within matrix were visualized for each data matrix separately to visualize the change in term of correlation using the Pearson correlation coefficient.

- **Robustness**

The assessment of the robustness was based on two dimensions: (1) the concordance and the (2) correlations. The concordance correlation coefficient (CCC) of Lin [18] was estimated to identify robust features [18, 19]. A threshold of ≥ 0.9 was used for the CCC based on McBride et al (*McBride et al, 2005*).

In addition, one of the goals of radiomics being to identify signatures, the Pearson correlation coefficient was calculated and used to identify the features that had the most stable correlations (Less than 10% change in correlation coefficient) with respect to the other features. All data analysis was carried out using R software version 4.1.3.

Combat batch effect correction methods

Combat works by modeling and removing batch-specific variations from the data through empirical Bayes frameworks. It adjusts expression values to create a harmonized dataset, reducing batch-related biases and enhancing the biological signal in high-throughput experiments.[12]

3 Results

Plotting the first two principal components of the full feature matrix revealed two clusters. The first cluster consisted of the standard kernel, ASIRs, ASIRV and soft kernel, and the second cluster consisted of the lung and bone kernels. Focusing on the standard, ASIR and ASIRV kernels, we found a gradient in the distance of these kernels to the standard kernel. This gradient grows from ASIR10 to ASIR 50 and from ASIRV1 to ASIRV5. Moreover, similarity found between ASIR20 and ASIRV1, and also between ASIR30 and ASIRV2 (Figure 1). The correlation matrix was still globally conserved with few changes in the ASIRs, the ASIRVs and the standard kernel; these small changes were more visible in the *skewness features*. However, very important changes appeared using the bone kernel and the lung kernel. These changes were larger for the *range, kurtosis, and skewness statistics*. The correlation between the energy, total energy and dispersion features were enhanced by the bone kernel and the lung kernel. In the other features class, the same trend was observed with very small changes in the ASIR. These findings highlight the ability of the ASIR and the standard kernel to conserve the functional relations between the first order statistics. In the neighborhood gray tone difference matrix, the correlations between the *contrast and the busyness, and the coarseness and the strength* were unstable, with a change of sign across different kernels. In addition, an analysis of the

correlation matrix of the neighborhood level dependance matrix showed notable changes in the *energy, entropy, gray level variance, and gray level non-uniform norm*. The correction of this batch effect by Combat resulted in kernel patterns disappearing after the application of the Combat correction on the ASIRs and the standard kernel. However, on the bone kernel and the lung kernel, Combat produced less improvement. Moreover, as shown in Figure 1, the application of the Combat correction methods to the kernel-related batch effects significantly improved the number of robust features from *13.7% to 73.2%*, where the first order statistics were still the most robust *(improvement from 22.72% to 81.82%)* and the size zone matrix features and the neighborhood gray level features were less robust.

4. Discussion

In this study, we have identified 40 features to be robust to CT image reconstruction kernels. The kernel batch effect affects the concordance and the correlation between the features in the same class. The application of Combat significantly improves the proportion of robust features. However, visual inspection reveals that its performance.

its performance is better when applied to ASIR, ASIRV and standard kernels, while bone and lung kernel harmonization still needs improvement.

The ability of Combat to improve the robustness of features has already been discussed and demonstrated in previous studies [12, 20, 21].

Evaluation of the robust features before correction revealed that the centrale tendance parameters were robust to the type of kernel. This robustness suggests that the batch effect does not significantly influence the centrale tendance and the global dispersion, although the dispersion parameters and skewness were changing. These measures detect variations in the shape of the distribution, suggesting changes in dispersion and symmetry of the intensity distribution between batches. Moreover, this contrast between the centrale tendance parameters and the dispersion and shape parameters highlight a non-linear and non-constant batch effect, with specific impact on certain regions of the image. This finding is consistent with the process of radiomics. In the processing, it is the voxels and the texture of an image that are affected by the batch effect not the extracted features directly.

Radiomics compared to some omics, extract features to characterize different aspect of an image. Generating then features that are mathematically related. However, most batch effect harmonization methods apply a form of feature-wise correction. Moreover, the batch effect takes a nonlinear and non-constant form. In this environment, the application of a linear or a constant correction model may overcorrect and unintentionally remove relevant radiomic features.

One of the main limitations of this study is the exclusive use of phantom data, as the findings must be validated in independent human datasets. In addition, the data analyzed in this study was extracted from a single scanner.

Future directions include the validation of these findings in a patient dataset and further exploration of the non-robust and robust feature sets.

5 Conclusion

In this study, we found that ASIRs and ASIRV were most similar to the standard kernel, while bone and lung kernels were dissimilar. Almost 14% of the features were robust to image reconstruction- kernel batch effects. This batch effect conserved the global voxel intensity distributions and the texture related distributions, while primarily affecting the dispersion parameters. From this pattern, we suggest a nonlinear and non-constant kernel batch effect model.

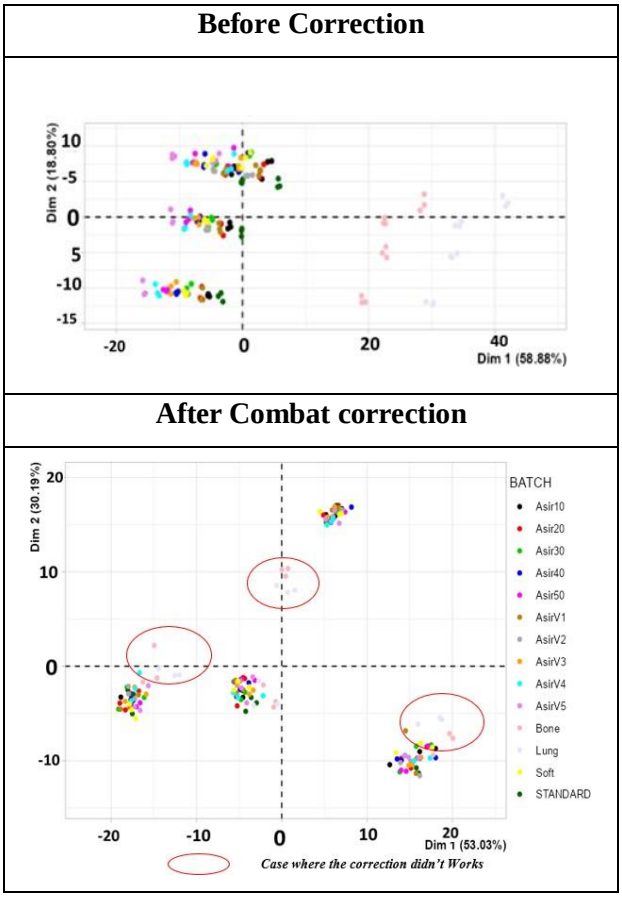


Figure 1: Principal components analysis before and after batch correction.

Table 1: Robust features before and after Combat

	ASIR					ASIRV				
	10	20	30	40	50	1	2	3	4	5
Before Combat	151	127	113	97	77	117	97	81	73	65
After combat	227	226	238	253	249	238	238	242	251	248

References

[1] Goh WWB, Wang W, Wong L. Why Batch Effects Matter in Omics Data, and How to Avoid Them. *Trends Biotechnol* 2017; 35: 498–507.

[2] Goh WWB, Wong L. Protein complex-based analysis is resistant to the obfuscating consequences of batch effects --- a case study in clinical proteomics. *BMC Genomics* 2017; 18: 142.

[3] Perng W, Aslibekyan S. Find the Needle in the Haystack, Then Find It Again: Replication and Validation in the 'Omics Era. *Metabolites* 2020; 10: 286.

[4] Liu Q, Walker D, Uppal K, et al. Addressing the batch effect issue for LC/MS metabolomics data in data preprocessing. *Sci Rep* 2020; 10: 13856.

[5] Castaldo R, Brancato V, Cavaliere C, et al. A Framework of Analysis to Facilitate the Harmonization of Multicenter Radiomic Features in Prostate Cancer. *J Clin Med* 2023;12:140

[6] Cavinato L, Massi MC, Sollini M, et al. Dual Adversarial Deconfounding Autoencoder for joint batch-effects removal from multi-center and multi-scanner radiomics data.Fornacon-Wood I, Mistry H, Ackermann CJ, et al. Reliability and prognostic value of radiomic features are highly dependent on choice of feature extraction platform. *Eur Radiol* 2020; 30: 6241–6250.

[7] Giriprakash PP. SYSTEMIC IDENTIFICATION OF RADIOMIC FEATURES RESILIENT TO BATCH EFFECTS AND ACQUISITION VARIATIONS FOR DIAGNOSIS OF ACTIVE CROHN'S DISEASE ON CT ENTEROGRAPHY.

[8] Hertel A, Tharmaseelan H, Rotkopf LT, et al. Phantom-based radiomics feature test–retest stability analysis on photon-counting detector CT. *Eur Radiol* 2023; 33: 4905–4914.

[9] Zhao B, Tan Y, Tsai WY, et al. Exploring Variability in CT Characterization of Tumors: A Preliminary Phantom Study. *Transl Oncol* 2014; 7: 88–93.

[10] Carles M, Fechter T, Martí-Bonmatí L, et al. Experimental phantom evaluation to identify robust positron emission tomography (PET) radiomic features. *EJNMMI Phys* 2021; 8: 46.

[12] Horng H, Singh A, Yousefi B, et al. Generalized ComBat harmonization methods for radiomic features with multi-modal distributions and multiple batch effects. *Sci Rep* 2022; 12: 4493.

[13] Ligero M, Jordi-Ollero O, Bernatowicz K, et al. Minimizing acquisition-related radiomics variability by image resampling and batch effect correction to allow for large-scale data analysis. *Eur Radiol* 2021; 31: 1460–1470.

[14] Lu L, Sun SH, Afran A, et al. Identifying Robust Radiomics Features for Lung Cancer by Using In-Vivo and Phantom Lung Lesions. *Tomography* 2021; 7: 55–64.

[15] Soliman MAS, Kelahan LC, Magnetta M, et al. A Framework for Harmonization of Radiomics Data for Multicenter Studies and Clinical Trials. *JCO Clin Cancer Inform* 2022; e2200023.

[16] Mahmood U, Apte AP, Kanan C, et al. Quality control of radiomic features using 3D-printed CT phantoms. *J Med Imaging* 2021; 8: 033505.

[17] Apte AP, Iyer A, Crispin-Ortuzar M, et al. Technical Note: Extension of CERR for computational radiomics: a comprehensive MATLAB platform for reproducible radiomics research. *Med Phys* 2018; 10.1002/mp.13046.

[18] Lin LI-K. A Concordance Correlation Coefficient to Evaluate Reproducibility. *Biometrics* 1989; 45: 255–268.

[19] Steichen TJ, Cox NJ. A Note on the Concordance Correlation Coefficient. *Stata J* 2002; 2: 183–189.

[20] Eshaghzadeh Torbati M, Minhas DS, Ahmad G, et al. A multi-scanner neuroimaging data harmonization using RAVEL and ComBat. *NeuroImage* 2021; 245: 118703.

[21] Mahon RN, Ghita M, Hugo GD, et al. ComBat harmonization for radiomic features in independent phantom and lung cancer patient computed tomography datasets. *Phys Med Biol* 2020; 65: 015010.

Leveraging hybrid FLT-PET/MR imaging to distinguish tumour progression from radionecrosis in brain metastasis patients post-radiosurgery

Michael Maddalena[1], Adam Farag[2], Noha Sinno[1], Paula Alcaide-Leon[2], Benjamin Schultz[3], Catherine Coolens[1,3]

[1]Department of Medical Biophysics, Faculty of Medicine, University of Toronto, Toronto, Canada
[2]Joint Department of Medical Imaging, University Health Network, Toronto, Canada
[3]Department of Radiation Oncology, Faculty of Medicine, University of Toronto, Toronto, Canada

Abstract An estimated 20% of stereotactic radiosurgery (SRS) patients relapse and experience tumour progression (TP) within 6 months of treatment. SRS patients may also be afflicted by radionecrosis (RN), a radiation-induced tissue injury virtually indistinguishable from TP via routine treatment follow-up scans. Invasive histopathology remains the only "gold-standard" TP/RN diagnostic method: non-invasive image-based protocols are urgently required to stratify patients while minimizing harm. Our study proposes leveraging FLT-PET/MR to achieve a superior diagnostic accuracy versus historical image-based protocols. Preliminary results suggest that the dynamic FLT-PET modelling-derived kinetic parameters Ki and Pf, as well as 10 MR-derived radiomic features, can significantly distinguish TP from RN. However, an increased sample number is required to confirm the generalizability of these findings. Our study underscores the potential of FLT-PET/MR in classifying metastatic brain lesions and suggest a path towards non-invasive diagnostic protocols to enhance patient stratification and favour outcomes in brain metastasis patients post-SRS.

1 Introduction

Despite excellent local control of brain metastases by the means of stereotactic radiosurgery (SRS), an estimated 20% of patients relapse within 6-months post-treatment[1]. Intracranial metastatic tumour progression (TP) is currently associated with a dismal overall survival rate of 8.1% at 24 months[2]. TP can be diagnosed by a combination of medical image-based diagnostic tools and behavioural tests, with patients exhibiting neurological symptoms including headaches, seizures and somnolence[3,4]. Conversely, patients exhibiting these symptoms may instead be afflicted by radionecrosis (RN), a necrotic tissue breakdown within the treatment region-of-interest (ROI) resulting from radiation exposure. RN is typically symptomatically indistinct from TP and is usually confined to the lesion treatment ROI, yet may appear in any irradiated region of the brain[5]. RN has been reported to affect up to 24% of radiosurgery patients[6]. Although the incidence of RN has been loosely correlated to the use of immunotherapy alongside SRS, there are no confirmed risk factors that may be employed as a predictor for susceptibility to RN[7]. Radionecrotic lesions are virtually indistinguishable from tumour progression when visualized using routine medical imaging, further complicating diagnosis[8]. Proper patient stratification is key in achieving positive treatment outcomes for cases of TP or RN, as either condition requires a significantly distinct regimen within which effective treatment options are available[9,10]. Currently, invasive histopathology remains the only "gold-standard" method of TP/RN differentiation[11]. The adoption of image-based diagnostic protocols with reduced harm has been hampered by poor diagnostic performance, with historical TP/RN differentiation efficacy lying at a modest 54%[12]. Thus, there is an urgent need for a strong non-invasive diagnostic protocol for TP/RN differentiation to improve patient stratification while minimizing harm. We propose leveraging a hybrid imaging platform, consisting of 18F-Fluorothymidine positron emission tomography (FLT-PET) to achieve improved diagnostic performance. FLT, a radiolabeled nucleotide precursor, functions as a non-invasive biomarker permitting the quantification of cellular proliferation, correlating strongly with the immunohistochemical proliferation marker Ki-67[13]. Additionally, we employ co-acquired anatomical magnetic resonance (MR) data from which radiomic features have been extracted to achieve a complimentary classification method. Radiomics permit the retrieval of quantitative metrics from medical images to be used as distinctive and informative features for tumor characterization[14]. Our study utilizes FLT-PET data fitted to compartmental models of tracer exchange alongside radiomics features extracted from co-acquired MR sequences with the aim of devising a protocol capable of achieving ≥80% diagnostic accuracy/sensitivity/specificity across all implemented classification methods.

2 Materials and Methods

To date, 8 adult patients with confirmed TP or RN have been enrolled in the context of a research ethics board-approved clinical study. Inclusion criteria for this study consist of a minimum age of 18 years old, a confirmed single or multiple brain metastasis diagnosis post-SRS treatment and a planned surgical resection. Lesion samples were collected from all enrolled patients during surgery, allowing histopathological confirmation of TP/RN. All patient datasets were co-acquired within the same session using the SIEMENS mMR Biograph hybrid PET/MR imaging platform (SIEMENS Healthineers, Erlangen, Germany). 18F-Fluorothymidine (FLT) was employed as a radiotracer with approval for experimental use, being produced jointly by the Centre for Probe Development and Commercialization (Hamilton, Canada) and CanProbe (Toronto, Canada). Patients were imaged immediately post-injection (p.i.) and surveyed over 20 minutes p.i. in a single bed position. Two PET datasets were acquired per patient:

a multi-timepoint dynamic PET (dPET) acquisition captured over the course of the entire imaging session in increasing intervals as well as a standard static PET acquisition at a 20 min p.i. (final timepoint). Attenuation-corrected PET datasets were analyzed using the PMOD software suite v.4.303 (PMOD Technologies LLC, Fällanden, Switzerland). PET datasets were registered to t2-weighted MR datasets to assist lesion contouring, using Normalized Mattes' Mutual Information as a cost metric. Volumes of interest (VOIs) were created on dPET data to match patient lesion contours available on planning t1-weighted MR datasets, with thresholding applied to lesion VOIs to reduce data outliers resulting from machine noise at early acquisition timepoints. Time-activity curves (TACs) describing lesion FLT uptake were generated from dPET data within the ROI and fitted to compartmental models. Arterial input functions were determined using an image-based method, via the creation of TACs from VOIs encompassing the common carotid artery. Lesion VOI uptake data were fitted to irreversible and reversible two-compartment models in the PMOD PKIN module. Kinetic parameters (K1, k2, k3 and k4) were derived from the best fitting model, determined via comparison of the Akaike Information Criterion. Additional kinetic parameters were calculated from the parameters obtained directly from modelling: influx rate (Ki), volume of distribution (Vd) and phosphorylation fraction (Pf) were inferred according to the following formulae:

$$Ki = \frac{K_1 \times k_3}{k_2 + k_3}$$

$$Vd = \frac{K_1}{k_2 + k_3}$$

$$Pf = \frac{k_3}{k_2 + k_3}$$

Kinetic parameters extracted from dPET modelling were subjected to analysis via the Pearson Correlation Coefficient to assess correlations between parameter values and histopathological ground truth lesion states.

Lesion contouring was performed anatomical MR sequences (t1w or t2w) used to plan radiotherapy treatment by a clinician with three years experience. All complimentary MR datasets were converted to Nearly Raster Raw Data format for integration into the Python 3 Simple ITK (sITK) open-source image analysis platform. Within sITK, complimentary MR datasets were resampled to align with the voxel spacings of the target planning image on a per-patient basis. All complimentary MR datasets were registered to the planning image to permit direct transposition of planning contours on co-acquired MR sequences. Three-dimensional rigid registration was achieved by maximization of Mutual Information using gradient descent optimization. Registered MR datasets were directly integrated into PyRadiomics, an open source radiomic feature extraction package. A comprehensive set of 107 radiomic features were extracted per dataset, forming a database of statistically mineable elements. Extracted radiomic features include spatially independent voxel intensity characteristics, referred to as first-order features, as well as spatially-correlative characteristics derived from computed gray-level matrices, referred to as higher-order features. Computed gray level matrices include gray-level co-occurrence matrices (GLCM), gray-level run-length matrices (GLCM), gray level size zone matrices (GLSZM), neighboring gray tone difference matrices (NGTDM) and gray level dependence matrices (GLDM). To improve computational efficiency, prevent range or scale-induced feature importance skewing and favour parameter interpretability, extracted radiomics data were normalized per-feature to a fixed range of [-1, 1]. Subsequently, a one-way ANOVA analysis was performed on each extracted radiomic feature with respect to the ground truth patient condition (TP/RN). The top 10 most correlative radiomic features were identified through inverse sorting of all collected ANOVA f-scores, with strict adherence to a predetermined two-tailed significance threshold.

3 Results

The FLT-dPET derived kinetic parameters Ki and Pf suggest that dPET analysis can distinguish TP from RN in post-SRS lesions significantly (0.063±0.043 and 0.304±0.204 versus 0.002±0.0 and 0.004±0.001 respectively, $p < 0.05$), whereas RN cannot be separated from background FLT uptake in healthy tissues ($p > 0.05$).

p#	Path.	Model	K1 [mL/ccm·min]	k2 [mL/g]	k3 [1/min]	k4 [1/min]	Ki [mL/ccm·min]	Vd [mL/ccm]	Pf	K1/k2 [g/ccm·min]
1	TP	2TK3	0.078±0.003	0.018±0.016	0.003±0.104	-	0.011	3.714	0.143	4.333
2	TP	2TK4	0.288±0.024	0.409±0.160	0.310±0.174	0.094±0.040	0.124	0.401	0.187	0.704
2	Hlthy	2TK3	0.467±0.449	22.593±22.476	0.118±0.065	-	0.002	0.020	0.005	0.021
3	TP	2TK4	0.067±0.015	0.653±1.159	1.499±2.747	0.396±0.206	0.047	0.031	0.697	0.103
3	RN	2TK3	0.925±0.335	27.390±9.918	0.074±0.023	-	0.002	0.034	0.003	0.034
4	TP	2TK4	0.136±0.032	1.000±1.038	1.124±0.957	0.207±0.056	0.072	0.064	0.529	0.035
5	TP	2TK3	0.282±0.022	0.309±0.060	0.067±0.018	-	0.050	0.750	0.178	0.913
6	TP	2TK3	0.099±0.040	0.180±0.025	0.027±0.013	-	0.013	0.478	0.130	0.550
8	TP	2TK4	0.456±0.099	1.393±0.646	0.501±0.218	0.162±0.043	0.121	0.241	0.265	0.327

Table 1: Kinetic parameters obtained from dynamic PET data fitting to both reversible (2TK4) and irreversible (2TK3) two-compartment models describing tracer distribution within target lesion ROIs. Model fitting performed on patients (p#) harbouring lesions with confirmed tumour progression (TP, n = 7), confirmed radionecrosis (RN, n = 1) as well as a healthy contralateral brain tissue control (Hlthy, n = 1).

Features were extracted from six out of eight patients within the study cohort; patient 1 was excluded owing to data unavailability, and patient 5 was excluded due to substantial artefacts within the dataset. One-way ANOVA analysis of radiomic datasets extracted from registered t2-weighted SPACE sequence images resulted in 10 significant features capable of differentiating TP from RN ($p < 0.05$). A

maximum f-score of 185.06 was produced by the NGTDM-extracted texture complexity feature.

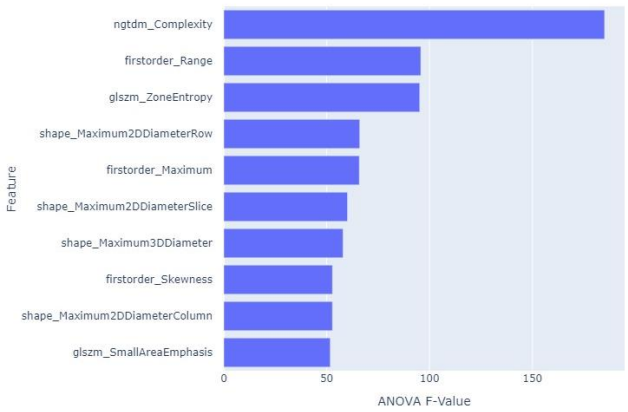


Figure 1: Top 10 significant radiomic features extracted from t2w-SPACE sequence MR datasets for TP/RN differentiation. Radiomic extraction performed on raw input images as well as gray level matrices computed from input images.

To increase result confidence and to permit further classification of data, expanding patient accrual to achieve a larger sample size (n) is required.

4 Discussion

In the exploration of parameters for stratifying TP and RN, Ki and Pf emerge as the most distinctive factors. k2, while approaching significant differentiation, is not fully capable of discerning TP from RN, but may achieve improved differentiation once an increased sample size is processed. Conversely, Vd, K1, and k3 show limited effectiveness in separating TP/RN cases. The absence of k4 in certain models renders it non-evaluable for TP/RN differentiation. Kinetically, FLT can operate reversibly within the tumour compartment, which theoretically warrants the use of a reversible model representing radiotracer de-phosphorylation (k4) across all cases[15]. Most FLT-PET kinetic studies presented in the literature employ a 60-minute dynamic acquisition window at minimum; our study protocol is constrained to a 20-minute acquisition due to the inclusion of hybrid magnetic resonance (MR) imaging alongside PET acquisitions, resulting in a requirement to reduce overall inter-modality and inter-sequence scan time to balance data depth with patient comfort. The use of a relatively short dPET acquisition time prevents cases with intrinsically slower tracer kinetics from achieving a full steady state, where the reverse step towards equilibrium is insufficiently captured. As a result, the value of k4 becomes negligible in select cases, allowing for the application of an irreversible model that excludes k4. While FLT may excel in distinguishing tumour recurrence extent in TP or mixed pathology cases, modelling radionecrosis proves challenging due to contouring requirements; while lesions composed mainly of TP are easily distinguishable from background signal, lesions containing primarily RN cannot be visually separated from healthy brain. FLT-PET remains a powerful diagnostic tool yet must be supplemented by MR scans to permit contouring and favour diagnosis, amending well to hybrid imaging platforms such as PET/MR. Overall, FLT-PET favours diagnosis of tumorigenic tissues presenting high radiotracer uptake, and lesions without specific FLT uptake and thus indifferentiable from background signal can be reasonably assumed to be radionecrotic on this platform.

Of the radiomic features extracted from the dataset, NGTDM-computed complexity was the most distinctive factor in differentiating TP from RN when considering the entire contoured region, followed by voxel intensity range and GLSZM-computed zone entropy. The top feature selected, NGTM complexity, alludes to a significant difference in the presence of specific irregular voxel intensity patterns between lesions harbouring TP versus RN. Notably, the top three features selected, as well as six out of the top ten features presented, do not directly relate to physical measurements or geometric characteristics of the lesion, aligning with challenges experienced by physicians in visually distinguishing TP/RN from geometric features on anatomical MR alone. Conversely, our results indicate that intensity-based features, which capture voxel values across various areas within the lesion, may harbor concealed higher-order information that could prove valuable in the differentiation of TP from RN. Utilizing these higher-order features, automated lesion classification may be feasible, facilitated by platforms such as machine learning or deep learning AI algorithms.

5 Conclusion

Our study examines the effectiveness of FLT-PET in categorizing metastatic brain lesions following stereotactic radiosurgery (SRS). While FLT-PET demonstrates high accuracy in identifying lesions primarily consisting of progressing tumors, it exhibits lower sensitivity in detecting radionecrotic areas, highlighting the necessity to incorporate data from additional imaging modalities to permit an accurate diagnosis of the condition of each lesion. Ultimately, the presented findings constitute a preliminary outlook at the proposed hybrid imaging protocol, as the current sample count is too small to draw significant conclusions regarding the true performance of the underlined imaging and data mining techniques. By increasing patient accrual and ultimately implementing machine learning and deep learning data classification methods within our protocol, we aim to advise future development of non-invasive diagnostic protocols for accurate differentiation of TP and RN; such advancements can potentially enhance patient outcomes through improved stratification of individuals post-SRS.

References

1. Patel, T. R. *et al.* A comprehensive review of MR imaging changes following radiosurgery to 500 brain metastases. *AJNR Am. J. Neuroradiol.* **32**, 1885–1892 (2011). doi: 10.3174/ajnr.A2668
2. Hartgerink, D. *et al.* LINAC based stereotactic radiosurgery for multiple brain metastases: guidance for clinical implementation. *Acta Oncol. Stockh. Swed.* **58**, 1275–1282 (2019). doi: 10.1080/0284186X.2019.1633016
3. Singh, K. *et al.* Update on the Management of Brain Metastasis. *Neurotherapeutics* **19**, 1772–1781 (2022). doi: 10.1007/s13311-022-01312-w
4. Lehrer, E. J. *et al.* The Cognitive Effects of Radiotherapy for Brain Metastases. *Front. Oncol.* **12**, (2022). doi: 10.3389/fonc.2022.893264
5. Hyperbaric Treatment of Brain Radiation Necrosis - StatPearls - NCBI Bookshelf ID: NBK431083
6. Chao, S. T. *et al.* Challenges with the diagnosis and treatment of cerebral radiation necrosis. *Int. J. Radiat. Oncol. Biol. Phys.* **87**, 449–457 (2013). doi: 10.1016/j.ijrobp.2013.05.015
7. Nichelli, L. & Casagranda, S. Current emerging MRI tools for radionecrosis and pseudoprogression diagnosis. *Curr. Opin. Oncol.* **33**, 597 (2021). doi: 10.1097/CCO.0000000000000793
8. Minniti, G. *et al.* Stereotactic radiosurgery for brain metastases: analysis of outcome and risk of brain radionecrosis. *Radiat. Oncol. Lond. Engl.* **6**, 48 (2011). doi: 10.1186/1748-717X-6-48
9. Truong, M. T. *et al.* Results of surgical resection for progression of brain metastases previously treated by gamma knife radiosurgery. *Neurosurgery* **59**, 86–97; discussion 86-97 (2006). doi: 10.1227/01.NEU.0000219858.80351.38
10. Vellayappan, B. *et al.* Diagnosis and Management of Radiation Necrosis in Patients With Brain Metastases. *Front. Oncol.* **8**, 395 (2018). doi: 10.3389/fonc.2018.00395
11. Detsky, J. S. *et al.* Differentiating radiation necrosis from tumor progression in brain metastases treated with stereotactic radiotherapy: utility of intravoxel incoherent motion perfusion MRI and correlation with histopathology. *J. Neurooncol.* **134**, 433–441 (2017). doi: 10.1007/s11060-017-2545-2
12. Telera, S. *et al.* Radionecrosis induced by stereotactic radiosurgery of brain metastases: results of surgery and outcome of disease. *J. Neurooncol.* **113**, 313–325 (2013). doi: 10.1007/s11060-013-1120-8
13. Chalkidou, A. *et al.* Correlation between Ki-67 immunohistochemistry and 18F-fluorothymidine uptake in patients with cancer: A systematic review and meta-analysis. *Eur. J. Cancer Oxf. Engl. 1990* **48**, 3499–3513 (2012). doi: 10.1016/j.ejca.2012.05.001
14. Mayerhoefer, M. E. *et al.* Introduction to Radiomics. *J. Nucl. Med. Off. Publ. Soc. Nucl. Med.* **61**, 488–495 (2020). doi: 10.2967/jnumed.118.222893
15. Kudomi, N., Maeda, Y., Hatakeyama, T., Yamamoto, Y. & Nishiyama, Y. Fully parametric imaging with reversible tracer 18F-FLT within a reasonable time. *Radiol. Phys. Technol.* **10**, 41–48 (2017). doi: 10.1007/s12194-016-0367-0

Towards fast plan adaptation for Proton Arc Therapy

Benjamin Roberfroid[1], Ana M. Barragán-Montero[1], John A. Lee[1] and Edmond Sterpin[1,2,3]

[1]Université catholique de Louvain – Center of Molecular Imaging, Radiotherapy and Oncology (MIRO), Brussels, Belgium
[2]Particle Therapy Interuniversity Center Leuven – PARTICLE, Leuven, Belgium
[3]KU Leuven – Department of Oncology, Laboratory of Experimental Radiotherapy, Leuven, Belgium

Abstract: Current approaches for dose optimization of proton arc therapy (PAT) plans suffer from optimization times ranging from several minutes to hours. In the context of adaptive proton therapy, we aim to speed up the plan reoptimization process. To achieve this, instead of optimizing a new PAT plan from scratch for each adaptive fraction, we propose to update the initial spot positions and their corresponding energies to the new anatomy of the day. Following this, we proceed to re-optimize only the most pertinent spot weights (less than 50% of all the spots). This results in less parameters to tune and thus in an accelerated plan adaption.

1 Introduction

Proton arc therapy (PAT) is an emerging modality that aims at delivering a continuously rotating proton beam around the patient, in opposition to the classical intensity modulated proton therapy (IMPT) which delivers beams on fixed angles. Consequently, PAT theoretically beneficiates of more degree of freedom than IMPT and has thus the potential to achieve better dose distribution. However, the problem of determining optimal energy layers for each angle is not trivial since it is necessary to minimize the treatment time (energy switch-ups are much slower than energy switch-downs for most of the proton installations), while at the same time, ensuring optimal dosimetric objectives [1]. This has given rise to several optimization approaches for PAT energy layers [1-5]. While approaching the problem from different perspectives, these methods suffer from optimization times ranging from several minutes to hours. In the context of adaptive proton therapy, we aim to speed up the plan reoptimization process. Instead of creating a new plan for each fraction, we propose updating spot positions and energies based on the daily anatomy. Following this, we then proceed to select only the most relevant spots to be re-optimized on this new anatomy, resulting in less parameters to tune and thus in an accelerated plan adaption. This follows an approach similar to what has been done by Botas et al. for IMPT using a different spot selection criteria [6].

2 Materials and Methods

The study has been retrospectively conducted on 3 head and neck patients who initially underwent a plan adaptation on a repeated CT acquired several days after the start of their treatment due to major anatomical changes. Local ethics committee approval by the CEHF (Comité d'Ethique hospitalo-facultaire) was obtained for the use of all patient data under the agreement "Learning from the past: the MIRO treatment planning database".

PAT plans were generated for the initial and repeated CT. Plans on repeated CT were generated in two different ways: classical, where a new PAT plan was optimized from scratch; and fast-adapted where the adapted plans resulted from re-optimization of a part of the initial spots that were displaced according to the new anatomy described below. Each plan featured a single 360-degree arc. Energy layers were defined with the ELSA algorithm [4] integrated in the RayStation treatment planning system (Raysearch Laboratories).

The fast-adaption method consists in the following steps:

1) Displace the initial spots on the new patient anatomy according a deformation vector field obtained by registering the initial CT on the repeated CT.
2) Update the energy of displaced spots according the new patient anatomy. Note that here the PAT energy layer pattern is constrained to be conserved i.e., energy switch-downs and switch-ups with respect to irradiation angles are preserved (see Figure 1).

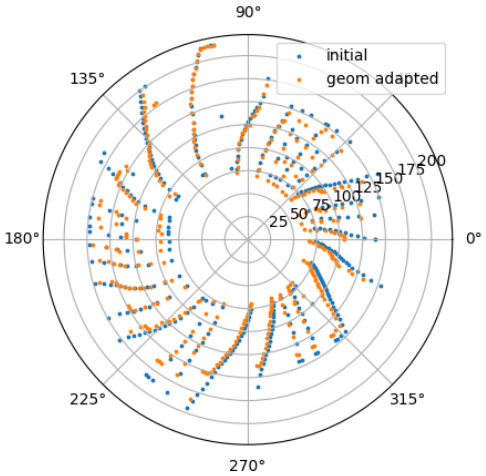


Figure 1: Example of polar plot for PAT energy layer pattern adaption (initially determined with ELSA here). Blue points denote the initial energy layers for each irradiation angle. Orange points denote the adapted energy layer while ensuring the initial PAT pattern. The angular position denotes the angle of irradiation while the radial position denotes the energy of the layer.

3) Select the most relevant spots for the re-optimization. The relevance criterion is defined by $\frac{\partial F_{obj}}{\partial w} > t$ where F_{obj} is the objective function of the optimization problem, w is the vector of spots weights and t is an arbitrary chosen threshold.
4) Continue the spot-weight optimization on the new anatomy only with the selected spots (the non-selected spots keep their dose contribution but their weights are not updated).

Planning of the different PAT plans as well as implementation of the fast adaption method was done on an open-source Python treatment planning system (TPS) [7].

Robust optimization was used for all cases with a range error of 2% and a setup error of [4.0,4.0,4.0] mm following a worst-case optimization approach. The robust constraints concerned the max and min dose to the CTV_p and CTV_n as well as the max dose to the spinal cord.

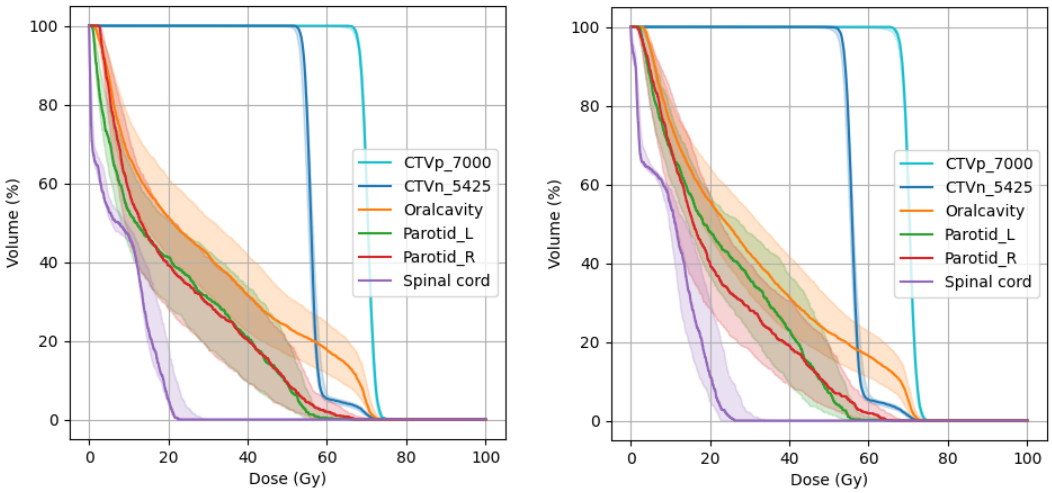


Figure 3: DVH for robust scenario of patient_1. Right: classical optimization; Left: fast-adaption

Concerning the optimization constraints, contraints from initial plan were re-used for the plan adaption. Comparison were thus done for both approaches (classical and fast-adaption) using these initial constraints.

3 Results

Early results show very similar dose distributions of fast adapted plan compared to classical optimization, e.g., no more than 2.44Gy difference to spinal cord max dose, 0.30Gy to CTV_n D95% or 0.55Gy to CTV_p D95%. Figure 1 illustrates this for the nominal case of patient_1. A similar trend for robust scenarios is observed and illustrated in Figure 2. Notice the respect of the robust constraints for every scenario in both plan optimization approaches.

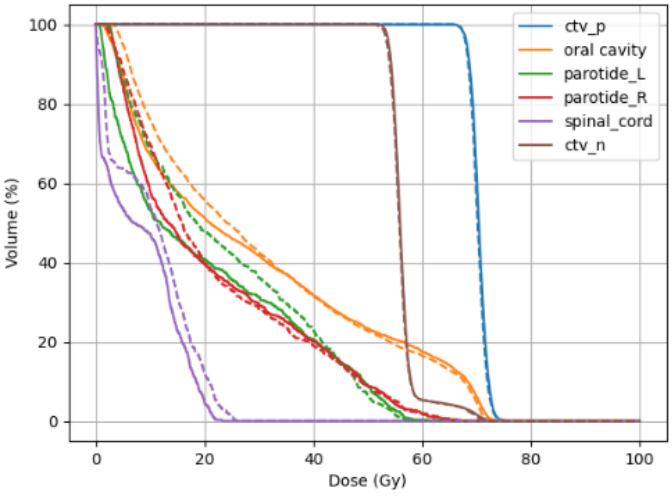


Figure 2: example of dose-volume histogram for PAT plan of patient_1. Solid lines represent the classical optimization. Dashed lines represent the fast-adaption.

Concerning the spots selection, currently the fast-adaption approach leads to re-optimizing less than 50% of them. This represents an acceleration of the optimization time from 3.5 to 5 times compared to the classical optimization.

4 Discussion

The goal of this study was to demonstrate the efficiency of our proposed fast-adaption method for PAT plan. We compared it to plan re-optimization fully from scratch. Results showed similar dose distribution achieved when comparing the two different optimization methods. Concerning the optimization time, limiting the number of spots to be re-optimized, substantially decreased the necessary time of optimization. However, future investigations should evaluate the results on a larger database to increase the confidence in the proposed method.

Besides, future investigations could include other energy layer determination approaches as their energy-angle pattern might drastically change from the ELSA one.

Finally, another interesting point would be to confront our method with other methods that have proven their value on IMPT, e.g., those based on analytical algorithm [8], those dose restoration-based [9], … on PAT plans in order to clearly highlight the strengths and weaknesses of each.

5 Conclusion

The proposed method successfully reduced the optimization time for PAT plan adaptation while achieving similar dose distribution showing that PAT could also beneficiate of faster plan adaption.

References

1. Wuyckens, S., et al., *Treatment planning in arc proton therapy: Comparison of several optimization problem statements and their corresponding solvers.* Computers in biology and medicine, 2022. **148**: p. 105609. DOI: 10.1016/j.compbiomed.2022.105609

2. Wuyckens, S., et al., *Bi-criteria Pareto optimization to balance irradiation time and dosimetric objectives in proton arc therapy*. Physics in Medicine & Biology, 2022. **67**(24): p. 245017. DOI: 10.1088/1361-6560/aca5e9
3. Ding, X., et al., *Spot-scanning proton arc (SPArc) therapy: the first robust and delivery-efficient spot-scanning proton arc therapy*. International Journal of Radiation Oncology* Biology* Physics, 2016. **96**(5): p. 1107-1116. DOI: 10.1016/j.ijrobp.2016.08.049
4. Engwall, E., et al., *Fast robust optimization of proton PBS arc therapy plans using early energy layer selection and spot assignment.* Physics in Medicine & Biology, 2022. **67**(6): p. 065010. DOI: 10.1088/1361-6560/ac55a6
5. Zhao, L., et al., *An evolutionary optimization algorithm for proton arc therapy*. Physics in Medicine & Biology, 2022. **67**(16): p. 16NT01. DOI: 10.1088/1361-6560/ac8411
6. Botas, P., et al., *Online adaption approaches for intensity modulated proton therapy for head and neck patients based on cone beam CTs and Monte Carlo simulations.* Physics in Medicine & Biology, 2018. **64**(1): p. 015004. DOI: 10.1088/1361-6560/aaf30b
7. Wuyckens, S., et al., *OpenTPS--Open-source treatment planning system for research in proton therapy.* arXiv preprint arXiv:2303.00365, 2023. DOI: 10.48550/arXiv.2303.00365
8. Matter, M., et al., *Intensity modulated proton therapy plan generation in under ten seconds.* Acta oncologica, 2019. **58**(10): p. 1435-1439. DOI: 10.1080/0284186X.2019.1630753
9. Borderias Villarroel, E., X. Geets, and E. Sterpin, *Online adaptive dose restoration in intensity modulated proton therapy of lung cancer to account for inter-fractional density changes.* Phys Imaging Radiat Oncol, 2020. **15**: p. 30-37. DOI: 10.1016/j.phro.2020.06.004

Fast robust dose evaluation for IMPT using deep learning

Hazem A. A. Nomer[1], Franziska Knuth[1], Joep van Genderingen[1], Dan Nguyen[2], Margriet Sattler[1], Uwe Oelfke[3], Steve Jiang[2], Linda Rossi[1], Ben Heijmen[1], and Sebastiaan Breedveld[1]

[1]Erasmus MC Cancer Institute, University Medical Center, Department of Radiotherapy, Rotterdam, Netherlands.
[2]UT Southwestern Medical Center, Medical Artificial Intelligence and Automation (MAIA) Laboratory, Department of Radiation Oncology, Dallas, USA.
[3]The Institute of Cancer Research and The Royal Marsden NHS Foundation Trust, Joint Department of Physics, London, United Kingdom.

Abstract Intensity Modulated Proton Therapy (IMPT) is susceptible to uncertainties in patient setup and proton range. Robust optimization is therefore employed in IMPT treatment planning to ensure sufficient coverage of the Clinical Target Volume (CTV) in predefined scenarios, albeit at a price of significantly increased planning times.
We investigated three Deep Learning (DL) strategies for dose predictions in all separate robustness scenarios for the chosen robustness settings in head-and-neck (H&N) cancer robust IMPT treatment planning: 1) a separate prediction model per scenario, 2) a separate model per scenario, but using transfer learning, and 3) an integrated single model for all scenarios.
We introduced scenario specific isocenteric distance maps as input to the models, to cope with resolution differences in the planning Computer Tomography (CT) scans and to account for different setup error scenarios.
For 392 H&N cancer patients, high-quality 9-scenario ground truth (GT) robust plans were fully automatically and consistently generated with wish-list driven Multi Criteria Optimization (MCO).
DL strategy 3, using a single prediction model for all scenarios, acheived best correspondence to the GT plans, with the predicted $V_{95\%}$ of the voxelwise minimum dose (VWmin) for the low dose CTV, within 1%-point of agreement with the GT for 91% of the test patients, compared to only 8% and 56% for strategies 1 and 2 respectively. For the high dose CTV, the within 1%-point agreements were 80%, 20% and 0% for strategies 3, 2 and 1 respectively.
With the current implementation (for 9 robustness scenarios), an IMPT dose distribution can be generated and assessed on robustness in under one minute.

1 Introduction

Intensity-modulated proton therapy (IMPT) enables highly conformal tumor coverage with improved sparing of adjacent normal tissues compared to photon therapy. IMPT is considered particularly beneficial for cancers of the head and neck (H&N) due to many organs at risk (OARs) involved and their proximity to the target volumes. However, IMPT is susceptible to uncertainties that arise from inter-fractional variations (e.g. patient setup) and proton range uncertainty [1].
Robust (worst-case) optimization is commonly used in IMPT treatment planning to obtain adequate Clinical Target Volume (CTV) coverage and limit high doses to OARs in a set of predefined scenarios [2–5]. However, robust optimization comes with the cost of extended planning time [6].
Deep learning (DL) models have been proposed recently to alleviate the problem of long planning times. Although DL has demonstrated to be capable of predicting realistic dose distributions in the order of seconds, published studies on using it in the context of IMPT plan robustness evaluations have limitations [7, 8].
In [7], DL is used to predict scenario dose distributions, but further dose mimicking is required to make the overall plan robust. In another study [8], DL is used to predict only the nominal scenario dose of a robust plan, and dose mimicking is employed to obtain a robust plan, considering all robustness scenarios.
In this study, we investigated three strategies that use DL for direct dose prediction in all separate scenarios for the chosen robustness settings in H&N cancer robust IMPT treatment planning: 1) a naive approach, training a separate model for each scenario individually, 2) a model for each scenario similar to strategy 1, but using transfer learning to speed-up the training, and 3) an integrated single model where labels were used to predict the dose for each specific scenario.
Scenario-specific distance maps with voxel distances to the isocenter were used as input for the prediction model(s) for all three strategies, playing a pivotal role in augmenting a model's capacity to incorporate setup error scenarios in the context of patients with various Computer Tomography (CT)-scan resolutions.

2 Materials and Methods

2.1 Data preparation

The dataset included contoured planning CT-scans of 392 H&N cancer patients with a variety of tumor locations. All patients were fully automatically robustly planned using Erasmus-iCycle for wish-list driven Multi Criteria Optimization (MCO) [9, 10], resulting in high-quality plans with consistent trade-offs. Nine scenarios were used for the robust planning: nominal scenario (no errors), 2 range errors (RE) scenarios (+3%/-3%) and 6 setup errors (SE) scenarios (+3/-3 mm) in the three directions X, Y & Z. Prescribed doses were 70 Gray equivalent (GyE) and 54.25 GyE for the high dose CTV and low dose CTV, respectively. For all patients, the clinical contours for both CTVs and OARs were used.
After the plan generation, the resulting dose distributions for all patients and all scenarios were stored as the ground truth (GT) for models training and evaluation. Before training

the models and as a preprocessing step, the Digital Imaging and Communications in Medicine (DICOM) files of the CT, Radiotherapy (RT) structure sets and the 9 scenario doses were converted into a voxelized version and downsampled to a grid size of 256 x 256 x 128 voxels. To determine the bounding box of the patients, a composite patient was constructed by overlaying all patients. This resulted in a voxel size of 1.8 x 1.3 mm^2 in XY (right-left and anterior-posterior) direction, and 2 or 2.5 mm in the Z direction, which was left unscaled to avoid large resampling artifacts. We introduced the isocentric distance map to handle patient-dependent Z resolution, see section 2.2.
Furthermore, the contoured targets, OARs and the patient external were converted into 3D binary masks, in which each patient voxel was set to 1 if the voxel was inside and 0 otherwise. The result of the data preparation process consisted of 10 binary masks for the structures (2 CTVs, 7 OARs and the patient external), 1 CT image, and the 9 scenario dose distributions. The CT and the binary masks were used as inputs for all of the prediction models.

2.2 DL strategies for scenario specific dose prediction

A naive way to predict robustness scenarios dose distributions is to train a DL model for each scenario independently (strategy 1) . The limitation of this approach is long computation times for both training of each model and for predictions.
An option to overcome this limitation, is to use transfer learning (strategy 2) in which we use the model trained to predict the nominal scenario dose distribution as a pretrained model and fine-tune it to predict the scenarios dose distributions, training again a model for each scenario. It would reduce the training times that would be required to train the models for the other scenarios from scratch.
As a third option (strategy 3), all scenarios were trained within one model, with a one-hot-vector (label) [11] added to explicitly indicate an individual scenario. The label represents the scenario index (in our case from 0-8). An embedding vector for the scenario index was produced using a learnable embedding layer. The label was used to steer the model towards the desired scenario dose distribution to be predicted.
The isocentric distance map was calculated by determining the Euclidean distance between the voxel at the isocenter and every voxel contained within the patient contour. The distance maps were provided as input to the models to introduce the isocenter information and handle different Z resolutions. For all strategies, the isocentric distance map was shifted according to the setup error scenario, i.e. the isocenter and thus the isocentric distance map was shifted +3/-3 mm in the respective directions. For the range uncertainty, the isocentric distance map was kept at the nominal scenario, but the HU of the CT is scaled with +3%/-3%. This approach mirrors the methodology employed in dose calculation engines, where simulation of SE scenarios involves deflecting beam directions and scaling the Hounsfield unit (HU) (or stopping power ratios) of the CT to compute doses for RE scenarios.

2.3 Experimental Setup

The dataset of 392 H&N patients was divided into training (283), validation (31) and test (78) sets. All models were based on the Hierarchical Dense (HD) UNet [12, 13]. Except for strategy 3, which used a label for each scenario passed to the decoding path of the HD-UNet, the inputs were the same for all strategies: the CT, the binary structure masks and the isocenter distance map for each patient.
To improve the training convergence, isocenter distance maps were normalized by 300 mm, the HU of the CT were clipped using a window of $[-1000, 1500]$ [14] and normalized by 1000, and the dose distributions were also normalized by the prescribed dose for CTV High (70 GyE).
All the experiments were done on a single NVIDIA A100 TensorCore 80 GB GPU. We used the Adam optimizer [15] with learning rate 10^{-4} with a batch size 1, and the voxelwise mean square error as a loss function. After every 10 epochs of stagnation and no improvement in the loss over the validation data, the learning rate is reduced by a factor of $\frac{3}{4}$.
For the transfer learning in strategy 2, in scenario models training we used the final state of the optimizer of the nominal model as the starting point of the optimizer for all scenario models, the last learning rate used in the last training epoch of the nominal model was used as the initial rate in the scenario models, to only allow fine-tuning of the model weights.
For training the single model (strategy 3) we randomly sampled a patient from the training set and a scenario each time.
Dose metrics for predicted target and OAR doses were computed and compared with the GT. For robustness evaluations VWmin and VWmax dose distributions were constructed from the 9 predicted dose distributions. The CTVs were evaluated on VWmin $V_{95\%}$ and VWmax near maximum $D_{0.03cc}$. Also robustly optimized OARs (brainstem and spinal cord) were evaluated on VWmax $D_{0.03cc}$. Statistical significance was tested using the Wilcoxon signed rank test with $p < 0.05$ as threshold for significance.

3 Results

Figure 1 shows for the 78 test patients, differences in predicted and GT VWmin and VWmax dose distributions. All strategies show high accuracy CTV coverage prediction, but strategy 3 clearly outperformed the other 2 (two upper panels in Fig. 1). For CTV Low, percentages of predicted plans with VWmin $V_{95\%}$ within 1%-point of the GT were 91% for strategy 3, compared to 8% and 56% for strategies 1 and 2, respectively. For CTV High, these percentages were 80%, 0% and 20%, respectively. Strategies 1 and 2 performed slightly better for VWmax $D_{0.03cc}$ of CTV High, while OAR

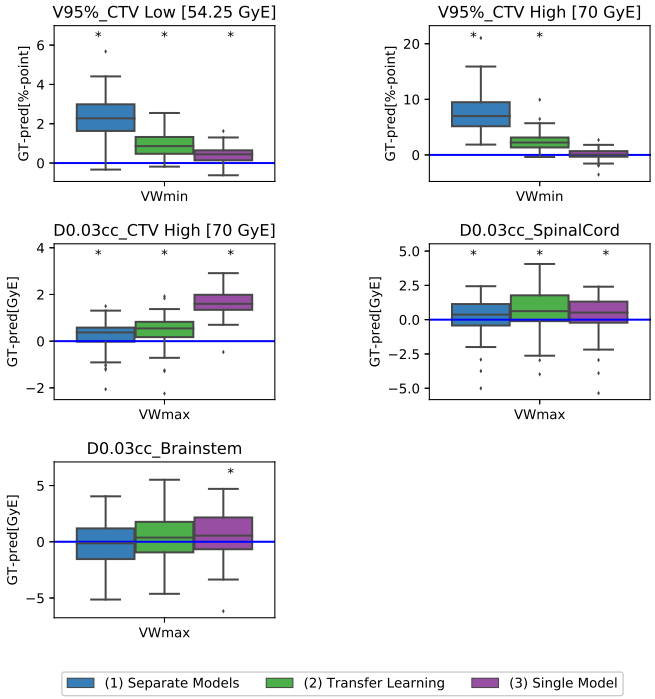


Figure 1: For the 78 test patients, pairwise differences between in dosimetric parameter values of voxelwise min or max dose distributions of predicted and GT plans. * = significantly different ($p < 0.05$).

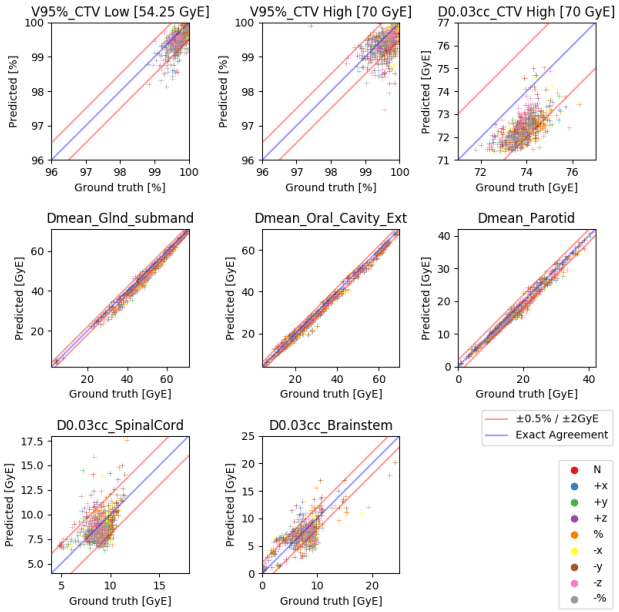


Figure 2: Predicted vs ground truth doses metrics for DL strategy 3. Each marker represents a predicted dose parameter with the corresponding ground truth value. Marker colours refer to investigated scenarios.

doses were highly similar. Figure 2 shows for the test patients excellent agreement between plan parameters of the individual scenario doses predicted with strategy 3 and GT parameters.

Strategy 3 was most efficient in terms of calculation times. Training times for strategies 1, 2 and 3 were 51, 24, and 16 hours respectively. Models loading times prior to predictions were 21.55 seconds for 1 and 2 and 3.85 seconds for loading the single model in 3. Average prediction times per scenario were 1.17 for 1 and 2 and 1.14 seconds for 3.

4 Discussion & Conclusion

Strategy 3, using a single DL model for predicting dose in all separate IMPT robustness scenarios with scenario labels as input, showed overall closest agreement between predicted doses and GT, outperforming the results of the other two investigated prediction strategies, in particular for target coverages. Only near-maximum doses were slightly underestimated. Given the complexity of the H&N treatment site and the limited training set size of 283 cases, it is expected that results will improve with increasing dataset size.

Strategy 3 was also most efficient in calculation times with training times of 16 hours (instead of 51 and 24 hours for strategies 1 and 2 respectively). For individual patients, the model loading time for the prediction of the scenario doses of a patient with strategy 3 was 3.85 seconds and prediction took on average 1.14 seconds per scenario, which results in a prediction time of 14 seconds for all scenarios. For both strategies 1 and 2, the overall prediction time is 32 seconds, as each scenario requires loading a new model to the GPU memory. Both in training time and robust prediction time, strategy 3 also has the best potential to scale up to larger numbers of scenarios, such as 21 or 29 as commonly used in IMPT. In the 21 scenario case, the overall prediction time would increase to 75 seconds for strategies 1 and 2 in contrast to 28 seconds for strategy 3.

References

[1] A. C. Kraan, S. Van De Water, D. N. Teguh, et al. “Dose uncertainties in IMPT for oropharyngeal cancer in the presence of anatomical, range, and setup errors”. *International Journal of Radiation Oncology* Biology* Physics* 87.5 (2013), pp. 888–896.

[2] W. Liu, X. Zhang, Y. Li, et al. “Robust optimization of intensity modulated proton therapy”. *Medical Physics* 39.2 (2012), pp. 1079–1091. DOI: https://doi.org/10.1118/1.3679340.

[3] J. Unkelbach and H. Paganetti. “Robust Proton Treatment Planning: Physical and Biological Optimization”. *Seminars in Radiation Oncology* 28.2 (2018). Proton Radiation Therapy, pp. 88–96. DOI: https://doi.org/10.1016/j.semradonc.2017.11.005.

[4] E. W. Korevaar, S. J. Habraken, D. Scandurra, et al. “Practical robustness evaluation in radiotherapy – A photon and proton-proof alternative to PTV-based plan evaluation”. *Radiotherapy and Oncology* 141 (2019), pp. 267–274. DOI: https://doi.org/10.1016/j.radonc.2019.08.005.

[5] R. M. Anakotta, H. P. van der Laan, S. Visser, et al. "Weekly robustness evaluation of intensity-modulated proton therapy for oesophageal cancer". *Radiotherapy and Oncology* 151 (2020), pp. 66–72. DOI: https://doi.org/10.1016/j.radonc.2020.07.015.

[6] R. G. Kierkels, A. Fredriksson, S. Both, et al. "Automated Robust Proton Planning Using Dose-Volume Histogram-Based Mimicking of the Photon Reference Dose and Reducing Organ at Risk Dose Optimization". *International Journal of Radiation Oncology*Biology*Physics* 103.1 (2019), pp. 251–258. DOI: https://doi.org/10.1016/j.ijrobp.2018.08.023.

[7] O. Eriksson and T. Zhang. "Robust automated radiation therapy treatment planning using scenario-specific dose prediction and robust dose mimicking". *Medical Physics* 49.6 (2022), pp. 3564–3573.

[8] E. Borderias-Villarroel, M. Huet Dastarac, A. M. Barragán-Montero, et al. "Machine learning-based automatic proton therapy planning: Impact of post-processing and dose-mimicking in plan robustness". *Medical Physics* 50.7 (2023), pp. 4480–4490.

[9] S. Breedveld, P. R. Storchi, P. W. Voet, et al. "iCycle: Integrated, multicriterial beam angle, and profile optimization for generation of coplanar and noncoplanar IMRT plans". *Medical physics* 39.2 (2012), pp. 951–963.

[10] S. van de Water, A. Kraan, S. Breedveld, et al. "Improved efficiency of multi-criteria IMPT treatment planning using iterative resampling of randomly placed pencil beams". *Physics in Medicine & Biology* 58.19 (2013), p. 6969.

[11] R. Gao, B. Lou, Z. Xu, et al. "Flexible-Cm GAN: Towards Precise 3D Dose Prediction in Radiotherapy". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 715–725.

[12] D. Nguyen, X. Jia, D. Sher, et al. "3D radiotherapy dose prediction on head and neck cancer patients with a hierarchically densely connected U-net deep learning architecture". *Physics in medicine & Biology* 64.6 (2019), p. 065020.

[13] D. Nguyen, T. Long, X. Jia, et al. "A feasibility study for predicting optimal radiation therapy dose distributions of prostate cancer patients from patient anatomy using deep learning". *Scientific reports* 9.1 (2019), p. 1076.

[14] N. T. Singh, C. Kaur, A. Chaudhary, et al. "Preprocessing of Medical Images using Deep Learning: A Comprehensive Review". *2023 Second International Conference on Augmented Intelligence and Sustainable Systems (ICAISS)*. 2023, pp. 521–527. DOI: 10.1109/ICAISS58487.2023.10250462.

[15] D. P. Kingma and J. Ba. "Adam: A method for stochastic optimization". *arXiv preprint arXiv:1412.6980* (2014).

Inverse robustness to control uncertainties in the RBE parameter in proton therapy planning

Mara Schubert[1]

[1]Department of Optimization, Fraunhofer Institute for Industrial Mathematics, Kaiserslautern, Germany

Abstract The relative biological effectiveness (RBE) of protons in comparison to photons depends on many factors. However, no highly accurate RBE model exists for varied cases that could be applied in treatment planning. Instead, in practice a constant RBE factor of 1.1 is used. Consequently, the effect of proton radiation might be over- or underestimated.

Here we introduce an inverse robustness model to maximize robustness of plans with respect to variations in the RBE parameter, without relying on a specific RBE model. Moreover, there is no need to specify a robustness measure beforehand. Importantly, our approach does not require a predefined robustness measure but instead focuses on maximizing the range of RBE values a plan can accommodate while ensuring that dose objectives remain within a predifined range.

We applied this concept to a prostate and a head and neck case. Our results show that it is possible to slightly shift the dose distribution to gain robustness against deviation from 1.1 in the RBE parameter for some structures, without significantly altering the overall treatment plan. Other structures demonstrated less robustness against variations in RBE.

Therefore, the outcomes of inverse robust optimization can be used to assess the robustness of a treatment plan against RBE variations or to identify structures that might need additional preventive measures.

1 Introduction

The biological effect of a given physical dose varies depending on the type of radiation. It is well-known that the relative biological effectiveness (RBE) of protons compared to photons is not constant but depended on factors such as dose, linear energy transfer (LET), and tissue parameters [1]. However, there are significant differences among various RBE models, and the uncertainty in tissue parameters further complicates the matter [2]. As a result, in clinical proton therapy planning, a constant RBE parameter of 1.1 is used, assuming that protons are 10% more effective than photons. Considering the uncertainties in the RBE is crucial when assessing the robustness of a treatment plan, as the actual RBE determines the effects of the radiation.

Many robustness concepts require to previously decide on the scenarios a plan should cover. The lack of a generally accepted model for the RBE parameter makes this difficult.

The concept of inverse robustness, as introduced in [3], does not require the a priori definition of an uncertainty set. Instead, it identifies the largest uncertainty set that can be accommodated when treating an uncertainty parameter as variable rather than fixed at a constant nominal value.

We applied this concept to the proton therapy planning problem with the RBE factor as uncertainty parameter. The aim is to find a treatment plan that is nearly as good as the treatment plan calculated with the nominal RBE value of 1.1, yet maximizes the measure of the set of acceptable RBE values. Hence, it enables treatment planners to find the plan that is the most robust with regards to variations in the RBE parameter and meets the dose objectives nearly as good as a conventionally chosen plan. The results then allow an evaluation whether the range of covered RBE values is sufficiently broad or if there is a need to find a different plan with better coverage or use further methods to limit the expected RBE variations to the interval covered by robustness.

2 Materials and Methods

2.1 Proton Therapy Optimization Framework

A typical proton therapy planning problem has the following structure:

$$\begin{aligned} \min_{x\in X} \quad & F_{\bar{r}}(x) = (f^1_{\bar{r}}(x), ..., f^n_{\bar{r}}(x)) \\ s.t. \quad & g^j_{\bar{r}}(x) \leq 0 \qquad \forall j \in \{1, ..., m\}. \end{aligned} \tag{PPP}$$

X denotes the set of feasible treatment plans. $F_{\bar{r}}(x)$ is vector valued. The single objective functions $f^i_{\bar{r}}(x)$ and the constraints $g^j_{\bar{r}}(x)$ depend on the beamlet intensities x. The RBE factor r determines how the dose is weighted. It is a constant parameter with nominal value $\bar{r} = 1.1$. The solutions are the best compromises between the objectives, the so called Pareto points.

2.2 The Inverse Robustness Problem

Given a solution of (PPP), the aim of inverse robustness is to maximize the volume of the set in which the uncertainty parameter can be vary, without degenerating the previously found solution. To achieve this, we fix the objectives values of the plan in the constraints of the inverse robustness problem.

Here, we treat the RBE factor as the uncertainty parameter. We allow a different value for the RBE parameter in each structure s. We measure the volume of the uncertainty set by summing the lengths of intervals, leading to the following

a) dose evaluation functions	
gEUD	$\sum_{i\in S}(r_s(D\cdot x)_i)^p$
OD	$\sum_{i\in S}(r_s(D\cdot x)_i - d_{ref})_+^p$
UD	$\sum_{i\in S}(d_{ref} - r_s(D\cdot x)_i)_+^p$
b) with relevant interval boundary	
gEUD	$\sum_{i\in S}(b_s(D\cdot x)_i)^p$
OD	$\sum_{i\in S}(b_s(D\cdot x)_i - d_{ref})_+^p$
UD	$\sum_{i\in S}(d_{ref} - a_s(D\cdot x)_i)_+^p$

Table 1: The dose evaluation functions we used for our model: gEUD: generalized equivalent uniform dose, OD: overdose, UD: underdose. D refers to the dose influence matrix. The product $D\cdot x$ is the dose in each voxel. The exponent p is a positive integer and the expression $(y)_+$ means $\max\{0,y\}$.

formulation:

$$\begin{aligned} \sup_{\substack{x\in X,\ W\subseteq U,\\ W=[a_i,b_i]^s}} \quad & \sum_{s\in S} b_s - a_s \\ s.t. \quad & f^i(x,r_s) \le f^{*i} + \varepsilon^i \qquad \forall r_s \in W,\ \forall i \in \{1,...,n\} \\ & g^j(x,r_s) \le 0 \qquad\qquad \forall r_s \in W,\ \forall j \in \{1,...,m\} \\ & \bar{r} \in W. \end{aligned} \tag{IRP}$$

U is the feasible set of the interval boundaries for the RBE parameter. We defined $U = \{[a_1,b_1]\times ... \times [a_s,b_s] : t_1^i \le a_i \le 1.1 \le b_i \le t_2^i\ \forall i \in S\}$, where S is the index set of the structures and t_1 and t_2 are the thresholds for the lower and upper interval boundaries, respectively. The objective values of the treatment plan chosen from (PPP) are f^{*i}.

The following objectives to evaluate dose were considered: generalized equivalent uniform dose, overdose and underdose (Table 1 a)). All are monotonous in r_s. Consequently, the constraints of (IRP) are satisfied for all $r_s \in W$ if they are met at the critical interval boundaries for each function. Table 1 b) shows which interval boundaries to use.

The constraints still hold if the RBE parameter is not constant in a structure, but varies within the interval $[a_s,b_s]$. This allows our approach to cover variable RBE values without the need to rely on a model for variable RBE.

Clearly, if only one dose evaluation function belongs to a structure s, the corresponding interval for r_s reaches one of the predefined threshold for the interval boundary. It is advisable to set tight limits for these structures. Otherwise they might outweigh intervals of structures where relatively little variation is possible.

2.3 Case Study

We applied the concept to two cases from the CORT dataset [4]: A prostate case with four gantry angles at (60°, 120°, 240°, 300°) and a head and neck case with four gantry angles at (45°, 90°, 270°, 315°).

For the prostate case we calculated the largest possible range of RBE values without plan degeneration and with an allowed relative deviation of 5% for maximal possible intervals of $[1,2]$ in all structures. Subsequently, we narrowed the allowed interval for some structures, and lastly we maximized the RBE interval for the rectum, both at a relative deviation of 5%.

For the head and neck case we used relative deviations of 5% and 10% from the nominal treatment plan. We used two sets of structures where the RBE range was maximized: One with several structures of the treatment site and one where the focus lay on structures whose dose was controlled by objective function in (PPP), rather than just by the constraint functions.

MatRad [5] was used to calculate the dose influence matrices and create the heat maps of the dose distribution. The optimization of (PPP) and (IRP) were conducted using python.

3 Results

When adhering strictly to the original dose objectives, the solution of the inverse robustness problem relies completely on degrees of freedom in the dose distribution to maximize the possible intervals for the RBE parameters.

In the prostate case, the right and left femoral heads demonstrated robustness against RBE values as high as 1.57 and 1.63, respectively. For the other structures (Rectum, Bladder and planning target volumes (PTV)) no deviation from 1.1 was tolerable.

When increasing the permissible deviation from the treatment plan to 5%, both femoral heads were robust against RBE variations in the whole considered interval of $[1,2]$. The upper threshold for the PTV56 increases to 1.14 and for the whole body to 1.32. However, the covered interval for the rectum, bladder, and PTV68 remained unchanged.

By reducing the upper interval boundaries to 1.2, the tolerance for the RBE in the rectum increased to 1.11, and in the PTV56 to 1.19.

To explore the largest possible interval for the rectum at 5% deviation, we optimized only the interval length for the rectum. This resulted in an RBE interval for the rectum of $[1,1.15]$. Figure 1 shows the dose volume histograms (DVHs) and a heat map for each scenario. The DVHs are similar, whereas slight shifts in the dose distribution lead to plans that are more robust to variations in the RBE factor.

In [6] a method to analyze RBE was presented and applied to clinical data of prostate cases. Depending on the case type and fractionation, they obtained 95% confidence intervals all within $[1.045, 1.213]$. Although they mentioned that their analysis did not take all aspects into account that influence RBE, it gives an orientation about the range of the RBE parameter. The absolute difference of 0.168 observed in their study aligns with the range our results also cover for some structures.

As second example we analyzed a head and neck case. Set-

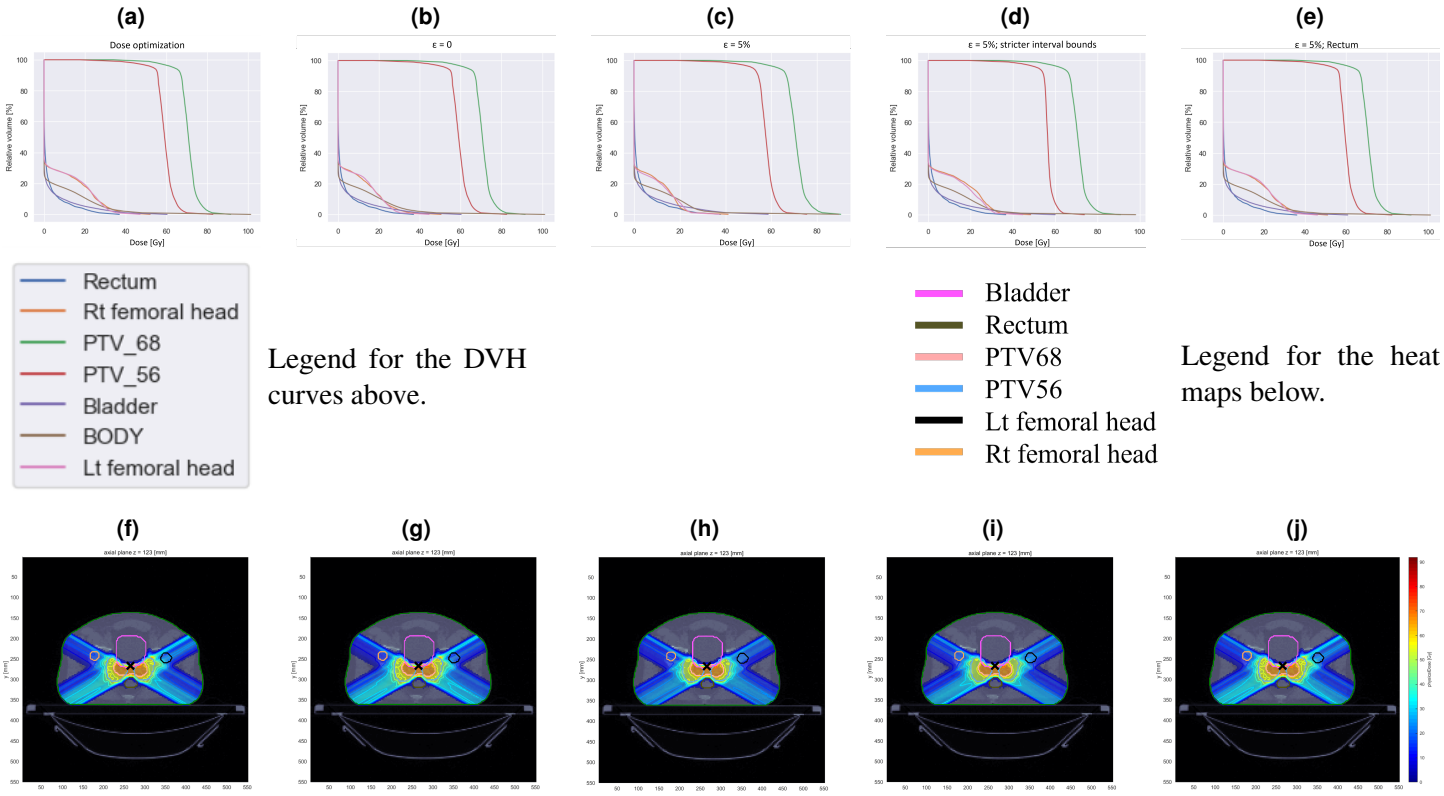


Figure 1: Prostate case. Dose is weighted with constant factor 1.1. Upper row: DVH curves, lower row: heat maps dose of an axial slice for a), f) dose optimization; inverse robustness with b), g) $\varepsilon = 0$, c), h) relative deviation of 5% and upper interval threshold of 2, d), i) relative deviation of 5% and upper interval threshold of 1.2 , e), j) relative deviation of 5% and only inverse robustness optimization for RBE parameter in the rectum.

ting the allowed relative deviation to 5% resulted in significant RBE intervals only for structures whose dose was controlled exclusively through constraints rather than minimized in (PPP).

We then focused on maximizing the RBE intervals for both parotid glands, PTV63, and PTV70, setting an upper interval threshold of 2 for the latter structures, and of 1.2 for the skin. For all other structures the interval length was not optimized. The skin was robust against RBE variations in the whole allowed interval. The resulting intervals for PTV63 and PTV70 were $[1.1, 1.136]$ and $[1.1, 1.143]$, respectively. The parotid glands were not robust against RBE variations.

Increasing the allowed deviation to 10%, while maintaining the same interval thresholds, increased the RBE intervals for PTV63 and PTV70 to 1.163 and 1.181, respectively. However, the intervals for the parotid glands could not be enlarged, reaffirming their sensitivity to RBE fluctuations. Figure 2 shows the DVH curves and heat maps for this case and the described settings for inverse robustness optimization.

4 Discussion

Robustness concepts typically necessitate trade-offs in the objective values. Unlike traditional methods where robustness levels are predetermined, inverse robustness explores achievable robustness levels for a given cost per objective. This is particularly useful for parameters where little is known about their distribution. However, it can be adapted to include additional knowledge. As observed for better results the search space, i.e. in our case the interval boundaries, should be limited to relevant values. The volume function can also be adjusted. Once it is known sufficiently well, the probability distribution of the RBE parameter could be used instead of the interval length.

Unsurprisingly, in both examples we observed structures where deviation from an RBE of 1.1 was infeasible. In our model this is especially likely to happen if the dose of a structure is tightly controlled by constraints, since we did not add slack variables to the constraints. If critical structures are affected, it could be an indicator to chose a different treatment plan or ensuring control over other RBE-influencing factors such as LET.

We are currently addressing the former issue. Furthermore, providing treatment planners with information on achievable RBE robustness during initial treatment plan selection, and, additionally considering the interval lengths for the RBE parameter of different structures as conflicting goals, leads to a hierarchical bi-level multicriteria problem. We have recently developed a method to solve this special subclass of bi-level problems as a common multicriteria optimization (MCO) problem and thus enable the use of established MCO methods [7].

Moreover, we are working on a method to focus on predefined

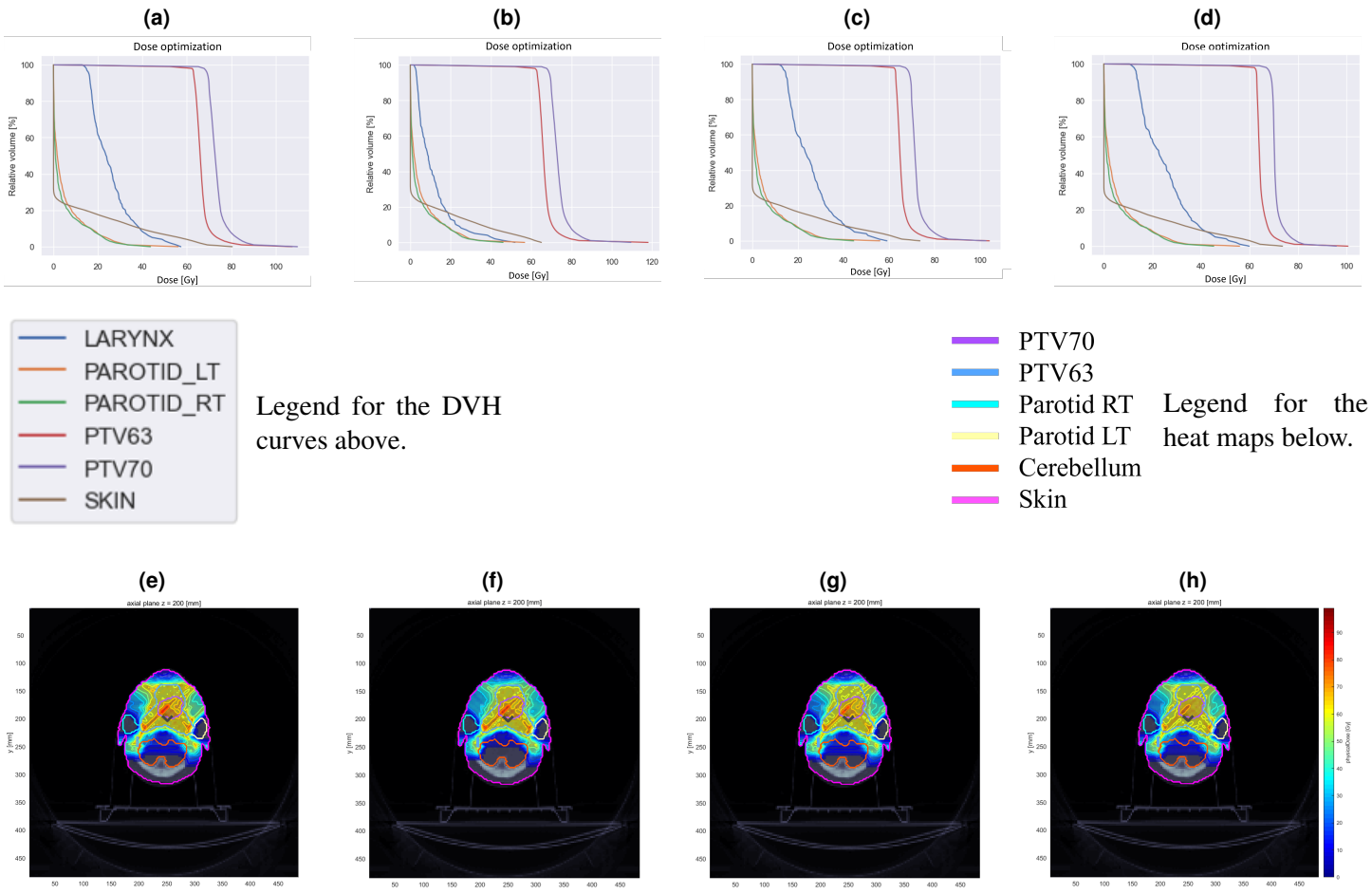


Figure 2: Head and neck case. Dose is weighted with constant factor 1.1. Upper row: DVH curves, lower row: heat maps of dose of an axial slice for a), e) dose optimization; inverse robustness with b), f) relative deviation of 5% and upper interval threshold of 2 c), g) relative deviation of 5%, only optimization for critical structures, d), h) relative deviation of 10% , only optimization for critical structures.

areas where RBE deviations are expected, such as the end of the Bragg peak [1], aiming to maximize the RBE range in these areas only. This approach should prioritize clinically relevant results and optimally utilizes the degrees of freedom in the dose distribution for more critical areas.

5 Conclusion

We introduced inverse robustness to address uncertainties in the RBE parameter for proton therapy planning. This method does not depend on RBE models or other parameters. By setting the price of robustness one is willing to pay, one obtains the largest possible intervals for variation of the RBE parameter that a plan can accommodate. If available, additional knowledge about the RBE can be included and can help to gain more degrees of freedom for the optimization. Thus, we believe that inverse robustness is a promising method for incorporating RBE into robust treatment planning for protons, even in the absence of a universal RBE model.

References

[1] H. Paganetti, A. Niemierko, M. Ancukiewicz, et al. “Relative biological effectiveness (RBE) values for proton beam therapy”. *International journal of radiation oncology, biology, physics* 53.2 (2002), pp. 407–421. DOI: 10.1016/s0360-3016(02)02754-2.

[2] S. J. McMahon. “Proton RBE models: commonalities and differences”. *Physics in medicine and biology* 66.4 (2021). DOI: 10.1088/1361-6560/abda98.

[3] H. Berthold, T. Heller, and T. Seidel. “A unified approach to inverse robust optimization problems”. *Mathematical Methods of Operations Research* (2024). DOI: 10.1007/s00186-023-00844-x.

[4] D. Craft, M. Bangert, T. Long, et al. “Shared data for intensity modulated radiation therapy (IMRT) optimization research: the CORT dataset”. *GigaScience* 3.1 (2014). DOI: 10.1186/2047-217X-3-37.

[5] H.-P. Wieser, E. Cisternas, N. Wahl, et al. “Development of the open-source dose calculation and optimization toolkit matRad”. *Medical physics* 44.6 (2017), pp. 2556–2568. DOI: 10.1002/mp.12251.

[6] J. Pardo-Montero, M. Pombar, A. Gómez-Caamaño, et al. “Variation of the relative biological effectiveness with fractionation in proton therapy: Analysis of prostate cancer response”. *Medical physics* 50.11 (2023), pp. 7304–7312. DOI: 10.1002/mp.16783.

[7] M. Schubert and K. Teichert. “Bi-level multi-criteria optimization to include linear energy transfer into proton treatment planning”. Optimization Online, 2023.

Machine Learning for Dose Prediction in Carbon Ion Therapy

Miriam Schwarze[1], Leo Thomas[1], Ali Kamal[1,2], Pichaya Tappayuthpijarn[1,2] and Hans Rabus[1]

[1]Physikalisch-Technische Bundesanstalt, Berlin, Germany
[2]Hochschule Anhalt, Köthen, Germany

Abstract Fast and accurate dose prediction in carbon ion therapy is crucial for optimizing treatment efficacy and minimizing adverse effects by precisely targeting tumors while sparing healthy tissues. This study explores machine learning (ML) models for accelerated dose prediction in carbon ion therapy. While GANs and transformer models have been previously introduced for dose prediction in other therapies, this work introduces the diffusion model as a novel approach. Trained on high-quality datasets from Geant4 simulations, these models exhibit promising performance results, showing a potential for both precision and speed in prediction dose distributions in mixed radiation fields of therapeutic carbon ions.

1 Introduction

Carbon ion therapy in radiotherapy utilizes the unique physical properties of carbon ions, including their high mass and charge. This enables a precise dose distribution in the tissue with a characteristic Bragg peak. Compared to protons, the Bragg peak of carbon ions has a higher energy deposition at the end of the range, allowing for even more precise local dose application. These characteristics allow for highly precise tumor treatment by focusing the radiation dose specifically on the tumor tissue while sparing the surrounding healthy tissue.

Due to the high precision, carbon ion therapy requires a very accurate dose calculation. Monte Carlo (MC) radiation transport simulations are considered as the most accurate dose calculation approach. However, the long calculation times of MC simulations limit their usefulness in everyday clinical practice, particularly for complex forms of therapy such as carbon ion therapy. The acceleration of MC simulations is therefore the focus of current research efforts, with a decisive emphasis on the application of machine learning (ML) methods.

There are numerous studies on the use of ML to accelerate MC simulations. Particularly fast and accurate results have been shown by approaches which predict the dose distribution in the patient based on patient geometry and beam properties. Generative Adversarial Networks (GANs) and transformer models have been proposed most recently as dose prediction models [1,2]. However, to the best of our knowledge, this methodology has not yet been applied to carbon ion therapy.

This work therefore examines whether this methodology can be transferred to dose calculation in carbon ion therapy. In addition to the application of the known GAN and Transformer models, a diffusion model for dose calculation is developed and compared with the other models.

2 Materials and Methods

High-quality data sets are crucial for machine learning, as they significantly influence the accuracy and performance of the model. To train the models, a dataset was created using the simulation software Geant4 [3]. CT images from the GLIS-RT data set of the TCIA [4], which depicts the head and neck region of various patients, were used as geometry. Six patients were randomly selected from the data set. A monoenergetic carbon ion beam of varying energy between 1250 MeV and 3000MeV and a diameter of 1 mm was simulated. For higher geometric variability, the same geometry was irradiated from different positions and the calculated dose and density matrices were also rotated in steps of 90°.

The GAN approach was published in [1]. The authors used a conditional GAN for dose prediction in microbeam radiation therapy. The GAN model consists of a generator and a discriminator. The generator produces the dose matrix based on a density matrix of the geometry and a dose matrix of the corresponding beam configuration in homogeneous water. The discriminator is trained to distinguish between real and generated data. Through this continuous rivalry, the generator constantly improves to produce more authentic dose maps, while the discriminator develops its detection capability.

The transformer model for dose prediction was proposed in [2] in the context of proton therapy. Transformer models can model complex dependencies in long sequences through self-attentive layers by utilizing attention mechanisms. In this approach the density and dose matrices are divided into a sequence of 2 dimensional slices along the beam direction. The transformer architecture is combined with a Convolutional Neural Network (CNN) for extracting information from the inputted density matrix and decoding the output dose matrix. The encoded density matrix is combined before the transformer module with a vector representing the energy of the carbon ions.

The Diffusion model approach disseminates information from the simulated dose matrix through a step-by-step stochastic diffusion process. A U-Net, conditioned on the density matrix and beam information, is trained to predict the noise added in one time step. The dose matrix is reconstructed by subtracting the predicted noise from the noisy image. After training, the U-Net is able to predict the dose distribution starting with a random noisy image and the conditional information.

Training involves both training and validation datasets. For testing, a distinct dataset is uniformly reserved across all geometries and energies. Additionally, to assess generalization, a complete brain geometry not encountered during the training phase was included in the testing process.

All models were optimized with respect to their hyperparameters.

3 Results

Table 1 presents the performance measurements for the optimized GAN model, including both the validation dataset and the two test datasets. The performance metrics include the averaged MSE per voxel, along with the average 1% and 3% passing rates.

To examine the network's generalization across various geometries, Figure 1 illustrates the MSE averaged across energies for different source positions in the unseen geometry.

The time required for one dose prediction using Geant4 ranges from 1 to 3 hours, depending on the beam energy and geometry factors. For the trained GAN model, the average computational time per prediction is 0.012 seconds on a CPU and 0.009 seconds on a GPU.

The Transformer and Diffusion Models are still under development.

Table 1: Average performance metrics of the optimized GAN model for the validation- and test dataset and for the held-out geometry.

Performance Metric	Validation Dataset	Test Dataset	Unseen Geometry
MSE [10^{-5}]	1.75 ± 0.44	1.53 ± 0.26	106.90 ± 9.76
1% passing rate [%]	99.48 ± 0.06	99.47 ± 0.05	99.60 ± 0.03
3 % passing rate [%]	99.80 ± 0.03	99.81 ± 0.03	99.71 ± 0.02

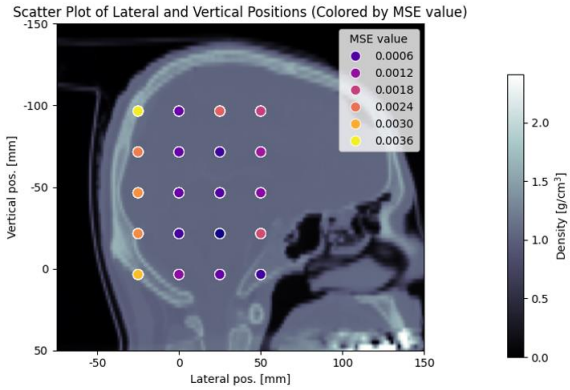


Figure 1: Average MSE at different source positions of the held-out test geometry.

4 Discussion

The GAN architecture demonstrates promising performance for fast dose prediction through machine learning. Dose prediction using a trained GAN model is five orders of magnitude faster than a Monte Carlo simulation with Geant4. In all cases, an average passing rate greater than 99% was achieved. Challenges arise in specific geometries, such as when the beam is close to the skull along most of its path. The elevated MSE values observed for the unseen geometry compared to the test dataset indicate the necessity of incorporating a broader range of diverse geometries, each characterized by its unique dimensions, into the training process.

5 Conclusion

Three different machine learning models for dose prediction in carbon ion therapy are explored and compared with each other. The available results for the GAN model showcase high passing rates exceeding 99% and an ability to generalize to the geometry. The results demonstrate that accurate and fast dose prediction through machine learning is viable even within the context of mixed radiation fields, such as in carbon ion therapy.

References

[1] Mentzel, Florian, et al. "Fast and accurate dose predictions for novel radiotherapy treatments in heterogeneous phantoms using conditional 3D-UNet generative adversarial networks." Medical Physics 49.5 (2022): 3389-3404. DOI : 10.1002/mp.15555.

[2] Pastor-Serrano, Oscar, and Zoltán Perkó. "Millisecond speed deep learning based proton dose calculation with Monte Carlo accuracy." Physics in Medicine & Biology 67.10 (2022): 105006. DOI: 10.1088/1361-6560/ac692e.

[3] Agostinelli, Sea, et al. "GEANT4—a simulation toolkit." Nuclear instruments and methods in physics research section A: Accelerators, Spectrometers, Detectors and Associated Equipment 506.3 (2003): 250-303. DOI: 10.1016/S0168-9002(03)01368-8.

[4] Shusharina, N. B. T., and T. Bortfeld. "Glioma Image Segmentation for Radiotherapy: RT targets, barriers to cancer spread, and organs at risk [Data set]." The Cancer Imaging Archive (2021). DOI: 10.1007/978-3-030-71827-5_1.

Investigating Associations Between Radiomics Features and Biomarkers in an Orthotopic Non-Small Cell Lung Cancer Mouse Model

Gabriel Giampa[1,2], Ranjan Yadav[1,2], Li Ming Wang[2,3], John Kildea[1,2], Martin Vallières[4], and Norma Ybarra[1,2,3]

[1]Medical Physics Unit, Gerald Bronfman Department of Oncology, McGill University, Montreal, Canada
[2]Cancer Research Program, Research Institute, McGill University Health Centre, Montreal, Canada
[3]Division of Experimental Medicine, Faculty of Medicine and Health Sciences, McGill University, Montreal, Canada
[4]Department of Computer Science, Faculty of Science, Université de Sherbrooke, Sherbrooke, Canada

Abstract The aim of this study is to investigate whether there exists a correlation between radiomics features and histological expression of biomarkers in an orthotopic non-small cell lung cancer (NSCLC) mouse model. Tumours were induced in thirty-two athymic male mice by implanting a human NSCLC line in their lungs. The mice were split into four treatment groups. Radiomics features were extracted from cone-beam computed tomography images taken two to three times per week over the course of two weeks of treatment. The mice were then euthanized and their tumours are currently undergoing immunohistochemical staining to detect several biomarkers. Thus far, many extracted features show significant variations across treatment groups, even before the different treatments begin. Further, many features show strong linear correlations with other features. We describe the experimental and analysis protocol and outline our efforts to account for these findings, including normalization of features by a baseline value and using feature selection protocols.

1 Introduction

Radiomics is an expanding branch of research that concerns itself with the extraction and analysis of quantitative characteristics, or features, from medical images [1]. Radiomics lends itself well to the study of solid cancer cases for several reasons. For one, medical images are ubiquitous to modern cancer treatment, and so a radiomics analysis can serve to increase the amount of information available to a patient's treatment team without demanding additional examinations. Radiomics analyses can also augment the information found within these medical images as certain features, such as those describing the texture of a tumour, are not meaningfully appreciable by the human eye on its own [2].

Radiomics, and medical imaging in general, are usually limited to describing macroscopic tumour characteristics. Biopsies may be performed to investigate microscopic characteristics, such as the tumour's microenvironment and the expression of histological biomarkers, but biopsies are invasive and can only provide information on the section of the tumour that is excised.

The goal of this project is to investigate whether there exists a correlation between radiomics features and histological biomarker expression in an orthotopic non-small cell lung cancer mouse model. If such a correlation is established, radiomics analyses of medical images could present a non-invasive way to probe a solid tumour for microscopic biomarker information. This analysis also presents an avenue for evaluating a tumour's response to treatment if the expressions of certain biomarkers are clinically relevant tumour response parameters.

2 Materials and Methods

2.1 The Mouse Model

For this study, orthotopic implantation of tumour cells was performed on 32 male athymic mice, by injecting H1299 non-small cell lung cancer (NSCLC) cells in either the left or right lung. Tumour development was monitored via cone-beam computed tomography (CBCT) and when visible tumours developed, the mice were randomly divided into four treatment cohorts: a control (CTRL), a group to receive radiation treatment (RT), a group to receive injections of L-carnosine (CAR), and a group to receive a combination of L-carnosine and radiotherapy (CAR+RT). L-carnosine is a naturally occurring dipeptide with anti-cancer properties [3] that has been shown to exhibit a radiosensitizing effect in lung cancer cells [4]. On Day 0, the mice in the RT and CAR+RT groups were treated with a single dose of 20 Gy. The mice were followed for at least two weeks thereafter, and underwent CBCT imaging two to three times per week. Finally, the mice were euthanized. Their lungs were resected and processed for immunochemistry.

2.2 Immunohistochemistry

Tissue sections from the tumour-bearing lungs are currently undergoing immunochemistry analysis for the detection of several biomarkers. Among these are Ki-67, a protein associated with cellular proliferation [5]; HIF-1α, a protein associated with hypoxia [6]; and TNF-α, an inflammatory cytokine associated with cellular immune response [7]. These biomarkers represent clinically relevant information about how each tumour responded to its treatment.

2.3 Radiomics Features

The CBCT images for each mouse were manually segmented at three time points: Day 0 (the day of radiation treatment),

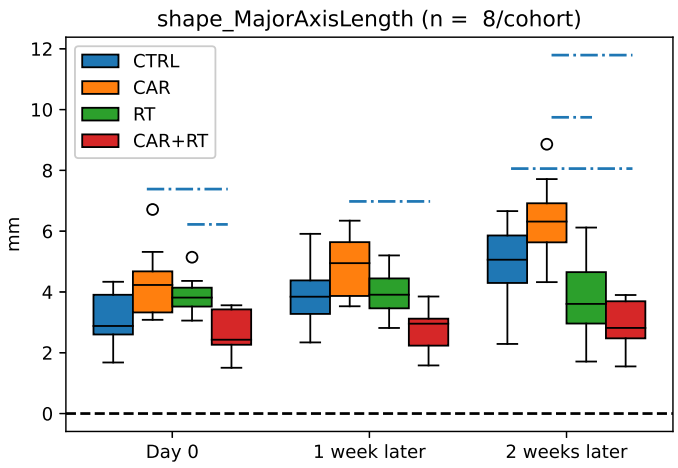


Figure 1: The major axis length is a radiomics feature in the shape family that describes the largest axis length of the smallest ellipsoid required to totally enclose the ROI.

one week thereafter, and two weeks thereafter. The segmentation consisted of defining a binary region of interest (ROI) in each image enclosing the tumour as conformally as possible. Then, using the open-source pyRadiomics package, 107 radiomics features were calculated from the pixels within each ROI and extracted [8]. The features can be categorized into three broad families: first-order features, which create a histogram of the pixel values within the ROI and then calculate statistics on the histogram; shape features, which describe the geometry of the ROI; and higher-order features, which create matrices based on the interrelationship between a pixel's values and the values of its neighbours to describe the texture of the ROI.

When comparing radiomics features across cohorts, we used a Kruskal-Wallis non-parametric test with a Dunn's post hoc test to identify statistically significant inter-cohort variation.

3 Results

3.1 Absolute Features

Many features show notable variation within and across treatment cohorts, even on Day 0. For example, Figure 1 shows the major axis length, a shape feature, of the tumours at the three contoured time points. Even on Day 0, there is large variation across the treatment cohorts, so much so that the major axis length of the tumours in the CAR+RT group deviates significantly from those of the tumours in the CAR and RT groups.

3.2 Feature Correlation

Many of the extracted features appear to be strongly linearly correlated to other features. For example, Figure 2 presents a clustermap of gray level co-occurrence matrix (glcm) features extracted from the tumours in the RT group at the 2 week time point. These are higher-order features that describe how

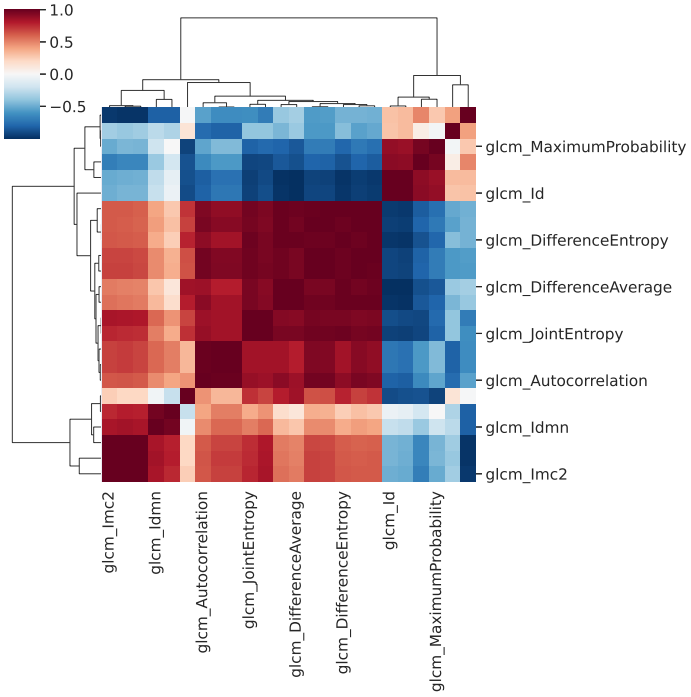


Figure 2: A clustermap representing correlations between gray level co-occurrence matrix (glcm) features in the RT group at the 2 week time point shows the linear Pearson R correlation between each pair of glcm features as a shaded box on the grid. Dark red boxes represent strongly positive correlations (R = 1), and dark blue boxes represent strongly negative correlations (R = -1). For visual clarity, every third feature is labelled by name.

often pairs of grayscale values are found near each other within the ROI. Very few boxes are light in color, indicating that the vast majority of glcm features in the RT group at the 2 week time point are strongly linearly correlated to several others.

4 Discussion

The inter- and intra-cohort variation of features at Day 0 is to be expected as there is variability both in the cell implantation process and in individual response and tumour development. However, this variation poses an obstacle when one wishes to compare the efficacy of one treatment versus another. To fairly assess the effect of each treatment on extracted features, the values of each feature will be normalized by a baseline value, such as the values on Day 0.

The strong correlation between features is also to be expected as many radiomics feature definitions are inherently sensitive to others. Further, when assessing changes to features in response to treatment, the underlying biological and physical processes are less numerous than the features extracted. For example, Figure 1 shows that after two weeks of treatment, the tumours in the CAR+RT group had significantly smaller major axis lengths than those in the CTRL and CAR groups. That radiation kills cells and thus retards growth in tumours is well known and is, in fact, the main reason radiotherapy is employed. These results are obviously linked to the cell

killing effect of radiation and one could easily predict that other shape-based features, like total tumour volume, will follow a similar trend. Indeed, an important step in employing radiomics is feature selection, whereby features that provide superfluous or redundant information are ignored from further analysis [9]. By using clustermaps, a representative feature from each cluster can be chosen to reduce the total number of features that require analysis.

5 Conclusion

We extracted over 100 features from non-small cell lung cancer tumours implanted orthotopically in an athymic mouse model and are in the process of comparing them. Thus far, we have shown that even at the beginning of the treatment period, the values of features within and across cohorts can be notably different. Further, we have shown that many features show strong linear correlations with other features, indicating that the number of features containing unique and clinically relevant information is likely quite small. We plan on comparing these extracted features to the expression of several biomarkers in the coming months.

6 Funding Acknowledgements

This research is being supported by the Quebec Ministère de la Santé et des Services sociaux, by the Canadian Graduate Scholarship (Master's) through the Canadian Institutes of Health Research, and by the CREATE Responsible Health and Healthcare Data Science (SDRDS) grant of the Natural Sciences and Engineering Research Council.

References

[1] M. E. Mayerhoefer, A. Materka, G. Langs, et al. "Introduction to Radiomics". *Journal of Nuclear Medicine* 61.4 (Feb. 2020), pp. 488–495. DOI: 10.2967/jnumed.118.222893.

[2] M. Avanzo, J. Stancanello, and I. El Naqa. "Beyond imaging: The promise of radiomics". *Physica Medica* 38 (June 2017), pp. 122–139. DOI: 10.1016/j.ejmp.2017.05.071.

[3] I. Jukić, N. Kolobarić, A. Stupin, et al. "Carnosine, Small but Mighty—Prospect of Use as Functional Ingredient for Functional Food Formulation". *Antioxidants* 10.7 (June 2021), p. 1037. DOI: 10.3390/antiox10071037.

[4] N. Ybarra and J. Seuntjens. "Radio-selective effects of a natural occurring muscle-derived dipeptide in A549 and normal cell lines". *Scientific Reports* 9.1 (Aug. 2019). DOI: 10.1038/s41598-019-47944-5.

[5] T. Scholzen and J. Gerdes. "The Ki-67 protein: From the known and the unknown". *Journal of Cellular Physiology* 182.3 (Mar. 2000), pp. 311–322. DOI: 10.1002/(sici)1097-4652(200003)182:3<311::aid-jcp1>3.0.co;2-9.

[6] J. E. Ziello, I. S. Jovin, and Y. Huang. "Hypoxia-Inducible Factor (HIF)-1 regulatory pathway and its potential for therapeutic intervention in malignancy and ischemia". en. *Yale J. Biol. Med.* 80.2 (June 2007), pp. 51–60.

[7] H. T. Idriss and J. H. Naismith. "TNF alpha and the TNF receptor superfamily: structure-function relationship(s)". en. *Microsc. Res. Tech.* 50.3 (Aug. 2000), pp. 184–195.

[8] J. J. van Griethuysen, A. Fedorov, C. Parmar, et al. "Computational Radiomics System to Decode the Radiographic Phenotype". *Cancer Research* 77.21 (Oct. 2017), e104–e107. DOI: 10.1158/0008-5472.can-17-0339.

[9] W. Zhang, Y. Guo, and Q. Jin. "Radiomics and Its Feature Selection: A Review". *Symmetry* 15.10 (Sept. 2023), p. 1834. DOI: 10.3390/sym15101834.

Towards patient-specific computational models of microvasculature to investigate radiation toxicity

Sophie Materne[1,2], Luca Possenti[1,*], Alessandro Cicchetti[1], Francesco Pisani[1], Alessandra Catalano[1], Piermario Vitullo[2], Paolo Zunino[2], and Tiziana Rancati[1]

[1]Data Science Unit, Fondazione IRCCS Istituto Nazionale dei Tumori, Milan, Italy
[2]Department of Mathematics, Politecnico di Milano, Milan, Italy
[*]Presenting/corresponding author

Abstract The microvasculature plays an essential role in the microenvironment, supplying nutrients, drugs and facilitating immune cell trafficking. In previous studies, microvascular health status was predictive of radiotherapy toxicities in different cohorts of patients. In this study, we present a computational workflow that generates synthetic microvascular networks leveraging data from a sublingual microscope collected in a cohort of 63 Head and Neck Cancer patients. The workflow facilitates the development of personalized computational models to analyze the role of the microvasculature in radiation toxicity. The synthetic networks replicate the microvascular density and vessel diameter of the patients, enabling mechanistic simulations of microvascular flow and drug/nutrient delivery. The results demonstrate the capability of the computational framework to generate relevant microvascular scenarios, with blood flow parameters aligning with physiological ranges. This model will serve as a valuable tool for mechanistically evaluating how the microvasculature shapes the microenvironment, contributing to current models on RT toxicity in diverse patient cohorts.

1 Introduction

The microvasculature plays an essential role in the microenvironment, delivering nutrients to tissues and facilitating drug delivery and immune cell trafficking. For these reasons, the microvasculature significantly influences cancer treatments, including chemotherapy, immunotherapy, and even radiotherapy (RT) through oxygen delivery and the potential presence of hypoxia within the tumor. The central role of the microvasculature in cancer treatment extends beyond the tumor to include healthy tissue toxicity. Acute and late toxicity define the therapeutic window, impacting therapeutic efficacy and quality of life for cancer patients.

In the context of radiotherapy, several trials have associated an increased risk of normal tissue toxicity with conditions that negatively affect the microvascular system, such as hypertension, diabetes, smoking, alcohol abuse, and obesity [1–3]. In a previous analysis, we demonstrated the predictive role of microvasculature healthiness for acute toxicity after RT in breast, prostate, and Head and Neck Cancer (HNC) patients, using a sublingual microscope [4]. Specifically, impaired microvasculature was associated with an elevated probability of toxicity. The functional information provided by the sublingual microscope can enhance the personalization of predictive models by tailoring them to the individual patient's functional status.

On the other hand, computational models can describe the state of the microvascular environment based on a set of parameters, including physical quantities (e.g., red blood cell (RBC) concentration, oxygen/drug concentration, permeability of the vascular wall) and microvascular morphology [5, 6]. Through computational modeling, we can assess the impact of these data on the microenvironment, such as interstitial pressure, oxygen availability, and drug/nutrient delivery [7]. This work aims to integrate patient data with computational modeling. To achieve this, we collected data from patients using the sublingual microscope to create a set of synthetic computational domains. In this study, our primary focus was on microvascular morphology, including microvascular radius and density. Using these data, we constructed computational microvascular networks. Then we simulated blood flow, describing parameters such as blood velocity, RBC transport, and blood pressure.

2 Methods

This work was carried out in three main steps: (i) collection of patients' data; (ii) generation of synthetic microvascular networks that match radius and density data; (iii) computational simulation of blood flow in the synthetic microvasculature.

2.1 Patients' data

We enrolled 63 HNC patients who received a prescribed total tumor dose ranging from 60 to 70 Gy, with 1.7-2.12 Gy/fraction. Using the GlycoCheck system, we measured microvasculature parameters in the sublingual region, including capillary density, capillary radius distribution, and RBC velocity. These data were obtained non-invasively using a handheld sublingual camera that recorded blood flow in the capillaries prior to RT treatment (Figure 1).

2.2 Building synthetic microvascular networks

As the initial step for the patient-specific microvascular simulation, we constructed computational domains designed to replicate the dimensions of the images obtained by the sublingual microscope, utilizing a $700 \times 950, \mu m$ rectangle. Within this domain, we created synthetic microvasculature,

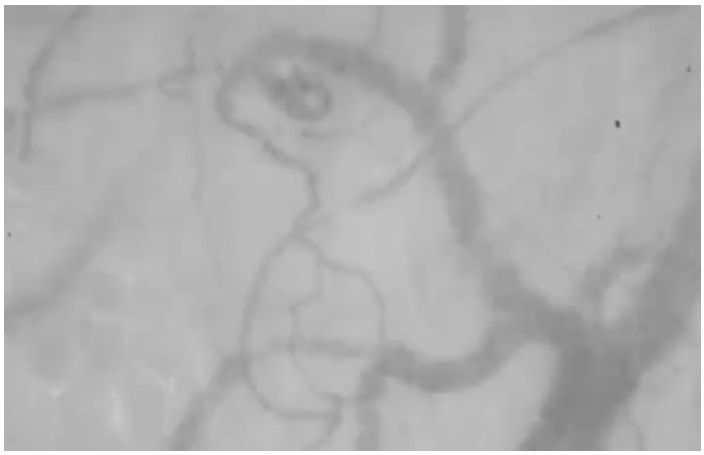


Figure 1: Frame extracted from the sublingual microscope recordings.

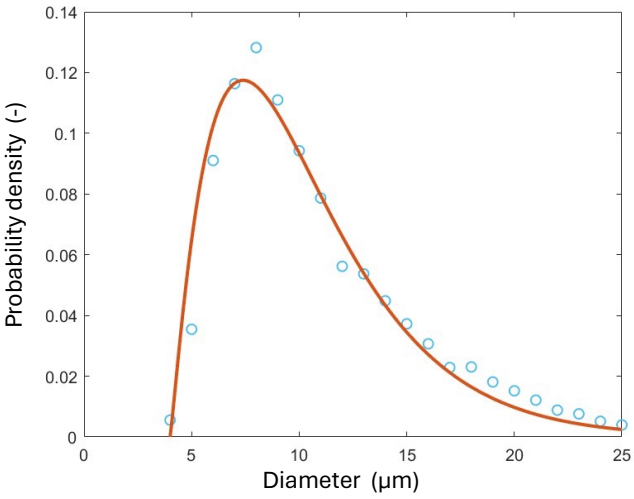


Figure 2: Microvascular diameter distribution from patients (dots) and its approximation with the γ distribution.

employing hierarchical two-level networks to mimic the data obtained from the microscope (Figure 1). The first level comprises a few larger vessels present in the images, while the second layer consists of smaller vessels originating from the first layer. Each of these layers was generated using a Voronoi-based approach, chosen for its bio-mimetic features, as described in previous works [6, 7].

In summary, we randomly generated a defined number of seeds and subsequently used them to perform Voronoi tessellation. Each edge represents a vessel, and the entire tessellation constitutes the microvascular network. The density of the vessel depends on the initial distribution of the seeds. The number of Voronoi seeds for the first layer is set to 4. On the contrary, the number of seeds for the second layer was adjusted to match the microvascular density. The sublingual microscope calculates the vascular density by considering a subset of vessels. Lee and colleagues [8] measured the total capillary density in a similar setup with a cohort of 6673 individuals, estimating it to be 43146 $\mu m/mm^2$.

We applied a correction based on this value to account for the total capillary density:

$$Total\,density = 43146\,\mu m/mm^2 \times \frac{desired\,density}{median\,density}. \quad (1)$$

Following this correction, we set the radius for each vessel to complete the morphological definition of the synthetic microvascular network. This step was performed on the second layer of the vessels, mirroring the approach taken by the sublingual microscope. Initially, we constructed a probability function based on radius data from our HNC cohort (Figure 2 - γ distribution with parameters = 2.15 and 2.95). Subsequently, for each vessel, we assigned the radius such that: (i) the resulting microvascular network mimics the length-weighted distribution obtained from patients; (ii) the diameter is strictly smaller than half the length of the considered capillary, emphasizing the length as the dominant dimension.

2.3 Simulation of microvascular flow

After the network generation, we simulated the blood flow within the synthetic microvasculature considering the presence of RBCs. Briefly, we calculate the blood pressure for each node and the flow rate for each branch. This involved solving a Poiseuille-like problem (Eq. 2), coupled with the mass conservation equation (Eq. 3):

$$p_{i,in} - p_{i,out} = R_i Q_i \qquad with\, i = 1,...,n_{branches} \quad (2)$$

$$\sum_{i \in j} Q_i = 0 \qquad with\, j = 1,...,m_{nodes}. \quad (3)$$

The resistance term is defined by including the Fahraeus-Lindqvist effect:

$$R_i = \frac{128\,\mu(H_i,d_i)\,L_i}{\pi\,d_i^4}. \quad (4)$$

RBC flow is determined considering mass conservation at each junction and neglecting potential fluid filtration through the capillary wall (Eq. 5). For bifurcations, the model accounts for the uneven partition of RBCs when flowing through them, known as the Zweifach-Fung effect (Eq. 6, 7):

$$\sum_{i \in j} Q_i H_i = 0 \qquad with\, j = 1,...,m_{nodes}; \quad (5)$$

$$H_\alpha = F_{QE\alpha}\,H_{father}\,\frac{Q_{father}}{Q_\alpha} \quad (6)$$

$$H_\beta = (1 - F_{QE\alpha})\,H_{father}\,\frac{Q_{father}}{Q_\beta} \quad (7)$$

An iterative method was used to solve the problem as a result of the interdependence of the RBC distribution on flow and vice versa (via viscosity). Further details on the model, including the definitions of $\mu(H_i,d_i)$ and $F_{QE\alpha}$, as well as the iterative solution strategy, can be found in [6]. The problem is then complemented with boundary conditions, defined as the inlet RBC concentration ($H_{in} = 0.1$, [9]), and a pressure

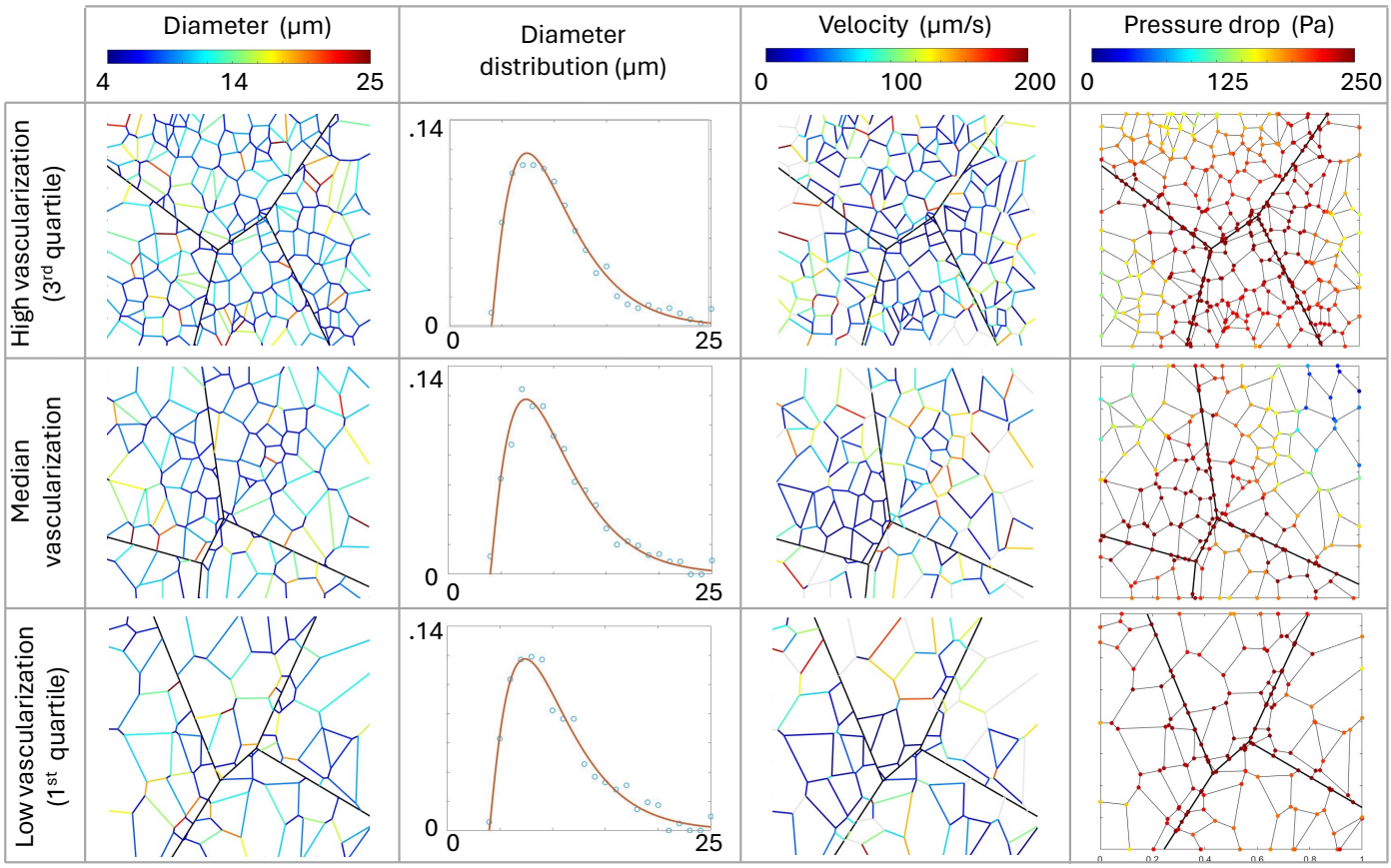


Figure 3: Synthetic microvascular networks based on the microvascular density data (Low = first quartile, Median, and High = third quartile. We report the diameter of the vessel within the network (first column), its distribution compared to the theoretical one (second column), and the computational estimation of the flow, quantified as velocity (third column) and pressure drop (fourth column). Blood velocity greater than 200 μm is indicated in grey.

drop ($\simeq$ 2 mmHg). The exact outlet pressure was defined considering the inlet-outlet distance, adjusting the pressure drop based on the shortest inlet-outlet path length. The inlet nodes are identified as the intersections of the first and second layers of the microvasculature, while outlet nodes are identified as the intersections of the second layer with the outer boundary of the domain.

3 Results

The results presented here are based on radius and density values extracted from the HNC cohort, focusing on the generation and calculation of microvascular networks. Three distinct cases are outlined, corresponding to the 1st, 2nd (median), and the 3rd quartile of capillary density in our population (Figure 3).

3.1 Microvascular networks

The morphology of the synthetic microvascular network is illustrated in the first column (Figure 3), delineating the first layer of the microvasculature (larger vessels, $d \simeq, 100\mu m$) in black and the second layer color-coded based on the diameter value ($4\mu m \leq d \leq 25\mu m$). By reducing the target density and, consequently, the number of seeds, we successfully obtained networks with lower total length and a reduced number of segments. The second column presents the diameter distribution, along with an approximated sample distribution, as a useful reference. These two types of data prove the generation of microvascular networks based on a target microvascular density and diameter distribution.

3.2 Blood velocity and pressures

Blood flow results are detailed as blood velocity (third column) and pressure (fourth column). First, these results demonstrate that the synthetic microvascular network is well-connected, and the computational workflow generates relevant scenarios. This observation is further supported by the velocity and pressure drop results, both falling within a physiological range ($v \simeq 50 - 200\,\mu m/s$) and Δp a few mmHg) [6, 10]. The velocity results also align with the same order of magnitude of the data collected using the sublingual microscope (Figure 4).

4 Discussion

In this study, we collected data acquired from a sublingual microscope in a HNC cohort consisting of 63 patients. Using

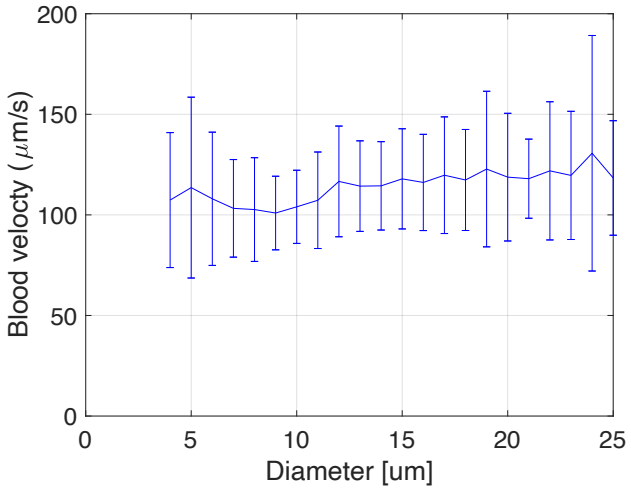


Figure 4: Blood velocity distribution obtained from patients' data showing the average values with standard deviation.

these data, we constructed a set of synthetic microvascular networks. Our results are based on population-based statistics, specifically focusing on microvascular density and vessel diameters. As a remark, our workflow is not restricted to population-based target values; rather, it can seamlessly integrate patient-specific target values. This adaptability enables the development of personalized computational models of the microenvironment, facilitating the analysis of the role of the microvasculature in radiation toxicity. In addition, pressure boundary conditions can be customized to replicate individual blood velocity measurements, which are also available for individual patients.

The flexibility of the presented workflow extends to modifications that involve employing different network generation strategies. For instance, Voronoi tessellation can be replaced or supplemented with alternative methods. Such modifications enable the creation of non-uniformly distributed networks, providing a more accurate representation of possible pathological states.

The proposed model serves as a valuable computational framework for a mechanistic evaluation of how the microvasculature shapes the microenvironment. These mechanistic insights contribute to analyses on the role of microvasculature in radiation toxicity in different cohorts of patients [4].

5 Conclusion

We introduced a computational workflow to generate synthetic microvascular networks based on a cohort of 63 HNC patients. These networks faithfully mimic the microvascular density and vessel diameter of the patients, enabling mechanistic simulations of microvascular flow and flow-mediated drug/nutrient delivery. The integration of this computational framework and patient data provides a valuable tool for unraveling the role of microvasculature in RT toxicity.

6 Acknowledgments

This work was supported by the AIRC Investigator Grant, no. IG21479 (TZ is the PI of the study). Paolo Zunino acknowledges the support of the grant MUR PRIN 2022 No. 2022WKWZA8 "Immersed methods for multiscale and multiphysics problems (IMMEDIATE)" part of the Next Generation EU program. Piermario Vitullo and Paolo Zunino are members of Gruppo Nazionale per il Calcolo Scientifico (GNCS) of Istituto Nazionale di Alta Matematica (INdAM). The present research is part of the activities of project Dipartimento di Eccellenza 2023-2027, Department of Mathematics, Politecnico di Milano.

References

[1] M. D. Santis, F. Bonfantini, F. D. Salvo, et al. "Factors influencing acute and late toxicity in the era of adjuvant hypofractionated breast radiotherapy". *The Breast* 29 (Oct. 2016), pp. 90–95. DOI: 10.1016/j.breast.2016.07.013.

[2] F. Palorini, T. Rancati, C. Cozzarini, et al. "Multi-variable models of large International Prostate Symptom Score worsening at the end of therapy in prostate cancer radiotherapy". *Radiotherapy and Oncology* 118 (1 Jan. 2016), pp. 92–98. DOI: 10.1016/j.radonc.2015.11.036.

[3] A. Cavallo, T. Rancati, A. Cicchetti, et al. "EP-1594: Development of multivariable models for acute toxicities in nasopharyngeal cancer radiotherapy". *Radiotherapy and Oncology* 123 (May 2017), S858–S859. DOI: 10.1016/S0167-8140(17)32029-7.

[4] T. Rancati, F. Badenchini, E. L. Rocca, et al. "The Microvascular Health Status Predicts Radio-Induced Acute Toxicities in Breast, Prostate and Head and Neck Cancer Patients". *International Journal of Radiation Oncology*Biology*Physics* 117 (2 Oct. 2023), S32–S33. DOI: 10.1016/j.ijrobp.2023.06.297.

[5] J. T. Celaya-Alcala, G. V. Lee, A. F. Smith, et al. "Simulation of oxygen transport and estimation of tissue perfusion in extensive microvascular networks: Application to cerebral cortex". *Journal of Cerebral Blood Flow & Metabolism* 41 (3 Mar. 2021), pp. 656–669. DOI: 10.1177/0271678X20927100.

[6] L. Possenti, S. di Gregorio, F. M. Gerosa, et al. "A computational model for microcirculation including Fahraeus-Lindqvist effect, plasma skimming and fluid exchange with the tissue interstitium". *International Journal for Numerical Methods in Biomedical Engineering* 35 (3 Mar. 2019), e3165. DOI: 10.1002/cnm.3165.

[7] P. Vitullo, L. Cicci, L. Possenti, et al. "Sensitivity analysis of a multi-physics model for the vascular microenvironment". *International Journal for Numerical Methods in Biomedical Engineering* (July 2023). DOI: 10.1002/cnm.3752.

[8] D. H. Lee, M. J. C. Dane, B. M. van den Berg, et al. "Deeper Penetration of Erythrocytes into the Endothelial Glycocalyx Is Associated with Impaired Microvascular Perfusion". *PLoS ONE* 9 (5 May 2014), e96477. DOI: 10.1371/journal.pone.0096477.

[9] F. Clavica, A. Homsy, L. Jeandupeux, et al. "Red blood cell phase separation in symmetric and asymmetric microchannel networks: Effect of capillary dilation and inflow velocity". *Scientific Reports* 6 (October 2016), pp. 1–12. DOI: 10.1038/srep36763.

[10] M. W. Dewhirst and T. W. Secomb. "Transport of drugs from blood vessels to tumour tissue". *Nature Reviews Cancer* 17 (12 Dec. 2017), pp. 738–750. DOI: 10.1038/nrc.2017.93.

Modeling the geometric distortion and fan beam imaging of an on-rails computed tomography scanner for 2D/3D registration

Giovanni Fattori[1], Riccardo Via[1], Antony John Lomax[1,2] and Sairos Safai[1]

[1]Center for Proton Therapy, Paul Scherrer Institut, Villigen, Switzerland [2]Department of Physics, ETH Zürich, Switzerland

Abstract Intraoperative CT scanners are increasingly present in high-precision radiotherapy units, enabling the acquisition of tomographic images of the treatment geometry. While providing high quality images, geometric uncertainties most often preclude the use of on-rails scanners as imagers for patient position verification. We studied the distortion of volumetric images and topograms, while modelling the fan-beam projective geometry for the calculation of digitally reconstructed topogram (DRT) images in the context of 2D/3D registration. Distortion models parameterized by comparison with laser tracker measurements enabled image-based localization with errors below 0.3 mm on individual directions in 3D. Finally, in a phantom study we evaluated the accuracy of a new 2D/3D positioning approach from orthogonal CT topogram image pairs, reporting similar accuracy and in conformity to clinical practice guidelines.

1 Introduction

Originally developed for interventional surgery, on-rails computed tomography (CT) scanners are nowadays gaining interest in high-precision radiotherapy. Indeed, their higher image quality makes them an attractive alternative to cone beam systems in the context of adaptive radiotherapy and proton therapy in particular, where the attenuation of the CT scan is crucial for the estimation of the beam range [1]. By and large, these scanners are combined with robotic positioning systems that allow intraoperative volumetric imaging of the patient on the surgical table in the exact treatment geometry. Despite the presence of several units in proton therapy facilities around the world, typically on-rail scanner use is limited to CT simulation and acquisition of control images, rather than for actual image guidance. The repurposing of on-rails technology as a patient positioning verification imager is not common and requires careful characterization of the geometric uncertainties associated with the moving gantry and fan beam geometry.

A major issue is that the reconstructed images are distorted in the longitudinal direction caused by non-linearity in the scanner motion during imaging and inaccurate readings of the gantry position encoders. Modelling the effect of such mechanical complexity on the reconstructed image coordinates and the consequent mapping of the CT space into real-world physical units is necessary for the use of on-rail scanners as for verification imaging in high-precision radiotherapy procedures.

On the other hand, the advantages of having a conventional CT in the treatment room go beyond image quality, and include the ability to reach long scan ranges, with images one meter or more in length, that are particularly useful for verifying extremities or the whole cranio-spinal axis. This applies not only to volumetric tomography but also to topogram images, which can be acquired in pairs at orthogonal angles and, provided that the projective geometry is known, be used for patient alignment in a 2D/3D registration approach.

We present the calibration and fan beam geometry model of an in-room, on-rails CT scanner (Sensation Open, Siemens Healthineers, Erlangen, Germany) installed for proton therapy at our institute. Firstly, we address the longitudinal geometric distortion of 3DCT imaging coordinates along the direction of scanner motion. This is followed by the topogram imaging characterization, for which we also introduce and calibrate a model of the projective imaging geometry for the computation of digitally reconstructed topogram (DRT) images. Finally, in a phantom study, we evaluate the accuracy of a novel 2D/3D positioning approach that makes use of CT topogram image pairs.

2 Materials and Methods

Modeling the geometric distortion of CT volumes

Repeated 3DCT images of a geometric phantom (Figure 1) have been acquired centering the scan range longitudinally at three different distances from the scanner parking position. The phantom holds 12 ceramic spheres of 2 cm diameter, facilitating their accurate measurement with a reference laser tracker equipment (Leica AT901 LR, Leica Camera AG, Wetzlar, Germany) and automatic localization in 3DCT images at 0.98x0.98 mm in-plane resolution and 1 mm slice thickness [2]. Data of each scan range have been processed individually to optimize the root mean square error (RMSE) of a point-based registration between the centers of spheres in the room (world) coordinates with respect to their corresponding imaging positions, including a linear correction along IEC-Y with scaling (s) and bias (b) free parameters:

$$P'_{Y_{IEC}} = P_{Y_{IEC}} + (s * P_{Y_{IEC}}) + b \qquad \text{(Eq. 1)}$$

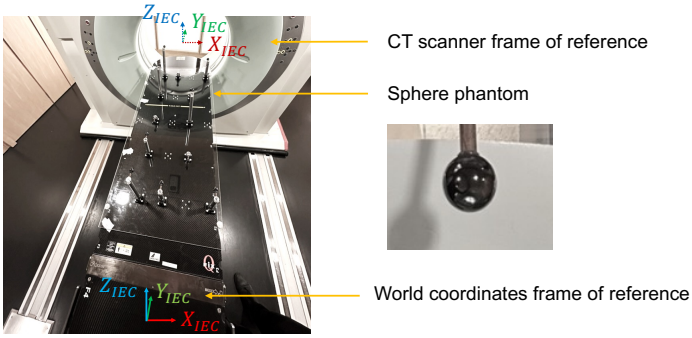


Figure 1: geometric phantom used for assessing the geometrical trueness of 3DCT imaging coordinates and reference systems involved.

A simpler version of this phantom with just two sphere points has been acquired for validation purposes, to compare the laser measurements with their undistorted image coordinates.

A model for fan beam topogram imaging

Modeling the fan beam projection geometry starts from the definition of the transform that maps the fixed CT scanner frame of reference {CT} and the digital model of the rotating (α) X-ray source {S} (Figure 2). At a source-to-image (SID) distance along the negative direction of the X-ray source reference system, we find the center of the DRT imaging plane, which in an ideal geometry is aligned to the CT gantry center of rotation. This assumption can be altered by defining the center of projection ($\boldsymbol{c}$), that offsets the digital model of source and panel.

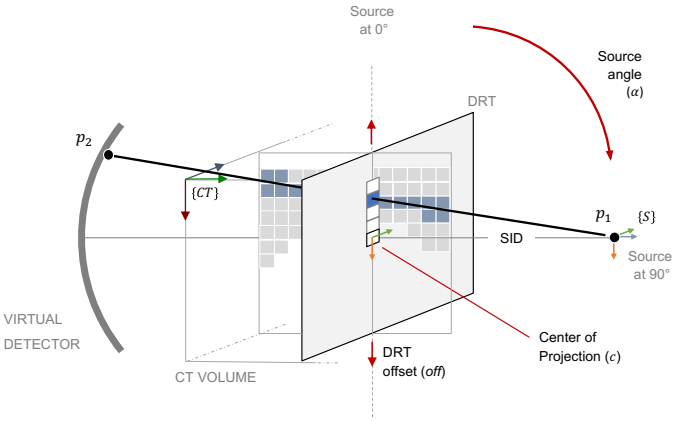


Figure 2: Projection geometry with the image volume {CT} and the X-ray model source {S} reference systems. One ray is casted from the source (p1), it crosses the whole volume and hits the detector in p2, contributing to the image intensity of one pixel of the DRT image located SID distance away from the source along the negative z-axis direction.

The mapping operator ${}^{S}_{CT}M$ is the result of composing an isocentric rotation of the source and its translations, including SID and the center of projection:

$$R_{SourceIso} = \begin{bmatrix} R_{Gantry} & R_{Gantry} \cdot \boldsymbol{c} - \boldsymbol{c} \\ 0 & 1 \end{bmatrix} \text{ and}$$

$$T_{SourceShift} = \begin{bmatrix} -c_x \\ SID - c_y \\ -c_z \end{bmatrix},$$

$$\text{With } R_{Gantry} = \begin{bmatrix} \cos(-\alpha) & -\sin(-\alpha) & 0 \\ \sin(-\alpha) & \cos(-\alpha) & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

From ${}^{S}_{CT}M$ it is possible to calculate the position of any arbitrary location (${}^{CT}_{\square}P$) in the CT field of view as it were seen from the local reference system mounted on the source (${}^{S}_{\square}P$). Afterwards, ${}^{S}_{\square}P$ is projected onto the DRT image plane, whilst correcting for the fan beam dispersion and using the intercept theorem as per equation 2.

$$ {}^{DRT}_{\square}P = \begin{bmatrix} {}^{S}_{\square}P_x * SID \,/\, abs({}^{S}_{\square}P_z) \\ {}^{S}_{\square}P_y \\ -SID \end{bmatrix} \qquad \text{(Eq. 2)}$$

Based on empirical evidence, an additional free parameter introducing a transversal shift of the DRT image plane is required to best replicate the actual topogram images. Such an offset (*off*) is applied only to the *x*-component of the DRT coordinates after projection, i.e. ${}^{DRT}_{\square}P'_x = {}^{DRT}_{\square}P_x + off$.

Modeling the geometric distortion of CT topogram images

Thirteen CT-dots (Beekley 119) of 2 mm diameter have been placed on blocks of light foam to sample pseudo-randomly the CT field of view. Both 3DCT and topogram images for lateral (LAT, $\alpha = 90°$) and posterior-anterior (PA, $\alpha = 180°$) projections have been acquired twice over the period of a year (dataset I and II). Topograms were acquired at two resolutions, using small and large field of view (FoV) for 1 mm and 2 mm in-plane pixel spacing images respectively. Fiducial positions have been localized manually in all datasets and the 3DCT positions were corrected for the longitudinal geometric distortion as per Eq. 1. Such undistorted fiducial positions have been then compared with the corresponding ones from topogram images to evaluate the longitudinal image distortion.

Calibration of fan beam projection geometry

The CT-dots dataset introduced before has been used for calibrating the fan beam projection geometry model by optimizing the projection error on the DRT plane of 3DCT points against the corresponding ones on topogram images acquired for the same setup. The dataset has then been extended by including 24 ceramic fiducials embedded on the two sides of the treatment couch top.
For the purpose of calibrating the fan beam projection geometry, the longitudinal direction is irrelevant, and therefore all locations have been collapsed as if they were lying onto a single axial image plane. The calibration routine is setup as a constrained optimization function using L-BFGS-B optimizer with 1e-13 tolerance for termination. Each projection angle and FoV has been calibrated individually.

Fast computation of DRT images

Digital image reconstruction is based on Siddon-Jacobs algorithm for incremental ray tracing [3]. Once the transformation that maps CT imaging coordinates into the source reference system (${}^{S}_{CT}M$) is known, its inverse can be used to map every pixel of the DRT image into the CT volume frame of reference. The same applies to the position of the source, and thus a ray direction can be defined from these two points in the {CT} reference system. The voxels of the 3DCT crossed by the ray are identified using Siddon's algorithm to compute the DRT pixel intensity value. In fan beam imaging, the lateral dispersion of the cone beam was

replaced with a sliding source model, with the ray calculation being adapted to reflect the geometry of topogram imaging with the virtual detector at the projection isocenter, rather than ray endpoint as it is for conventional radiograph imagers. Of note, the actual detector characteristic is not relevant to this parametrization as pixels of the DRT image represent rays going from the source *p1* to a point *p2* on the detector which is not linked to the physical elements of the CT array (Figure 2). Our implementation is a new interpolator filter for the Insight Toolkit (ITK) library.

Patient alignment with 2D/3D topogram registration

Clinical positioning using pairs of orthogonal topogram images has been simulated using a phantom patient head (CIRS, Norfolk, Virginia, United States). A ceramic sphere placed cranially to the phantom has been set as treatment isocenter and its laser tracker measurement used as a control point. After acquiring a planning 3DCT image (0.98 mm in-plane and 1 mm slice thickness), an arbitrary translational offset has been applied to the phantom before taking positioning topogram images. From these data, 2D/3D image-based correction has been computed in the CT scanner frame of reference and corrected for topogram-specific distortions before comparison with laser tracker reference measurements in world coordinates (Figure 1). The setup correction was obtained iteratively minimizing Mattes mutual information (50 bins) between two pairs of DRT and topogram images (Figure 3) with the Nelder-Mead optimizer (simplex delta=2, convergence tolerance = 0.1). The registration is based on the multi-threaded framework of ITK and makes use of the DRT model from the previous section. The procedure is DICOM compliant and supports standard radiotherapy formats.

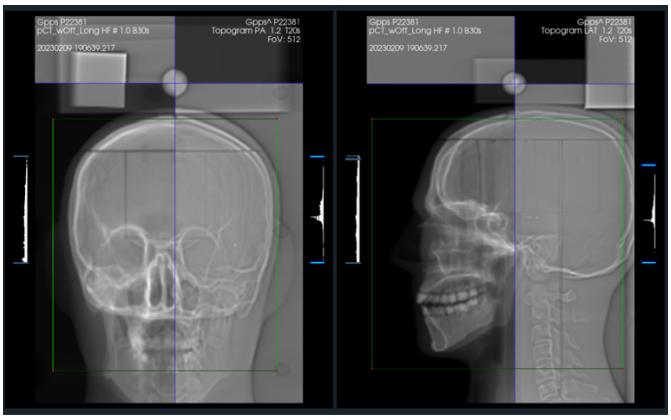


Figure 3: CT topogram (1st and 3nd quadrants) and DRT (2nd and 4th quadrants) images of the phantom head at two orthogonal projections. Above the head is the 2 cm sphere used as target control point.

3 Results

Geometric distortion of CT volumes

Optimal bias and scaling parameters (Eq. 1) have been computed from the dataset for the three scan ranges and assigned to the center-of-volume (CoV) of the respective images. These distortion parameters, scaling in particular, are not a linear function of the volume center – the further the scanner moves away from the parking position the larger the distortion (Figure 4). As such, the three experimental data points were fit with a second order polynomial curve (RMSE << 1e-5) to allow extrapolation of distortion parameters for any realistically possible CoV. The RMSE of the registration, based on the points corrected for longitudinal distortion was 0.27 mm on the calibration dataset and similarly less than 0.2 mm on two different validation datasets acquired with a CoV at 984 mm and 1015.5 mm distance from the CT parking position.

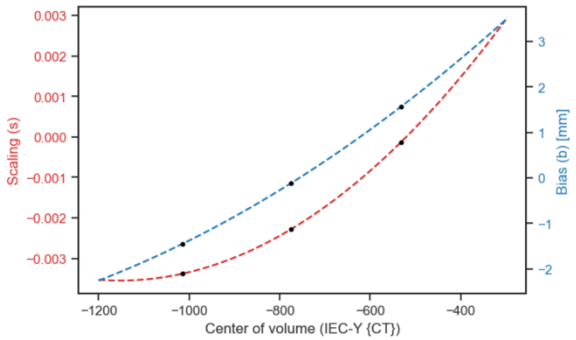


Figure 4: Scaling and bias experimental data as a function of the CT image center of volume and second order fit for paramters extrapolation.

Geometric distortion of CT topogram images

The geometric distortion model presented above has been used to compute the reference dot positions in the longitudinal direction from 3DCT images before comparison with corresponding topogram points. All data was fitted with Huber L2-regularized linear regression (RMSE: 0.33 mm) whose parameters describe the distortion correction of the topogram image coordinates relative to the scanner frame of reference (Figure 5). Residual errors (distance from fit line) on the calibration dataset were within ±0.25 mm of the corresponding (undistorted) reference 3DCT for 50% of the points; all data were within ±1 mm of such a reference.

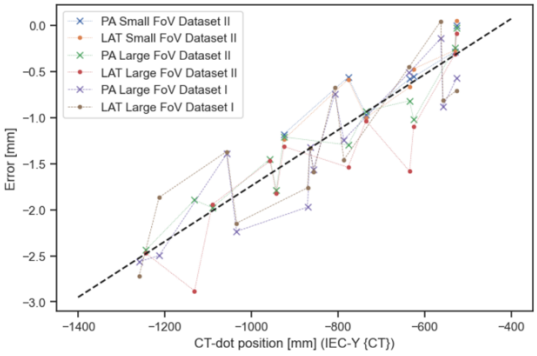


Figure 5: point error in longitudal direction as a function of the distance from the scanner frame of reference in parking position for PA and LAT images at two FoV sizes. Points connected with dotted lines belong to the same acquisition and imaging settings. Fit line is dashed.

Fan beam calibration

Lateral and posterior-anterior projections, each with small and large FoV, were calibrated individually. The patient position relative to the imaging equipment, e.g. head-first-supine / feet-first-supine, was also found to play a role in the topogram image representation, and therefore required specific calibration.
Overall, the RMSE of the corresponding point distances on the calibration dataset were below 0.1 and 0.3 mm for topograms acquired with small or large FoV respectively. A complete end-to-end validation is presented in the following section.

2D/3D registration of head phantom

Multiple registrations have been performed in sequence using topogram images at different resolutions. An initial manual registration that focused on the isocenter sphere has been followed by two automatic runs using topogram image pairs with small and large FoV. Only the phantom head region, excluding the control point sphere, has been considered for automatic registration. Sub-millimeter residual registration errors on the target sphere center with respect to laser tracker have been measured, with discrepancies below 0.3 mm using higher resolution topogram images.

Registration	FoV	x [mm]	y [mm]	z [mm]
Manual	Small	-0.18	-0.01	0.12
Automatic	Small	-0.27	-0.01	-0.05
Automatic	Large	-0.36	0.55	0.13

Table 1: residual errors following manual and automatic 2D-3D registration with CT topogram images with respect to the laser tracker measurement. Directions in IEC-world frame of reference (Figure 1).

4 Discussion

We present the geometrical characterization of a clinical on-rails CT scanner installed in a proton therapy treatment room for patient positioning verification. The procedures and methods for deriving a geometry model to correct the non-linearity in the longitudinal direction have been described for both 3DCT and topogram images, together with the computation of digitally reconstructed images for 2D/3D registration.
Although the concept is the same, i.e. linear scaling of the longitudinal axis coordinates, for 3DCT, the parameters for correcting distortion have a dependency on the center of volume position, whereas for topogram imaging a unique scaling was used. To a large extent, this is due to the difference in the mechanical operation of the scanner between the two imaging modalities. With its internal components spinning at high speed during 3D tomographic imaging, the CT scanner has a strong angular momentum opposing its longitudinal motion. Topogram imaging is different from the mechanical point of view of the scanner, as the source-detector ring is fixed to the angle of projection and this changes the dynamics of the movement. Second-order effects on topogram image geometry may have been masked by the use of manually identified reference positions, which suffer from significant geometric distortions despite using fiducials of small dimension. Higher quality topogram calibration data could allow for more refined models, although their necessity must be evaluated in the context of the overall accuracy of the system for the purpose of patient positioning, which has been shown to be already well within clinical tolerances.
The DRT projection geometry has been studied before in the context of Hi·Art TomoTherapy [4]. Unlike such physics-based models, we have here taken a computer graphics oriented approach, that although may compromise a little on absolute accuracy of DRT image intensity, on the other hand is detector agnostic and allows for implementation as a fast rendering routine. To the best of our knowledge, this is the first report of 2D/3D registration using topograms for clinical application.
The mechanical characterization of the robotic treatment couch, and its sag between imaging and treatment position, is a mandatory requirement to profit from intraoperative CT as a verification imager. However, this aspect is extremely specific to the design of the manipulator and the treatment room, and is beyond the scope of this contribution.

5 Conclusion

We introduced a model of the geometric distortion and fan-beam projection geometry of a CT on-rails. The model was further integrated into a 2D/3D registration framework with orthogonal topogram image pairs.
Once fitted to experimental data from our scanner, we report positioning errors of less than 0.3 mm in single directions for most common clinical settings, supporting the use of this intraoperative scanner as an imaging system for patient positioning in high-precision radiotherapy based on topogram imaging.

References

[1] K. Nesteruk, et al. “CT-on-Rails Versus In-Room CBCT for Online Daily Adaptive Proton Therapy of Head-and-Neck Cancers”. Cancers 13.23 (2021), p.5991. DOI: 10.3390/cancers13235991.

[2] G. Fattori et al. "Automated Fiducial Localization in CT Images Based on Surface Processing and Geometrical Prior Knowledge for Radiotherapy Applications". IEEE TBME 59.8 (2012), p. 2191-99. DOI: 10.1109/TBME.2012.2198822.

[3] R. L. Siddon. "Fast calculation of the exact radiological path for a threedimensional CT array" MedPhys 12 (1985), p. 252-5. DOI: 10.1118/1.595715.

[4] X. S. Qi, et al. “Clinical Development and Evaluation of Megavoltage Topogram for Fast Patient Alignment on Helical Tomotherapy”. Adv Radiat Oncol 5 (2020), p. 1334-41. DOI: 10.1016/j.adro.2020.05.014.

Bladder Cancer Hypoxia: Combining OE-MRI and Genomics for Biomarker Discovery

Rekaya Shabbir[1], Brian A Telfer[1], Ben Dickie[1], Muhammad Babur[1], Kaye Williams[2], Catharine ML West[1], Ananya Choudhury[1,3], Tim AD Smith[1].

[1] Division of Cancer Sciences, The University of Manchester, Manchester, UK.
[2] Division of Pharmacy & Optometry, The University of Manchester, Manchester. UK
[3] The Christie Hospitals NHS Foundation Trust, Manchester, UK

Abstract

Introduction: Hypoxic bladder patients benefit from hypoxia modification added to radiotherapy, but no biomarkers are used clinically for patient stratification. We aimed to investigate the potential of OE-MRI to identify hypoxia in xenografts derived from MIBC.

Methods: MIBC cells were inoculated into nude mice. Mice were imaged 1h after receiving a pimonidazole injection with an Agilant 7T 16cm bore magnet interfaced to a Bruker Avance III console and a T2-TurboRARE sequence using a dynamic MPRAGE acquisition. Dynamic Spoiled Gradient Recalled Echo images were acquired with 0.1mmol/kg Gd-DOTA injection. Tumours were harvested, sectioned, RNA extracted for transcriptomic analysis.

Results: Hypoxia scores were higher for pimo-high vs pimo-low samples (p=0.0002). Expression of known hypoxia-inducible genes were higher: *CA9* (p=0.012) and *SLC2A1* (p =0.012).

Conclusion: Hypoxic regions determined using OE-MRI were greater in the larger tumour. OE-MRI reflects other measures of successfully implemented in MIBC derived xenografts and differences in their transcriptomic profiles.

Introduction

The development of cancer is a multi-step process driven by both genetic and environmental factors [1]. Hypoxia is a feature of most solid tumours associated with genomic instability and treatment resistance particularly to radiotherapy [2]. Improved responses can be achieved by hypoxia modification during radiotherapy [3]. To date, there are no biomarkers in clinical use that identify patients with hypoxic tumours. Hypoxia scores derived from the collective expression of multiple genes have been shown to predict benefit of hypoxia modification [4-6]. Advanced functional imaging modalities have the ability to identify hypoxia and study genomic instability and heterogeneity of tumors associated with tumors [7]. Medical imaging techniques can identify regions within tumours for additional treatment dose and are suitable for serial measurements [8]. Combining two magnetic resonance imaging (MRI) techniques, oxygen-enhanced MRI (OE-MRI) and Dynamic Contrast Enhanced (DCE-MRI), enables the non-invasive mapping of hypoxia in tumours. [9]. DCE-MRI identifies regions within tumours that are perfused whilst OE-MRI show regions that utilise oxygen.
Imaging genomics links medical imaging with molecular profiling for biomarkers discovery. Radiogenomics offers a promising avenue for personalised medicine, enabling non-invasive characterization of tumours, prediction of treatment response, and improved patient outcomes [7, 10]. The aim of this study was to implement OE-MRI/DEC-MRI in a xenograft model of bladder cancer for future hypoxia biomarker target discovery to enhance the radiotherapy outcomes.

2 Materials and Methods

HT1367 MIBC cells were inoculated subcutaneously into the flanks of female CD nude mice with (1×10^6 cells in 100µl saline) (0.1ml vol). The injection was done under isoflurane anaesthesia. All experiments were carried out under UK Animals (Scientific Procedures) Act 1986, UK Home Office regulations. When the tumors reach 300 mm^3 (small) or 700 mm^3 (large) in diameter, mice were injected (ip) with the hypoxia marker pimonidazole (pimo) 1h before imaging. The Mice were then placed in an Agilant 7T 16cm bore magnet MRI interfaced to a Bruker Avance III console. The mice were first imaged with OE-MRI to obtain the oxygen signal following by DCE-MRI to identify the regions of hypoxia based on the diffusion. Animals were scanned using T2-TurboRARE, oxygen-enhanced MRI (OE-MRI), and dynamic-contrast enhanced MRI (DCE-MRI). OE-MRI scans were acquired using a dynamic MPRAGE acquisition: TR/TE=6000/3.10 ms, TI=2000 ms, voxel size=0.234 x 0.234 x 0.5 mm, 30 slices with total duration of 32 minutes. The anaesthetic carrier gas was switched from medical air to 100% O_2 midway through the OE-MRI scan, and kept at 100% O_2 for the remainder of the scanning session. Following OE-MRI, DCE-MRI was performed to determine perfused volume pixel (voxels). Dynamic Spoiled Gradient-Recalled (SPGR) images were acquired for 5 minutes 52 seconds (flip angle=30°, TR/TE=10.2/1.9 ms), with 0.1mmol/kg Gd-DOTA (Dotarem, Guerbet) injected after 60 seconds at 1mL/min with a tail vein cannular to intravenous (i/v), under anaesthesia, during the scan as a contrast enrichment.
Voxel size and field of view (FOV)of DCE-MRI and OE-MRI scans matched. The voxels with significant post-contrast enhancement ($p<0.05$) in DCE-MRI scans were classified as perfused and tissue was split into pOxyE and pOxyR regions. The tumours were harvested in liquid N_2, sectioned and RNA extracted from two 10um sections for transcriptomic analysis using Clariom S microarrays.

3 Results

1. Identification of oxygen-rich and perfused regions

Figure 1 illustrates regions of enhancement oxygen in tumours using OE-MRI (i) and prefusion using DCE-MRI (ii). The regions of oxygen enhancement and perfusion in were identified white (iii) in small and large xenograft bladder tumours.
Figure 1B shows montage plots of the oxygen-enhancement (OE-MRI) maps and perfusion (DCE-MRI) in slices from left to middle and from right to middle of the small and large tumour. Regions that are perfused normoxic (pOxy-E) are shown in yellow and regions that are perfused hypoxic (pOxy-R) are shown in blue.

Table 1 shows the overall measurement of normoxia (P-OxyE) and hypoxia (P-OxyR) and the proportion of tumours that are necrotic in the small and large tumour. Both hypoxia and necrosis are higher in the larger compared with the smaller tumour.

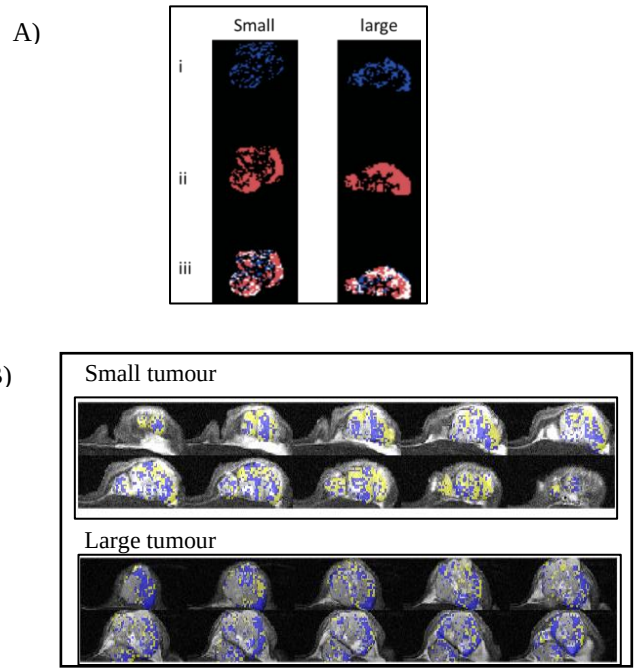


Figure 1: (A) Regions of oxygen enhancement (i) in blue, perfusion (ii) in red and a Montage (iii) of (i) and (ii) illustrating regions of perfusion and oxygen enhancement in white. (B) Montage of images derived by OE-MRI and DCE- MRI for the small tumour and large tumour. Top image in each panel reflects the right side of the tumour and bottom image reflects the left side of the tumour. Perfused normoxic areas (pOxyE) (yellow), perfused hypoxic areas (pOxyR) (blue) and not perfused areas identified (grey).

	Small tumour	Large tumour
Perfused	73%	48%
Perfused OxyE (normoxic)	49%	24%
Perfused OxyR (hypoxic)	51%	76%
Non-perfused (necrotic)	27%	52%

Table 1: Percentage of small and large tumour that is perfused, perfused with O_2-enhancement (normoxic), perfused without O_2-enhancement (hypoxic) and not-perfused (necrotic).

2. Identification of hypoxic, non-hypoxic and necrotic regions

Figure 2 shows regions stained with the pimo-antibody in sections immediately adjacent to the upper and lower sections micro-dissected for RNA extraction. The locality of both pimo-staining is similar in both indicating that the regions of hypoxia will be consistent though the micro-dissected.

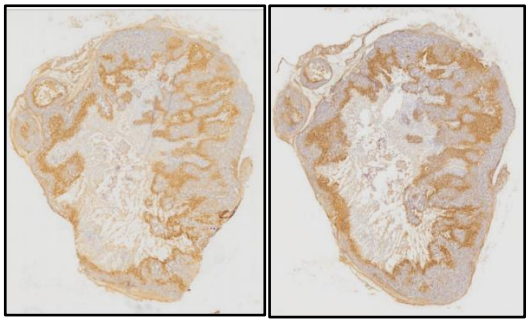


Figure 2: Sections above and below micro-dissected series of sections. RNA extraction from xenografted tumours.

3. Expression of genes associated with hypoxia

Figure 3 shows the increase in the expression of known hypoxia associated genes, *CA9* (p=0.012), *SLC2A1* (Glut-1) (p =0.012) and *VEGFA* (p=0.0152) in pimo-positive (red) compared with pimo-negative regions (blue).

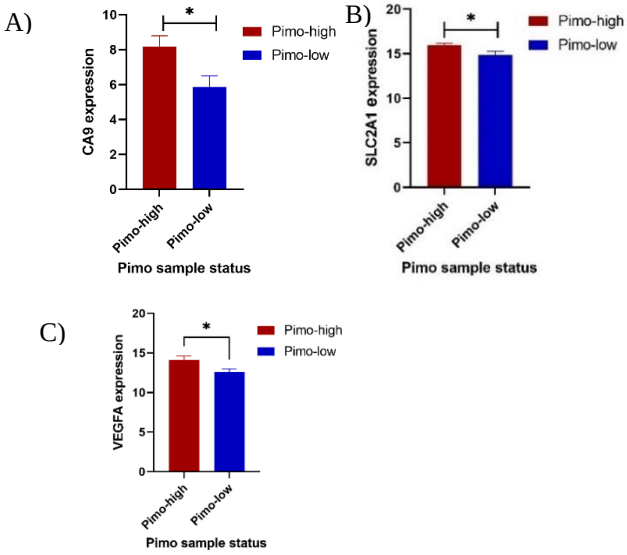


Figure 3: Expression of genes associated with hypoxia. Expression of (A) *CA9*, (B) *SLC2A1* and (C) *VEGFA* in pimo-positive and pimo-negative regions. *p<0.05, ***p<0.001.

4. Hypoxia Score (HS) generated using the West 24 gene hypoxia signature

Hypoxia scores (HS) were generated from both tumours using a 24-gene expression bladder cancer hypoxia signature (Yang et al 2017). Figure 4 shows the HS of each (A) individual sample and (B) mean of the pimo-positive

(n=3) and pimo-negative (n=3) samples. HS was higher (p=0.0052) in the pimo-positive (red circle) than in pimo-negative (blue circle) samples.

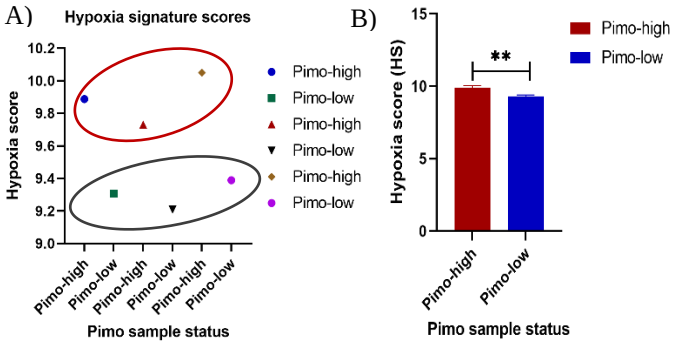


Figure 4: Hypoxia Scores (HS) generated from RNA extracted from pimo-positive and pimo-negative samples. (A) The HS in samples from each pimo-positive and pimo-negative sample. (B) Mean and SD of HS differences (** P.Value<0.0052) between Pimo-positive and Pimo-negative.

4 Discussion

Hypoxia is an important prognostic factor in solid tumors mediating resistance to radiotherapy. The need for hypoxia biomarkers is exemplified by Spiegelberg et al. who found that the failure of phase III trials can be fully explained by lack of patient classification based on tumour hypoxia status [11]. Combined DCE-MRI and OE-MRI for the identification of hypoxia has been validated in clinical [12] as well as pre-clinical [13] work, and is therefore a reliable tool to combine with genomics for hypoxia biomarker discovery [9]. The higher proportion of hypoxia in the larger tumour compared with the smaller tumour demonstrated using OE-MRI/DCE-MRI agrees with other studies [11, 14, 15]. Hypoxic tumours and necrosis pose significant challenges for radiotherapy [16]. Here, necrosis was also higher in the large tumour and was evident in over half the volume in agreement with previous studies reporting on size association with necrotic fraction [11, 17]. Several studies have shown that the presence of the necrotic regions in tumours associated with hypoxia [18, 19] and that patients with high necrosis benefit with carbogen-nicotinamide (CON) modified therapy [18]. Identifying the relationship between hypoxia and its role in promoting cancer progression and resistance to radiotherapy were performed. Tumour hypoxia induces angiogenesis through up-regulation of *VEGF* expression. Other mechanisms that enable cancer cells to thrive in hypoxic conditions include up regulation of cell surface expression of Glut-1 thus increasing glucose uptake to fuel glycolysis and *CA9* to reduce the consequent intracellular acidification. Genes encoding the cell surface proteins CAIX, GLUT-1 and *VEGF3* (*CA9*, *SLC2A1* and *VEGF3*), associated with hypoxia [20, 21], had increased expression in pimo-positive regions. Clinical work has shown that both *CA9* and Glut-1 co-localise with pimondazole (GLUT1: P<0.0001) and (*CA9*: P<0.0001) in patients with MIBC [20]. Studies have demonstrated that overexpression of CAIX and GLUT-1 protein in squamous cell Head and Neck Cancer and MIBC is associated with poor response to chemotherapy [22]. Genomic analysis helps to better understand the biology of bladder cancer leading to the identification of disease at early stages, facilitates patient stratification and predicts patent outcome. Combining hypoxia imaging techniques with gene expression signatures are likely to provide robust biomarkers of tumour hypoxia leading to improved patient treatment selection and develop personalised radiotherapy [23]. DCE-MRI parameters, like the pharmacokinetic marker ABrix, can pinpoint tumour hypoxia, a key factor in treatment resistance, while gene expression profiles reveal the underlying molecular landscape.

This approach allows for the construction of personalised gene signatures that predict response to radiotherapy, enabling tailored treatment strategies. This personalised approach, informed by both functional imaging and molecular fingerprints, has the potential to improve tumuor control, reduce side effects, and ultimately enhance radiotherapy outcomes for cancer patients.

5 Conclusion

Hypoxic regions determined using OE-MRI were greater in the larger tumour. OE-MRI reflects other measures of successfully implemented in MIBC derived xenografts and differences in their transcriptomic profiles determined in

same tumours. OE-MRI should be further investigated in bladder cancer.

References

1. Ding, L., et al., *Perspective on oncogenic processes at the end of the beginning of cancer genomics.* Cell, 2018. 173(2): p. 305-320. e10. DOI: 10.1016/j.cell.2018.03.033
2. Telarovic, I., R.H. Wenger, and M. Pruschy, *Interfering with tumor hypoxia for radiotherapy optimization.* Journal of Experimental & Clinical Cancer Research, 2021. 40(1): p. 1-26. DOI: 10.1186/s13046-021-02000-x
3. Hoskin, P.J., et al., *Radiotherapy with concurrent carbogen and nicotinamide in bladder carcinoma.* Journal of clinical oncology, 2010. 28(33): p. 4912-4918. DOI: 10.1200/JCO.2010.28.4950
4. Yang, L., et al., *A Gene Signature for Selecting Benefit from Hypoxia Modification of Radiotherapy for High-Risk Bladder Cancer Patients.* Clin Cancer Res, 2017. 23(16): p. 4761-4768. DOI: 10.1158/1078-0432.CCR-17-0038
5. Eustace, A., et al., *A 26-gene hypoxia signature predicts benefit from hypoxia-modifying therapy in laryngeal cancer but not bladder CancerGene signature predicts benefit from antihypoxia therapy.* Clinical cancer research, 2013. 19(17): p. 4879-4888. DOI: 10.1158/1078-0432.CCR-13-0542
6. Forker, L.J., et al., *Technical development and validation of a clinically applicable microenvironment classifier as a biomarker of tumour hypoxia for soft tissue sarcoma.* Br J Cancer, 2023. DOI: 10.1038/s41416-023-02265-3
7. Lee, G., et al., *Radiomics and its emerging role in lung cancer research, imaging biomarkers and clinical management: State of the art.* European Journal of Radiology, 2017. 86: p. 297-307. DOI: 10.1016/j.ejrad.2016.09.005
8. Grosu, A.L., et al., *Hypoxia imaging with FAZA-PET and theoretical considerations with regard to dose painting for individualization of radiotherapy in patients with head and neck cancer.* Int J Radiat Oncol Biol Phys, 2007. 69(2): p. 541-51. DOI: 10.1016/j.ijrobp.2007.05.079
9. O'Connor, J.P., et al., *Oxygen-Enhanced MRI Accurately Identifies, Quantifies, and Maps Tumor Hypoxia in Preclinical Cancer Models.* Cancer Res, 2016. 76(4): p. 787-95. DOI: 10.1158/0008-5472.CAN-15-2062
10. Mazurowski, M.A., *Radiogenomics: what it is and why it is important.* J Am Coll Radiol, 2015. 12(8): p. 862-6. DOI: 10.1016/j.jacr.2015.04.019
11. Hildingsson, S., et al., *Hypoxia in relationship to tumor volume using hypoxia PET-imaging in head & neck cancer - A scoping review.* Clin Transl Radiat Oncol, 2022. 36: p. 40-46. DOI: 10.1016/j.ctro.2022.06.004
12. Dubec, M.J., et al., *First-in-human technique translation of oxygen-enhanced MRI to an MR Linac system in patients with head and neck cancer.* Radiotherapy and oncology : journal of the European Society for Therapeutic Radiology and Oncology, 2023. 183: p. 109592. DOI: 10.1016/j.radonc.2023.109592
13. Salem, A., et al., *Oxygen-enhanced MRI Is Feasible, Repeatable, and Detects Radiotherapy-induced Change in Hypoxia in Xenograft Models and in Patients with Non–small Cell Lung Cancer.* Clinical Cancer Research, 2019. 25(13): p. 3818-3829. DOI: 10.1158/1078-0432.CCR-18-3932
14. SHIBAMOTO, Y., et al., *Variation in the hypoxic fraction among mouse tumors of different types, sizes, and sites.* Japanese Journal of Cancer Research GANN, 1986. 77(9): p. 908-915.
15. Stanley, J., W. Shipley, and G. Steel, *Influence of tumour size on hypoxic fraction and therapeutic sensitivity of Lewis lung tumour.* British journal of cancer, 1977. 36(1): p. 105-113. DOI: 10.1038/bjc.1977.160
16. Ellingsen, C., et al., *DCE-MRI of the hypoxic fraction, radioresponsiveness, and metastatic propensity of cervical carcinoma xenografts.* Radiotherapy and Oncology, 2014. 110(2): p. 335-341. DOI: 10.1016/j.radonc.2013.10.018
17. Belli, J.A. and J.R. Andrews, *Relationship between tumor growth and radiosensitivity.* Journal of the National Cancer Institute, 1963. 31(3): p. 689-703. DOI: 10.1093/jnci/31.3.689
18. Eustace, A., et al., *Necrosis predicts benefit from hypoxia-modifying therapy in patients with high risk bladder cancer enrolled in a phase III randomised trial.* Radiotherapy and Oncology, 2013. 108(1): p. 40-47. DOI: 10.1016/j.radonc.2013.05.017
19. Song, Y.P., et al., *Long-term outcomes of radical radiation therapy with hypoxia modification with biomarker discovery for stratification: 10-year update of the BCON (Bladder Carbogen Nicotinamide) phase 3 randomized trial (ISRCTN45938399).* International Journal of Radiation Oncology* Biology* Physics, 2021. 110(5): p. 1407-1415. DOI: 10.1016/j.ijrobp.2021.03.001
20. Hoskin, P., et al., *GLUT1 and CAIX as intrinsic markers of hypoxia in bladder cancer: relationship with vascularity and proliferation as predictors of outcome of ARCON.* British journal of cancer, 2003. 89(7): p. 1290-1297. DOI: 10.1038/sj.bjc.6601260
21. Giatromanolaki, A., et al., *Expression of hypoxia-inducible carbonic anhydrase-9 relates to angiogenic pathways and independently to poor outcome in non-small cell lung cancer.* Cancer research, 2001. 61(21): p. 7992-7998.
22. Koukourakis, M.I., et al., *Hypoxia-regulated carbonic anhydrase-9 (CA9) relates to poor vascularization and resistance of squamous cell head and neck cancer to chemoradiotherapy.* Clinical Cancer Research, 2001. 7(11): p. 3399-3403.
23. Liu, Z. and J. Zhang, *Radiogenomics correlation between MR imaging features and mRNA-based subtypes in lower-grade glioma.* BMC Neurology, 2020. 20(1): p. 259. doi: 10.1186/s12883-020-01838-6

Quantifying changes in ventilation for photon- and proton-treated patients according to dose and pneumonitis development

Rebecca Lim[1,2], Tien Tang[2], Austin Castelo[2], Caleb O'Connor[2], Yulun He[1,2], Zhongxing Liao[4], Radhe Mohan[3], Kristy Brock[2,3]

[1]Graduate School of Biomedical Sciences, The University of Texas MD Anderson Cancer Center UTHealth Houston, Houston, U.S.A.
[2]Department of Imaging Physics, The University of Texas MD Anderson Cancer Center, Houston, U.S.A.
[3]Department of Radiation Physics, The University of Texas MD Anderson Cancer Center, Houston, U.S.A.
[4]Department of Radiation Oncology, The University of Texas MD Anderson Cancer Center, Houston, U.S.A.

Abstract Treatment of non-small cell lung cancer with proton therapy allows for delivery of high radiation dose while sparing critical normal tissue structures compared to conventional photon therapy. However, there are currently no established clinical endpoints to study the biological effect between these radiation modalities. Computed tomography (CT) imaging allows for extraction of ventilation through deformable image registration of 4DCTs and may provide information on imaging-based functional changes in the lung. We hypothesize that patients treated with protons exhibit different changes in ventilation compared to those treated with photons. For 20 photon and 20 proton cases, ventilation was extracted using 4DCT Jacobian-based methods and correlated with dose and radiation-induced pneumonitis development. Ventilation was tracked over the course of treatment on a voxel-wise basis and stratified into high and low function. Ventilation differences across modalities were compared by function and pneumonitis development. Patients treated with protons exhibited statistically significant improvements in ventilation change of low functioning voxels compared to photon patients. Future work will evaluate ventilation change specifically in the lung regions that developed pneumonitis.

1 Introduction

Lung cancer is the leading cause of cancer death and the second most common form of cancer in the U.S [1]. 80-85% of lung cancer is non-small cell lung cancer (NSCLC) [1], and while surgical resection provides the best outcomes, radiotherapy (RT) is a more feasible option for many patients, especially those advanced in age. However, the lungs are particularly radiosensitive, and RT often leads to toxicities such as radiation-induced pneumonitis (RP) which reduce patient quality of life after treatment [2], [3].

Proton RT is a promising option to reduce normal tissue toxicity, but research to date has failed to demonstrate a clear difference in toxicity development between patients treated with photon or proton RT [4], [5]. Current clinical practice for toxicity modeling and outcome prediction in proton RT relies solely on regional dosimetric factors, such as dose volume histograms (DVHs) and does not incorporate patient-specific biological factors. However, since proton RT is known to produce a different biological effect compared to photon RT [6], [7], [8], biological factors should be considered when designing toxicity models.

Computed tomography ventilation imaging (CTVI) uses deformable image registration (DIR) of 4DCT images to calculate ventilation and has emerged as an efficient alternative to typical ventilation imaging such as single-photon emission computed tomography (SPECT) or positron emission tomography (PET). It has been validated in both animal and human studies [9], [10], and clinical trials have begun to show the potential of using CTVI to design treatment plans that reduce dose to highly functional areas of the lung for toxicity reduction [11], [12]. In particular, it has been shown that pre-treatment ventilation is correlated with dose and pneumonitis development for photon patients [13], [14]. While research has shown that CTVI can be feasibly incorporated into functional avoidance plans for proton therapy [15] and can improve toxicity models for photon patients [16], [17], [18], toxicity models for proton RT have not yet incorporated such functional changes to predict outcomes at a voxel level.

In this study, we extracted ventilation from 4DCTs using Jacobian-based ventilation for both photon- and proton-treated patients. We evaluated how each patient's ventilation changes on a voxel-wise basis over the course of treatment across treatment modality, RP development, and dose. We hypothesize that there will be significant differences in ventilation change over treatment across radiation modalities. In particular, since proton RT is known to reduce normal tissue toxicity, we hypothesize that proton-treated patients will have a greater capacity for ventilation improvement due to greater sparing of normal lung tissue.

2 Materials and Methods

40 patients treated for NSCLC were chosen retrospectively from a randomized clinical trial [4]. 20 patients were treated with photon RT, 12 of which developed RP, and 20 patients were treated with proton RT, 10 of which developed RP. Photon patients were treated with intensity modulated radiation therapy (IMRT) and proton patients were treated with passive scattering proton therapy (PSPT), each for 37 fractions at a prescription dose of 74 Gy. All patients received concurrent chemotherapy. Each patient had 4DCTs acquired for planning and weekly up to the final week of treatment, as well as a PET/CT scan acquired 3-12 months after treatment, which was used to diagnose RP. Dose plans for each patient were calculated on the exhale (T5) scan of the planning week.

The Jacobian determinant, shown below, is a representation of local volume change and can be calculated using DIR [9].

$$J(x,y,z) = \begin{vmatrix} 1+\frac{\partial u_x(x,y,z)}{\partial x} & \frac{\partial u_x(x,y,z)}{\partial y} & \frac{\partial u_x(x,y,z)}{\partial z} \\ \frac{\partial u_y(x,y,z)}{\partial x} & 1+\frac{\partial u_y(x,y,z)}{\partial y} & \frac{\partial u_y(x,y,z)}{\partial z} \\ \frac{\partial u_z(x,y,z)}{\partial x} & \frac{\partial u_z(x,y,z)}{\partial y} & 1+\frac{\partial u_z(x,y,z)}{\partial z} \end{vmatrix}$$

The Jacobian determinant of the displacement vector field (DVF) from each registration was used to determine each voxel's differential expansion or contraction, enabling the calculation of the specific volume change of the voxel, *J(x,y,z) – 1*. An additional scaling factor using the Hounsfield Unit (HU) values of the associated T5 scan of the 4DCT was incorporated to provide a more physiologically accurate measurement of ventilation and converted specific volume change to specific ventilation [19], denoted by *V*:

$$V = \left(\frac{-HU}{1000}\right)^{-1} * (J(x,y,z) - 1)$$

Figure 1 shows the workflow for determining voxel-wise correlation of the Jacobian over treatment. First, intra-fraction DIR using a hybrid intensity-based algorithm was performed between the T0 and T5 scans of the 4DCTs for the planning week and final week of treatment separately. Focus ROIs were used to align the lung contours during the registration process. The resulting intra-fraction DIRs were used to calculate the Jacobian determinant. Next, an inter-fraction DIR was performed between the T5 scans of the planning and final weeks of treatment. This step allowed for voxel-wise correspondence across timepoints. For photon patients, a biomechanical-based algorithm was used to perform the inter-fraction DIR, and a hybrid intensity-based algorithm was used for the proton patients. The DVF from the inter-fraction DIR was applied to the Jacobian map from the final week to map it to the planning week coordinate space. The resulting Jacobian maps were then scaled using the HU correction factor to calculate ventilation. Since the planned doses were also done on the T5 scan of the planning week, the dose maps match the Jacobian maps on a voxel level. Lastly, the gross tumor volume (GTV) was segmented using an in-house segmentation model and removed from the evaluation to avoid erroneous ventilation calculations. Grills et al. found that GTV contours underestimate composite pathologic tumor size by 1.2 mm [20], so all GTV contours were expanded by 3 mm to ensure consistency before removal from the lung contour.

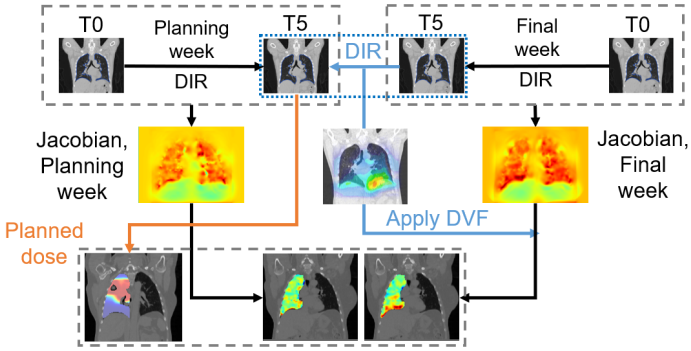


Figure 1: Workflow figure to extract voxel-wise change in ventilation over the course of treatment and accumulated dose using deformable image registration (DIR) and deformation vector fields (DVFs).

The change in ventilation between beginning and end of treatment for each voxel was calculated for each patient. Voxels were then stratified by function, toxicity, and treatment modality. To stratify voxels by function, the median Jacobian value was determined from the planning week ventilation map for each patient. A high functioning voxel was defined as a voxel that had a ventilation value above the median value in the planning week, and a low functioning voxel was defined to have a ventilation value below the median value in the planning week. After stratifying by function, voxels were binned by dose in 5 Gy increments and the ventilation differences were averaged in each bin. The averaged differences were plotted against dose for proton and photon patients, split into RP vs. non-RP patients for all voxels, then split again into high vs. low functioning voxels. T-tests were conducted on the ventilation differences between photon and proton patients, within each dose bin as well as across all dose bins.

3 Results

Boxplots of the average differences in ventilation for photon and proton patients are shown in Figure 2. Average differences further split into high and low functioning voxels are shown in Figure 3.

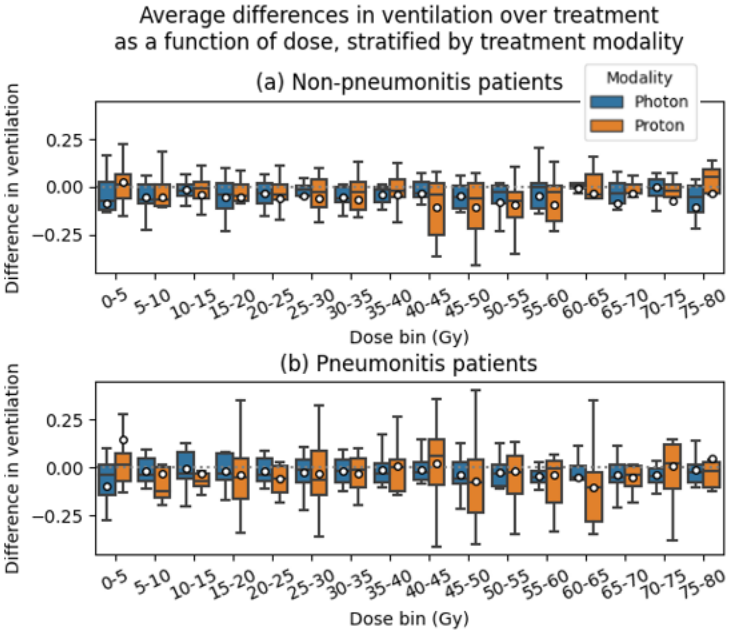


Figure 2: Boxplots of average differences in ventilation for patients that did not develop pneumonitis (a) and patients that did (b). Patients were split into photon and proton cohorts, shown in blue and orange respectively. Dose bins are 5 Gy in width and range from 0 to 80 Gy. Ventilation function is determined by the median of the ventilation values from the planning week.

High functioning voxels decreased in ventilation over the course of treatment while low functioning voxels increased in ventilation, with relatively small changes in overall ventilation averaged across function. In the high functioning voxels, ventilation tended to decrease by a greater magnitude for proton treated patients compared to photon treated patients. In contrast, proton patients

demonstrated an increase in ventilation in low functioning voxels where photon treated patients did not.

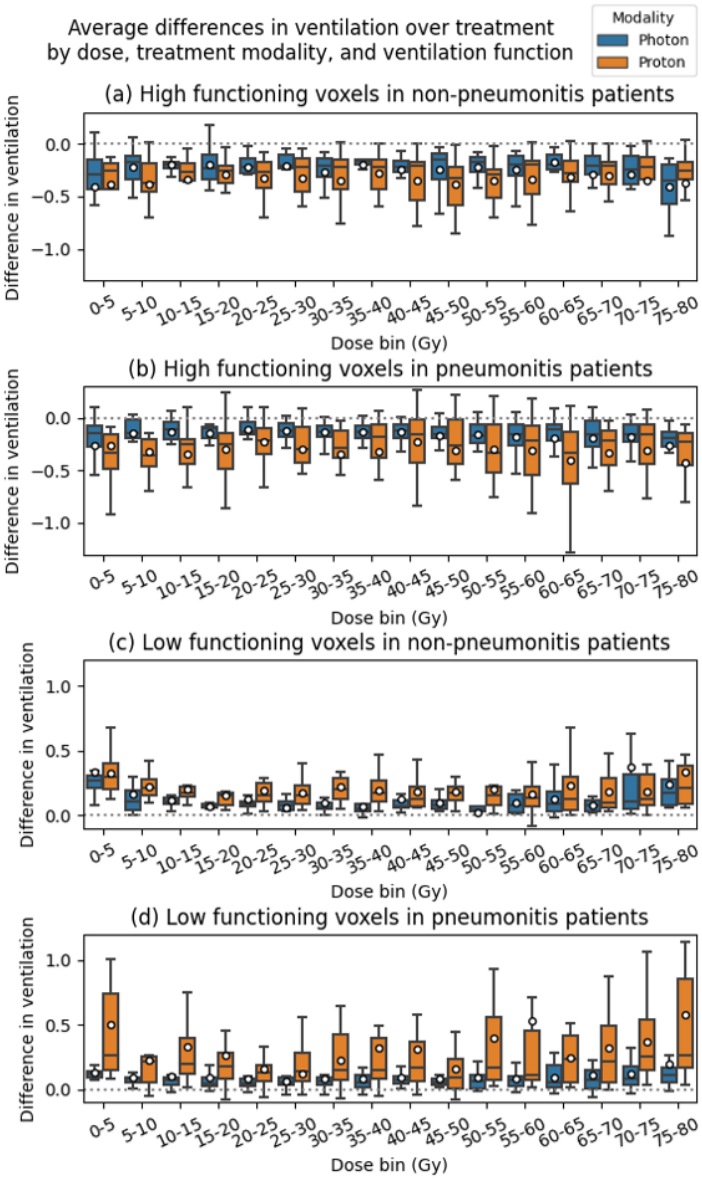


Figure 3: Boxplots of average differences in ventilation for pneumonitis (RP) and non-RP patients, split by toxicity and ventilation function. (a) and (b) show high functioning voxels in patients that did not and did develop RP respectively. (c) and (d) show low functioning voxels in patients that did not and did develop RP. Patients were split into photon and proton cohorts, shown in blue and orange respectively. Dose bins are 5 Gy in width and range from 0 to 80 Gy. Ventilation function is determined by the median of the ventilation values from the planning week.

Ventilation differences for all voxels prior to stratification by function were not found to be significant. T-test p-values for the high and low functioning voxels corresponding to Figure 2 are shown in Table 1. Ventilation differences were found to be statistically significant ($p < 0.05$) when considering all dose bins. Within each dose bin, statistical significance was found only in the low functioning voxels. In particular, the 15-20 Gy, 25-30 Gy, and 30-35 Gy dose bins were found to be statistically significant when comparing patients that did not develop RP, and the 70-75 Gy dose bin was significant when comparing patients that developed RP.

Table 1: P-values for t-tests conducted across treatment modalities. T-tests were performed for each cohort within each dose bin and over all dose bins combined. Statistically significant p-values are shown in bold.

Dose bin (Gy)	Low function, non-RP, photon vs. proton	Low function, RP, photon vs. proton	High function, non-RP, photon vs. proton	High function, RP, photon vs. proton
0-5	0.923	0.061	0.9097	0.989
5-10	0.408	0.137	0.1582	0.247
10-15	0.063	0.062	0.1107	0.064
15-20	**0.032**	0.161	0.3099	0.156
20-25	0.299	0.309	0.2504	0.149
25-30	**0.030**	0.667	0.2977	0.105
30-35	**0.037**	0.111	0.5015	0.093
35-40	**0.019**	0.172	0.3334	0.375
40-45	0.413	0.160	0.3549	0.419
45-50	0.327	0.319	0.2546	0.377
50-55	0.058	0.067	0.1319	0.209
55-60	0.430	0.200	0.4536	0.310
60-65	0.304	0.054	0.1033	0.172
65-70	0.090	0.088	0.9728	0.285
70-75	0.415	**0.047**	0.6856	0.313
75-80	0.551	0.100	0.8147	0.376
All bins	**0.003**	**1.76E-08**	**0.003**	**1.96E-05**

4 Discussion

Statistical significance was found when comparing all dose bins, indicating that there is a significant difference in ventilation overall when comparing modalities. High functioning voxels exhibited a decrease in ventilation over the course of treatment for all patients while low functioning voxels increased over the course of treatment. These findings were similarly demonstrated in Patton et al. [13], who showed that higher functioning voxels decreased by 4.8% per 10 Gy compared to lower functioning voxels, which decreased by 2.4%. Additionally, statistical significance within dose bins was found mostly above 20 Gy, which has also been shown to be a significant threshold for toxicity development by King et al. [21] and Graham et al. [22].

Proton patients tended to have more extreme changes in ventilation overall as demonstrated by the t-tests. In the low functioning voxels, the increase in ventilation was significantly higher for proton patients, supporting the hypothesis that the higher sparing of normal tissue in proton RT allows for greater increase in function, due to sparing of normal tissue.

Some limitations of this study include the lack of spatial information for each voxel and the large amounts of averaging for each patient. Future work will include a more in-depth analysis of the voxels within the regions of the lung that developed RP for patients that were diagnosed with RP. Since different DIR strategies were employed across time points for photon patients compared to proton patients, we will also be comparing results achieved with the same inter-fraction DIR. The photon patients in our dataset also have accumulated dose plans, and a future study will compare how ventilation changes for planned and accumulated doses. Lastly, additional patients will be incorporated for further analysis.

5 Conclusion

We developed a workflow to calculate voxel-wise change in lung ventilation and dose using Jacobian-based methods. Proton patients exhibited more extreme changes in ventilation in comparison to photon patients. Future work will incorporate more patients and work towards establishing CT based lung ventilation as a functional endpoint when evaluating differences between photon and proton therapy.

References

[1] R. L. Siegel, K. D. Miller, N. S. Wagle, and A. Jemal, "Cancer statistics, 2023," *CA A Cancer J Clinicians*, vol. 73, no. 1, pp. 17–48, Jan. 2023, doi: 10.3322/caac.21763.

[2] X.-J. Zhang *et al.*, "Prediction of radiation pneumonitis in lung cancer patients: a systematic review," *J Cancer Res Clin Oncol*, vol. 138, no. 12, pp. 2103–2116, Dec. 2012, doi: 10.1007/s00432-012-1284-1.

[3] M. Aiad *et al.*, "Comparison of Pneumonitis Rates and Severity in Patients With Lung Cancer Treated by Immunotherapy, Radiotherapy, and Immunoradiotherapy," *Cureus*, Jun. 2022, doi: 10.7759/cureus.25665.

[4] Z. Liao *et al.*, "Bayesian Adaptive Randomization Trial of Passive Scattering Proton Therapy and Intensity-Modulated Photon Radiotherapy for Locally Advanced Non–Small-Cell Lung Cancer," *JCO*, vol. 36, no. 18, pp. 1813–1822, Jun. 2018, doi: 10.1200/JCO.2017.74.0720.

[5] S. Mesko and D. Gomez, "Proton Therapy in Non-small Cell Lung Cancer," *Curr. Treat. Options in Oncol.*, vol. 19, no. 12, p. 76, Nov. 2018, doi: 10.1007/s11864-018-0588-z.

[6] H. Paganetti, "Relative biological effectiveness (RBE) values for proton beam therapy. Variations as a function of biological endpoint, dose, and linear energy transfer," *Phys. Med. Biol.*, vol. 59, no. 22, p. R419, Oct. 2014, doi: 10.1088/0031-9155/59/22/R419.

[7] E. T. Vitti and J. L. Parsons, "The Radiobiological Effects of Proton Beam Therapy: Impact on DNA Damage and Repair," *Cancers*, vol. 11, no. 7, Art. no. 7, Jul. 2019, doi: 10.3390/cancers11070946.

[8] R. Mohan and D. Grosshans, "Proton therapy – Present and future," *Advanced Drug Delivery Reviews*, vol. 109, pp. 26–44, Jan. 2017, doi: 10.1016/j.addr.2016.11.006.

[9] J. M. Reinhardt, K. Ding, K. Cao, G. E. Christensen, E. A. Hoffman, and S. V. Bodas, "Registration-based estimates of local lung tissue expansion compared to xenon CT measures of specific ventilation," *Medical Image Analysis*, vol. 12, no. 6, pp. 752–763, Dec. 2008, doi: 10.1016/j.media.2008.03.007.

[10] "The VAMPIRE challenge: A multi-institutional validation study of CT ventilation imaging - Kipritidis - 2019 - Medical Physics - Wiley Online Library." Accessed: Feb. 02, 2024. [Online]. Available: https://aapm.onlinelibrary.wiley.com/doi/full/10.1002/mp.13346

[11] E. Vicente *et al.*, "Functionally weighted airway sparing (FWAS): a functional avoidance method for preserving post-treatment ventilation in lung radiotherapy," *Phys. Med. Biol.*, vol. 65, no. 16, p. 165010, Aug. 2020, doi: 10.1088/1361-6560/ab9f5d.

[12] T. Yamamoto *et al.*, "Four-Dimensional Computed Tomography Ventilation Image-Guided Lung Functional Avoidance Radiation Therapy: A Single-Arm Prospective Pilot Clinical Trial," *International Journal of Radiation Oncology*Biology*Physics*, vol. 115, no. 5, pp. 1144–1154, Apr. 2023, doi: 10.1016/j.ijrobp.2022.11.026.

[13] T. J. Patton, S. E. Gerard, W. Shao, G. E. Christensen, J. M. Reinhardt, and J. E. Bayouth, "Quantifying ventilation change due to radiation therapy using 4 DCT Jacobian calculations," *Medical Physics*, vol. 45, no. 10, pp. 4483–4492, Oct. 2018, doi: 10.1002/mp.13105.

[14] M. Otsuka, H. Monzen, N. Kadoya, M. Inada, K. Matsumoto, and Y. Nishimura, "Evaluation of Lung Toxicity Risk with Computed Tomography Ventilation Functional Image for Lung Stereotactic Body Radiation Therapy and Three-Dimensional Conformal Radiation Therapy," *International Journal of Radiation Oncology*Biology*Physics*, vol. 99, no. 2, Supplement, p. E708, Oct. 2017, doi: 10.1016/j.ijrobp.2017.06.2306.

[15] Q. Huang *et al.*, "Dosimetric feasibility of 4DCT-ventilation imaging guided proton therapy for locally advanced non-small-cell lung cancer," *Radiat Oncol*, vol. 13, no. 1, p. 78, Apr. 2018, doi: 10.1186/s13014-018-1018-x.

[16] A. M. Faught *et al.*, "Evaluating Which Dose-Function Metrics Are Most Critical for Functional-Guided Radiation Therapy," *International Journal of Radiation Oncology*Biology*Physics*, vol. 99, no. 1, pp. 202–209, Sep. 2017, doi: 10.1016/j.ijrobp.2017.03.051.

[17] Y. Vinogradskiy *et al.*, "Use of 4-Dimensional Computed Tomography-Based Ventilation Imaging to Correlate Lung Dose and Function With Clinical Outcomes," *International Journal of Radiation Oncology*Biology*Physics*, vol. 86, no. 2, pp. 366–371, Jun. 2013, doi: 10.1016/j.ijrobp.2013.01.004.

[18] A. M. Faught *et al.*, "Evaluating the Toxicity Reduction With Computed Tomographic Ventilation Functional Avoidance Radiation Therapy," *International Journal of Radiation Oncology*Biology*Physics*, vol. 99, no. 2, pp. 325–333, Oct. 2017, doi: 10.1016/j.ijrobp.2017.04.024.

[19] R. Castillo, E. Castillo, J. Martinez, and T. Guerrero, "Ventilation from four-dimensional computed tomography: density versus Jacobian methods," *Phys. Med. Biol.*, vol. 55, no. 16, pp. 4661–4685, Aug. 2010, doi: 10.1088/0031-9155/55/16/004.

[20] I. S. Grills *et al.*, "Clinicopathologic Analysis of Microscopic Extension in Lung Adenocarcinoma: Defining Clinical Target Volume for Radiotherapy," *International Journal of Radiation Oncology*Biology*Physics*, vol. 69, no. 2, pp. 334–341, Oct. 2007, doi: 10.1016/j.ijrobp.2007.03.023.

[21] M. T. King, P. G. Maxim, M. Diehn, B. W. Loo, and L. Xing, "Analysis of Long-Term 4-Dimensional Computed Tomography Regional Ventilation After Radiation Therapy," *Int J Radiat Oncol Biol Phys*, vol. 92, no. 3, pp. 683–690, Jul. 2015, doi: 10.1016/j.ijrobp.2015.02.037.

[22] M. V. Graham *et al.*, "Clinical dose-volume histogram analysis for pneumonitis after 3D treatment for non-small cell lung cancer (NSCLC)," *Int J Radiat Oncol Biol Phys*, vol. 45, no. 2, pp. 323–329, Sep. 1999, doi: 10.1016/s0360-3016(99)00183-2.

Introducing a co-creation design methodology for AI-based error detection in radiotherapy

Authors: L. Heising[1,2], F. Verhaegen[1], V. Mottarella[3], Y. Chong[3], M. Zamburlini[3], B. Nijsten[1], A. Swinnen[1], M. Oellers[1], M. Jacobs[1,2], C.X.J. Ou[2], C.J.A. Wolfs[1]

[1]Department of Radiation Oncology (Maastro), GROW School for Oncology and Reproduction, Maastricht University Medical Centre+, Maastricht, The Netherlands

[2]Department of Management, Tilburg School of Economics and Management, Tilburg University, Tilburg, The Netherlands

[3]Varian, a Siemens Healthineers Company, Baden, Switzerland

Abstract In vivo dosimetry (IVD) using the electronic portal imaging device (EPID) can increase accuracy of radiotherapy treatment (RT)by catching treatment errors, monitoring dose delivery, and assist in treatment adaptations. As manual assessment of information generated by the EPID is labor intensive work, artificial intelligence (AI) has been proposed to catch treatment errors. To foster the translation from research to clinical implementation, a human-centric design process consisting of a five-day design sprint was taken, yielding a prototype of an interface to interact with the AI, including AI certainty metric, explainable AI (XAI), and a decision-aid tool. The prototype was well received by end-users. Further work will focus on the AI certainty metric, and development of appropriate XAI.

1 Introduction

In vivo dosimetry (IVD) aims to catch treatment errors, assist in treatment adaptation, and monitor the actual dose delivered to the patient [1]. For external beam radiotherapy, the electronic portal imaging device (EPID) lends itself well for IVD. In practice, IVD using EPID is rarely performed, as assessing the complex and multitude of EPID-generated information is too time consuming to act upon treatment errors in time as required for dose-guided radiotherapy (DGRT). Recently, developments using artificial intelligence (AI) allow for rapid processing of these large quantities of complex EPID generated data to extract relevant information for further decision making [2].

In practice, only few AI systems developed for supporting the radiotherapy (RT) workflow reach the clinic. To ensure maximum utility of the AI and foster human-AI collaboration, the development of clinical AI systems interacting with humans requires a human-centric approach [3]. To promote IVD for clinical practice, we designed an AI-based system to evaluate treatment accuracy and detect treatment errors by automatically analyzing EPID images, while presenting human-interpretable information to users.

2 Materials and Methods

Design thinking is a systematic innovation process that prioritizes needs and challenges of the end user [4], instead of first progressing through a lengthy development process before presenting a finished product to the end users. A team of experts with expertise in user experience design, AI and medical physics participated in a five-day design sprint using the principles of design thinking to develop a prototype of the proposed system. In total, eight people actively participated in the design process which was a mix of researchers and experts from industry. An additional six experts, four user experts and two industry experts, were consulted during the exploration and testing phase. The co-creation design methodology covers the phases of empathize, define, ideate, prototype and test (Figure 1).

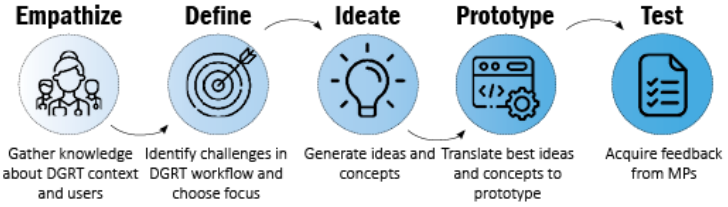


Figure 1 Steps of the co-creation design process.

The goal of the interface development was collectively drawn up and reads as follows: *"With DGRT.AI based on EPID & AI in 5 years, the medical physicist will be able to confidently detect and verify treatment errors, including root cause analysis, quickly enough to support decision-making, and taking actions when needed, always before the next fraction, to achieve safer treatment for every patient."*

In the empathize phase, three medical physicists (MPs) and three experts on radiotherapy quality assurance were interviewed. During the interviews, the goal of the interface and workflow mapping was validated, and challenges or bottlenecks in the DGRT workflow were identified.

2.1 Materials

A rapid prototype was developed using Figma inc [5]. The prototype allows for interaction with users. Synthetic images of the projected errors were generated using an in-house tool to introduce errors into the CT image, allowing for realistic synthetic images [6].

2.2 Evaluation

Finally, the prototype was tested with the end-users. The users were asked to provide feedback and to rate the system on a 7-Likert scale based on the technology acceptance model [7].

3 Results

Resulting from the design process, some novel features to interact with the AI were drawn up. These feature take into account information overload by providing a compact view of the fractions that need assessment. An in depth investigation is possible by clicking through the fraction, presenting all the information the MP needs to assess the fraction for errors.

3.1 AI-based treatment error detection

To detect the treatment error, a convolutional neural network is employed trained on time-integrated 2D gamma maps [2]. With the current performance, the AI is able to detect ten different error classes with 86.6% accuracy on synthetic data. Ongoing work aims to translate these results to real-world data. In case of an error, the error type is shown in the general overview of the developed prototype.

3.2 Overview fractions

After logging into the system, the user is provided with a list of all fractions that were flagged by the software. General information about the patient is provided, such as name, patient ID, and how far along the patient is in their treatment. The latter is important information for the user to judge how urgent the patient's fraction needs to be assessed, e.g., the last fraction of a treatment will be less urgent than a fraction in the middle of a treatment. The view is automatically ordered on error status so that the fractions containing an error are on top of the list. On the start page, the MP can see in a blink of an eye which fractions have the highest priority for further investigation as the AI-detected error is shown, including status and AI certainty score. If the AI certainty score is low (threshold to be determined), the fraction requires manual assessment by the MP (Figure 2). The information provided in the first view is kept at a minimum to prevent information overload. This was confirmed during the evaluation phase, where the MPs noted the simplicity of the system and highlighted the ease of use as their first reaction.

Error	Status	AI Certainty
Tumor regression	Failed	99.0%
Tumor shift	To be investigated	44.3%

Figure 3 Snippet of the AI information given in the first overview. The system shows which error was detected. In case of an error with high certainty, the system will tell the MP that the fraction failed. In case of low AI certainty, manual investigation by the MP is required.

3.3 Detailed fraction information

By clicking on the fraction, information relevant to the AI-classified treatment error is provided. In the example of tumor regression, the most relevant information needed to confirm the AI error classification is anatomical information, so the planning CT and CBCT images are shown (Figure 3). If desired, the user can find additional information in the left panel, such as dose information, PDIs including gamma analysis, dose recalculation based on log files, or a 3D rendering of the CT and dose.



Figure 2 Detailed view of the prototype showing the anatomical information of the patient. The MP can choose another view in the left panel. Upper red dashed area shows the AI-classified error, including the certainty score, and the options the MP has (review AI or Make decision).

3.4 AI certainty

An important property of the system is the AI certainty. By providing the AI certainty, the user may get an idea which fractions need to be manually assessed. As one of the expert users pointed out, the combination between the error status and the AI certainty gives an idea of the case complexity. While this feature was generally well received by the end-users, one of the users pointed out that the certainty score can cause some confusion. To prevent this, a detailed explanation should be available which describes how the AI certainty metric was established and how the metric should be interpreted.

3.5 Decision making aid

In practice, it is not always clear what the best decision is after detecting a certain treatment error. In our design, we propose to follow up based on protocols embedded in the system. We envision the embedded protocols to be developed based on predefined guidelines by the radiotherapy community (e.g., by ESTRO or AAPM), however, clinics can adjust the protocol to match their own protocol. Another option for the MP is to view the hospital's most used decisions for the detected treatment error to evaluate which actions have been taken in similar cases. This is a process to establish the human-AI shared decision making.

3.6 Explainable AI and feedback

If the user is skeptical about the AI decision (e.g., in case of low AI certainty), they can choose to evaluate the AI by looking at the AI explanation. In our prototype, a heatmap is overlaid over the gamma map highlighting the areas of the image that the AI considered most important for the error detection, as an explainable AI (XAI) method. The heatmap in the prototype was artificially developed. To avoid information overload to the user, the XAI information is available upon request. If the user agrees with the AI prediction, they can accept. The end-users expressed to appreciate the XAI feature to gain trust in the AI decisions.

4 Discussion

Using a co-creation strategy we were able to, within a week, prototype a novel human-interpretable system to detect RT treatment errors and aid the DGRT decision making process. Where practitioners typically choose to focus on IGRT or DGRT, our prototype allows for a combination of both to exploit their advantages, as well as AI-powered error detection which can substantially aid in streamlining these workflows. The prototype sparked enthusiastic responses from end-users. Regardless, some considerations for future work should be kept in mind.

The AI certainty metric played an important role in the MP's trust in the AI system. Computing a trustworthy AI certainty metric has not been the focus of our prototype development, but presents a challenge for the development of the presented system. To account for an accurate AI certainty metric, we endeavor to develop an AI verification framework that is able to output a certainty estimate that is accurate enough to be used in clinical practice. Additionally, in a follow up study, we aim to investigate how the explainable AI feature should be designed to best fit the user needs. XAI is a novel line of research that aims to explain why the AI makes a certain prediction. The XAI requirements are very dependent on the user's knowledge and needs. The MPs liked the XAI feature, but the current solution was difficult to interpret. Designing an optimal AI explanation was outside of scope for the presented prototype, but will be incorporated in future work.

One of the challenges identified during the exploration phase was the trend analysis of errors. The current design does not include this feature. In the evaluation phase, however, it was mentioned that trend analysis would be an interesting feature to add. In future work, we endeavor to study how AI can detect systematic and random inter-fraction treatment errors.

5 Conclusion

By undertaking a novel approach in the design of an interface for automatizing AI for radiotherapy, we developed a prototype to aid IVD. The system is designed to provide the user with enough information to evaluate each fraction in a first glance, but allows for further exploration of the data collected each fraction and AI tool. To empathize, define, ideate, prototype and test the system, the users participate the whole design cycle to leverage both human intelligence and artificial intelligence to make clinical process with more informed, efficient and effective decisions. Novel components of the system include an AI certainty estimation, XAI and feedback component, and an AI-driven decision aid.

References

[1] I. Olaciregui-Ruiz *et al.*, "In vivo dosimetry in external beam photon radiotherapy: Requirements and future directions for research, development, and clinical practice," *Phys. Imaging Radiat. Oncol.*, vol. 15, pp. 108–116, Jul. 2020, doi: 10.1016/j.phro.2020.08.003.

[2] C. J. A. Wolfs, R. A. M. Canters, and F. Verhaegen, "Identification of treatment error types for lung cancer patients using convolutional neural networks and EPID dosimetry," *Radiother. Oncol.*, vol. 153, pp. 243–249, Dec. 2020, doi: 10.1016/j.radonc.2020.09.048.

[3] H. Chen, C. Gomez, C.-M. Huang, and M. Unberath, "Explainable medical imaging AI needs human-centered design: guidelines and evidence from a systematic review," *Npj Digit. Med.*, vol. 5, no. 1, Art. no. 1, Oct. 2022, doi: 10.1038/s41746-022-00699-2.

[4] J. P. Roberts, T. R. Fisher, M. J. Trowbridge, and C. Bent, "A design thinking framework for healthcare management and innovation," *Healthcare*, vol. 4, no. 1, pp. 11–14, Mar. 2016, doi: 10.1016/j.hjdsi.2015.12.002.

[5] P. Faria, *figma: Web Client/Wrapper to the "Figma API."* 2023.

[6] C. J. A. Wolfs *et al.*, "Detection of anatomical changes in lung cancer patients with 2D time-integrated, 2D time-resolved and 3D time-integrated portal dosimetry: a simulation study," *Phys. Med. Biol.*, vol. 62, p. 6044, Aug. 2017, doi: 10.1088/1361-6560/aa7730.
[7] F. D. Davis, "Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology," *MIS Q.*, vol. 13, no. 3, pp. 319–340, 1989, doi: 10.2307/249008.

Dose driven gaze optimization in ocular proton therapy

Daniel Björkman[1,2], Riccardo Via[1], Tony Lomax[1,2], Maria De Prado[1], Alessia Pica[1], Guido Baroni[3], Damien C. Weber[1,4,5], and Jan Hrbacek[1]

[1]Center for Proton Therapy (CPT), Paul Scherrer Institute, Villigen, Switzerland
[2]Department of Physics, Eidgenössische Technische Hochschule (ETH), Zürich, Switzerland
[3]Politecnico di Milano, Department of Electronics Information and Bioengineering, Milano, Italy
[4]Inselspital, Bern University Hospital, University of Bern, Department of Radiation Oncology, Bern, Switzerland
[5]University Hospital Zurich, Department of Radiation Oncology, Zurich, Switzerland

Abstract
This study introduces a novel approach to ocular proton therapy treatment planning of ocular tumours. The aim is to autonomously equip planners with information derived from recent normal tissue complication probability (NTCP) models and predictive factors for radiation induced side effects, to assist the planners in sparing of healthy tissue. This is accomplished by utilizing our in-house dose engine OptisTPS, where we integrate machine learning (ML) and a tailored cost function to streamline computations for identifying an attainable eye orientation for treatment. The ML predictions efficiently reduce the search area for attainable eye orientations which has been demonstrated for a cohort of 190 patients. Results from the integration of the cost function and NTCP models are displayed for sampled gaze fixation points, where color represents either cost of arbitrary unit or the probability of radiation-induced toxicity, across 12 investigated patients.

1 Introduction

Ocular proton therapy (OPT) planning, established with Eyeplan [1] over four decades ago, aims to ensure target coverage while preserving delicate eye structures' functionality. Planners select gaze fixation points to determine suitable eye orientations for isocentric treatment. This study aim to identify a close to optimum eye orientation for treatment, prioritizing the sparing of healthy tissue, whilst also preserving target coverage. This is accomplished by applying penalties to dose points linked to elevated risks of radiation-induced toxicity, as outlined in [2], through dose calculations performed by our in-house developed dose engine which is a part of the OptisTPS framework. ML is used to minimize the required computational time by initially predicting a starting point for the optimization algorithm. Dose calculations are subsequently analyzed for the probability of radiation-induced side effects.

2 Materials and Methods

We integrate dose calculations from our in-house treatment planning system, OptisTPS, with a patient 3D point cloud model (Figure 1) exported from Eyeplan. This integration involves incorporating an ML algorithm to reduce the search area for determining a favourable eye orientation. All data used in this study originates from patients previously treated at our facility, the Centre for Proton Therapy (CPT), located at the Paul Scherrer Institute (PSI) in Switzerland. OptisTPS calculates 3D dose distributions based on proton range and modulation length, derived from target and patient surface models. The dedicated dose engine in OptisTPS follows the principles outlined in equation 5 of [3], integrated into an object-oriented framework with design decisions drawing inspiration from the matRad system [4]. Recently, OptisTPS was validated against Eyeplan and compared to measured patient specific depth dose profiles for 50 patients [5].

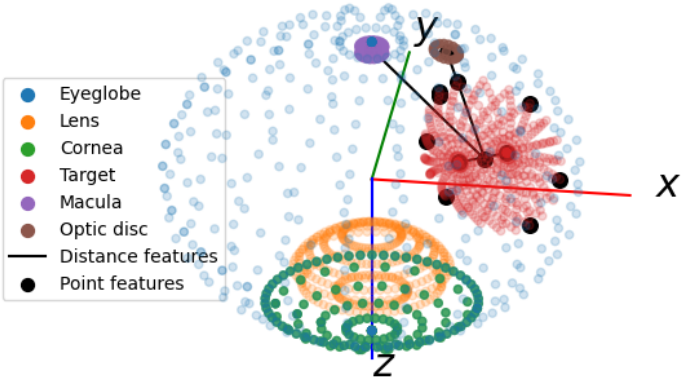


Figure 1: Point cloud patient model exported from Eyeplan in the gaze centered reference frame with highlighted features for gaze point prediction.

The 3D point cloud model of the patient is characterized by a spatial arrangement of points that represent the target and eye structures at risk, including the lens, cornea, macula, optic disc, and the entire eye globe from which the target is subtracted. Its orientation is determined by its rotation point and the direction of its gaze vector. We define a cost function that incentivizes the minimization of specific dose points associated with an increased risk of inducing secondary toxicities in eye structures at risk. The primary objective of including the ML algorithm is to predict an initial attainable gaze fixation point thus narrowing the search area for determining a favourable eye orientation for treatment. In this investigation, we examine a grid comprising 60x60x60 voxels, each with a resolution of approximately 0.59 mm/voxel, similar to the grid parameters in the corresponding clinical treatment plan defined in Eyeplan.

2.1 Optimization

In the realm of OPT, optimization is the pursuit of identifying the gaze fixation point corresponding to the eye orientation with optimal sparing of healthy eye structures for the patient. This is because nozzle setting is defined in such a way that target coverage is ensured. The objective of the optimization can therefore be formulated with equation 1 as:

$$\underset{O}{\text{minimize}}, \mathscr{C}(O) \tag{1}$$

where O represents an attainable eye orientation for treatment, and $\mathscr{C}(O)$ is the cost calculated from the dose simulated by our dose engine for eye orientation O.

2.1.1 Cost Function for healthy tissue sparing

The cost function, as illustrated in equation 2, attempts to emulate the decision-making process of the planner. It calculates $\mathscr{C}$ by applying penalties to the dose deposition attributed to healthy structures of the eye.

$$\begin{aligned}\mathscr{C} = &\sum_{i=1}^{N} \frac{w_i}{M} \sum_{j=1}^{M} Dose_{ij} + \\ &w_{macula} \cdot D_{2\%_{macula}} + \\ &w_{opticdisc} \cdot D_{20\%_{opticdisc}} + \\ &w_{cornea} \cdot D_{20\%_{cornea}} + \\ &w_{lens} \cdot \max(V_{5\%_{\text{lens}}} - 10GyE, 0)\end{aligned} \tag{2}$$

Where:

$Dose_{ij}$ represents the dose-volume points for the i-th structure and j-th volume point,
N is the total number of considered eye structures,
M is the number of dose-volume points considered for each structure (e.g., V5, V10, ..., V100),
w_i is the tunable weight assigned to the i-th structure.

In formulating the cost function, a deliberate strategy has been employed to preserve healthy tissue by leveraging the ocular structures present in the clinical model. This strategy is informed by recent findings on dose-response to healthy tissue [2, 6, 7]. The first term of the cost function penalizes the area beneath the dose-volume histogram curves for the considered eye structures. Though the area itself does not correspond to a clinical outcome, it imposes penalties on dose attributed to those eye structures that are known contributors to neovascular glaucoma (NVG) [6], as well as the dose to the optic disc and macula. The second, third, and fourth terms penalize $D_{2\%_{\text{macula}}}$, $D_{20\%_{\text{optic disc}}}$, and $D_{20\%_{\text{cornea}}}$, respectively. These are based on the findings of Espensen et al. (2021), which demonstrated the predictive capability of these parameters for visual acuity deterioration, optic neuropathy, and NVG respectively [2]. The fifth term is formulated to impose a penalty on the V5 of the lens when exceeding the threshold of 10 GyRBE, as this elevation in dose is associated with an elevated risk of vision-impairing cataracts (VIC) [7]. Furthermore, the cost function imposes an indirect penalty on the dose administered to the ciliary body due to its anatomical proximity to the lens and cornea. This intentional measure is grounded in the demonstrated association between dose to the ciliary body and the onset of cataract [2].

2.1.2 Optimization procedure

This optimization procedure involves generating a randomized set of eye orientations with the target aligned to the beam axis which together define a diverse set of gaze directions to populate the gaze fixation plane. The set of orientations are then filtered by polar limit of 25 degrees and proximity to the ML prediction. The remaining orientations are subsequently assessed for treatment suitability by computing the corresponding dose to the eye structures, cost and NTCP value. The orientations under consideration form 2D maps in the fixation point plane, represented spatially by their polar and azimuth angles and color-coded to the particular costs or NTCP model under investigation.

2.2 Gaze Prediction

Calculating the dose for each eye orientation is a computationally expensive endeavor. Therefore, an ML algorithm is first employed to reduce the necessary search area in which to densely sample the space of attainable fixation points. This gaze fixation prediction function is comprised of a set of random forest regression algorithms, which have been trained on a dataset of 1730 patients previously treated at PSI. This function predicts gaze fixation points by extracting geometric features from the patient's anatomy as represented by the patient model. These features include characteristics such as target boundaries, the center of mass of the target, and the proximity of the target to critical eye structures. Visual representation of a subset of selected features, along with the patient model, are depicted in Figure 1.

3 Results

The outcomes for 12 patients are first presented, with two patient cases studied in more detail in subsection 3.1. All results for the investigated patient cases were examined by a licensed medical physicist in OPT. This analysis concluded that all identified gaze fixation spots are acceptable. The cost function can identify a region with similar costs and comparable NTCP values to the clinical fixation point selected by the planner for all examined patients. Differences in NTPC between the clinical fixation point of the fixation point of least cost position are better or within 0.13, 0.14 and 2.81 % of for all investigated patients for optic neuropathy, neovascular glaucoma and visual acuity respectively, based on the findings of Espensen et al [2]. While the examined

fixation space is not exhaustive, the limited ML-assisted cost evaluated sampling was able to identify significant findings in terms of reduced probability of radiation induced toxicities for some patients. Specifically, for optic neuropathy, two patient cases exhibited a decrease in probability of 2 %, and one case showed a reduction of 9 %. In the context of neovascular glaucoma, two patient cases demonstrated a reduction in probability of 5 and 6 %, respectively. Regarding visual acuity, one case showed a 3 % decline.

3.1 Optimization examples

Examples for two patients are shown in figures 2 and 3. In each subplot, explored fixation points are colour coded depending on their cost or probability of radiation-induced toxicity [2]. Both figures show gradients which slope towards favourable healthy tissue sparing, suggesting gaze angle modifications for reducing radiation exposure to health tissue. Specifically figure 2 showcases a potential to reduce the risk of neovascular glaucoma by approximately 5% units with a polar angle increase of 6 degrees. In Figure 3, a case is presented where a gradient of reduced probability of neovascular glaucoma is evident manifested by reduced levels of $D_{20\%}$ to the cornea, however the same change risks increasing the likelihood of visual acuity deterioration.

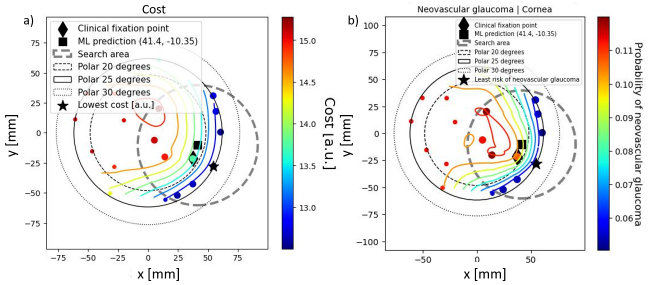


Figure 2: Investigated fixation points for the first patient example are color-coded based on their respective costs (a) or probability of neovascular glaucoma (b) [2].

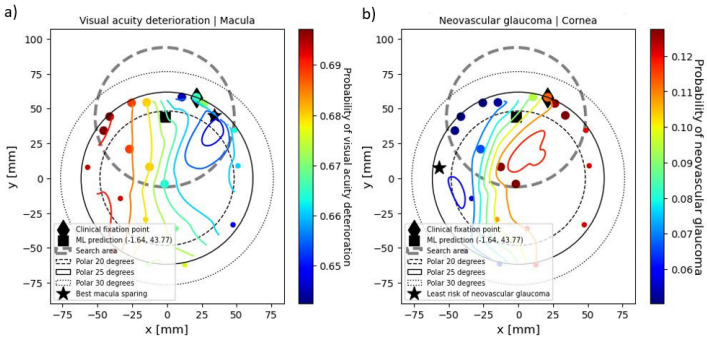


Figure 3: Investigated fixation points for the first patient example are color-coded based on their respective probability of visual acuity deterioration (a) and probability of neovascular glaucoma (b) [2].

3.2 Gaze prediction

The ML predictions closely align with the clinical orientation as shown in Figure 4, deviating by no more than 5 degrees in three quartiles of predictions for 190 patients. The black lines in the figure illustrates the proximity between the predicted fixation point and the clinical fixation point, which often display a slight underestimation of the polar angle. The observed feature importance in our ML algorithm is primary driven by the centre of mass of the target in relation to the centre and rotation point of the model in the gaze centered reference frame.

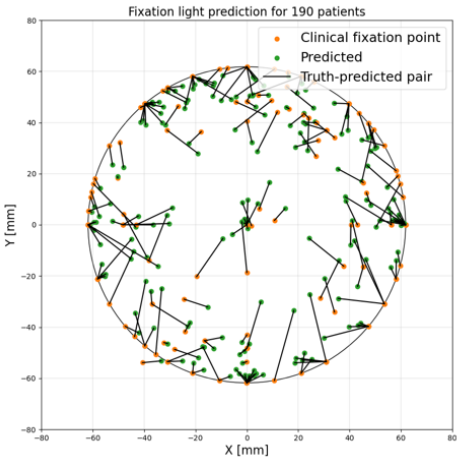


Figure 4: ML predictions for 190 patients and their corresponding clinical fixation point connected with a black line indicating correspondence.

4 Discussion

We have demonstrated that our algorithm can optimize healthy tissue sparing for ocular proton therapy patients by expressing specific penalties associated with the onset of radiation induced toxicities to healthy tissue using a cost function and NTCP models. Identified fixation points of the optimization algorithm represent a theoretical close to optimal orientation expressed in terms of cost or NTCP value. However, its clinical attainability may be hindered by clinical constraints such as avoiding to treat through an eyelid or the patient's ability to gaze according to the treatment plan. The continuous gradients given by the cost function or parameterized NTCP dose-response curves, allow the planner to track iso-lines of similar healthy tissue sparing in order to to strike a balance between sparing healthy tissue and ensuring clinical feasibility.

Additionally we have demonstrated capacity to consistently predict a fixation point near to the clinically defined point, signifying a reduction in the computational time necessary for the optimization algorithm to identify a suitable fixation point. The required optimization time can vary, typically taking a few minutes, depending on the user's choice for terminating the simulation script. However, it's important to note that the

prediction, while informed by historical patient data at PSI, does not guarantee optimality from an NTCP perspective. The ML predictions, nevertheless, efficiently constrain the search area, eliminating the need to explore all potential gaze fixation points, a methodology employed in previous studies by [8]. Figure 4 illustrates the ML algorithm's capacity to learn from the dataset. The feature importance indicates that the predominant information influencing the predictions is the spatial data concerning the horizontal and vertical aspects of the target center of mass in relation to the centre of the model. The observed learning reflects an understanding of the planner's intent to choose a gaze that minimizes dose exposure to the anterior segments of the eye by rotating them out of the treatment beam. Additionally, the learning captures the planner's intention to rotate distal targets closer to the surface, thereby reducing the proton range. It also seems to recognize that certain superficial targets in proximity to the cornea are treated with a gaze directly aligned with the axis of the treatment beam.

Our methodology refines the previous work by [8] by introducing a tunable cost function and mapping of NTCP curves. Both elements contribute to providing gradients of healthy tissue sparing. This approach preserves a more nuanced understanding of the patient positioning compared to the binary choices for each structure suggested by Henning et al. Additionally, our approach features a more accurate dose calculation engine, and achieves faster results despite employing a finer voxel grid by incorporating ML predictions based on patient anatomy.

Among the examined patient cases, NTCP value gradients of diminishing risks of radiation induced toxicities have been identified for 4 out of 12 patient cases next to the clinical fixation point, suggesting potential strategies for mitigating side effects of the treatment. One identified gradient suggests that adjusting the polar angle by 6 degrees could reduce the probability of neovascular glaucoma by 5% due to reduced exposure to the cornea. It should however be considered that small reductions in probability of radiation-induced toxicities fall within the uncertainty of the NTCP models. For half of the examined patients, no discernible benefits, NTCP changes, or cost gradients in proximity to the clinical orientation were identified. However, the cost function identifies an iso-area of similar cost in proximity to the clinical fixation point, with potential utilization for adapting treatment plans.

5 Conclusion

This work suggests that the next-generation OPT treatment planning systems could be equipped with NTCP models and adjustable expressions of cost functions to provide guidance for the treatment planner to optimize OPT treatment plans.

References

[1] M. Goitein and T. Miller. "Planning proton therapy of the eye". *Medical Physics* 10 (3 1983). DOI: 10.1118/1.595258.

[2] C. A. Espensen, J. F. Kiilgaard, A. L. Appelt, et al. "Dose-Response and Normal Tissue Complication Probabilities after Proton Therapy for Choroidal Melanoma". *Ophthalmology* 128 (1 2021). DOI: 10.1016/j.ophtha.2020.06.030.

[3] N. C. Koch and W. D. Newhauser. "Development and verification of an analytical algorithm to predict absorbed dose distributions in ocular proton therapy using Monte Carlo simulations". *Physics in Medicine and Biology* 55 (3 2010). DOI: 10.1088/0031-9155/55/3/019.

[4] H.-P. Wieser, E. Cisternas, N. Wahl, et al. "Development of the open-source dose calculation and optimization toolkit matRad". *Medical Physics* 44.6 (2017), pp. 2556–2568. DOI: 10.1002/mp.12251.

[5] E. Learmonth. *Validation and Development of a Treatment Planning System for Proton Therapy of Intraocular Tumours*. Master's thesis. Glasgow, Scotland, Jan. 2024.

[6] K. K. Mishra, I. K. Daftari, V. Weinberg, et al. "Risk factors for neovascular glaucoma after proton beam therapy of uveal melanoma: A detailed analysis of tumor and dose-volume parameters". *International Journal of Radiation Oncology Biology Physics* 87 (2 2013). DOI: 10.1016/j.ijrobp.2013.05.051.

[7] J. Thariat, S. Jacob, J. P. Caujolle, et al. "Cataract avoidance with proton therapy in ocular melanomas". *Investigative Ophthalmology and Visual Science* 58 (12 2017). DOI: 10.1167/iovs.17-22557.

[8] F. Hennings, A. Lomax, A. Pica, et al. "Automated Treatment Planning System for Uveal Melanomas Treated With Proton Therapy: A Proof-of-Concept Analysis". *International Journal of Radiation Oncology*Biology*Physics* 101 (3 July 2018), pp. 724–731. DOI: 10.1016/J.IJROBP.2018.02.008.

Neural Graphics Primitives-based Deformable Image Registration for On-the-fly Motion Extraction

Xia Li[1,2], Fabian Zhang[2], Muheng Li[1,3], Damien Weber[1,4,5], Antony Lomax[1], Joachim Buhmann[2], and Ye Zhang[1]

[1]Center for Proton Therapy, Paul Scherrer Institut, Villigen, Switzerland
[2]Department of Computer Science, ETH Zurich, Zurich, Switzerland
[3]Department of Physics, ETH Zurich, Zurich, Switzerland
[4]Department of Radiation Oncology, University Hospital of Zurich, Zurich, Switzerland
[5]Department of Radiation Oncology, Inselspital, Bern University Hospital, University of Bern, Bern, Switzerland

Abstract

Intra-fraction motion in radiotherapy is commonly modeled using deformable image registration (DIR). However, existing methods often struggle to balance speed and accuracy, limiting their applicability in clinical scenarios. This study introduces a novel approach that harnesses Neural Graphics Primitives (NGP) to optimize the displacement vector field (DVF). Our method leverages learned primitives, processed as splats, and interpolates within space using a shallow neural network. Uniquely, it enables self-supervised optimization at an ultra-fast speed, negating the need for pre-training on extensive datasets and allowing seamless adaptation to new cases. We validated this approach on the 4D-CT lung dataset DIR-lab, achieving a target registration error (TRE) of 1.15 ± 1.15 mm within a remarkable time of 1.77 seconds. Notably, our method also addresses the sliding boundary problem, a common challenge in conventional DIR methods.

1 Introduction

Deformable image registration (DIR) is a critical technique in radiotherapy, essential for accurate motion modeling, dose accumulation [1], and image alignments. The complexity of modeling the displacement vector field (DVF) and the absence of a definitive ground truth make DIR a challenging endeavor. The evolution of DIR methodologies has transitioned from foundational techniques such as optical flow and elastic models [2] to more sophisticated approaches like the demon algorithm [3] and B-spline-based methods [4]. These advancements, including the adoption of map-based approaches, have significantly enhanced precision in capturing complex image deformations. However, they are limited by their computational time demands and the need for precise initial alignment and parameter settings.

The introduction of deep learning (DL) into DIR has been a game-changer. Notable studies like VoxelMorph [5] have pioneered the use of unsupervised learning frameworks employing convolutional neural networks (CNNs). Despite their advances, DL-based methods still suffer from the need for extensive training datasets and face challenges in generalizing to diverse clinical scenarios. Besides, CNNs are sensitive to the input resolution (physical spacing). More recently, Implicit Neural Representations (INR) [6] have emerged, combining the benefits of case-specific optimization with neural network efficiency. Innovations such as IDIR [7], and ccIDIR [8] have collectively pushed the boundaries in DIR, improving processing speed and accuracy. However, they still fall short of achieving real-time performance, using at least 15 seconds for one case.

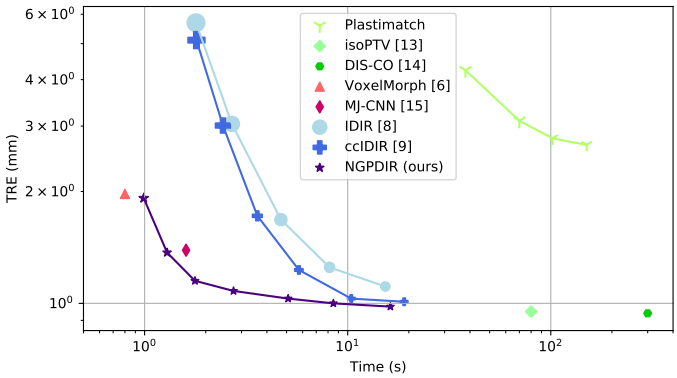


Figure 1: Landmark error, averaged over all cases, for different DIR methods. Conventional and DL-based methods are denoted in green and red, respectively, with reported values from the literature. INR/NGP based (blue) have been implemented by us and exclusively trained on the lung region for comparison.

In our work, we introduce the application of Neural Graphics Primitives (NGP) [9] for DIR, specifically adapted for 4D-CT scans in the DIR-lab dataset. By leveraging NGP for optimizing DVF, our approach combines efficiency and accuracy, utilizing learned primitives and a shallow neural network for rapid deformation field estimations. This method not only represents a significant advancement in DIR by offering quick estimations and maintaining high accuracy but also shows substantial promise for real-time clinical applications. Our preliminary findings demonstrate notable improvements in landmark accuracy, image warp accuracy, and alignment of organs-at-risk (OARs) masks, addressing challenges such as sliding boundaries. In this work, the capabilities of NGP for DIR have been assessed by applying it to the deformation of exhalation to inhalation CT data, extracted from 4DCT studies in the DIR-lab dataset [10].

2 Materials and Methods

2.1 Dataset

In this study, we employed DIR-lab that comprises 4D-CT images for the radiotherapy of thoracic malignancies. This

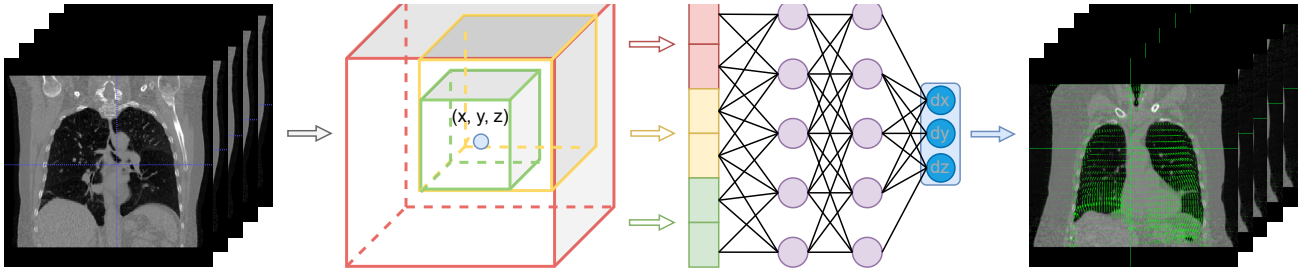


Figure 2: Schematic of the NGPDIR framework. A random point (x, y, z) is represented as the interpolations of multi-level primitives, then mapped to the displacement vector by a small network. The primitives and the network are optimized with respect to each image.

dataset includes 10 cases, from which the exhalation and inhalation phases have been extracted for this study, with the inhalation images as the fixed images. We preserved the original resolution of these 4D-CT scans, which varies across cases, and refrained from normalizing intensity values, as our proposed method is resolution-agnostic and value-agnostic.

2.2 Method

2.2.1 INR for DIR

The preliminary approach of our method is IDIR [7], which adopts INR for DIR. INR fundamentally functions by mapping 3D coordinates to corresponding values or intensities by a learnable neural network, which is a crucial aspect of reconstructing and manipulating complex spatial data. In the context of DIR, INR is adapted to map these 3D coordinates to displacement vectors $(\delta x, \delta y, \delta z)$, constituting the DVF to warp the moving image (exhale phase) towards the fixed image (inhale phase). The training process of INR within DIR involves random sampling of points within the 3D space, followed by the computation of the Normalized Cross-Correlation (NCC) loss between the warped image and the fixed image, together with a Jacobian regularization. Unlike DL-based methods that require training on large-scale datasets, INR-based DIR only needs to fit one network per case, so it combines the advantages of instance optimization with neural representation.

2.2.2 NGP for DIR

The multi-resolution encoding in NGP employs a hierarchical hash table mechanism to represent 3D coordinates at varying levels of detail, which can be formulated as follows:
Given a 3D coordinate $\mathbf{x} = (x, y, z)$, the encoding process is defined by a set of hashing functions $\{h_1, h_2, ..., h_L\}$, where each h_i corresponds to a different resolution level i. These functions map $\mathbf{x}$ to a series of indices that reference feature vectors within the respective hash tables:

$$\mathbf{f}_i = \text{HashTable}_i[h_i(\mathbf{x})], \tag{1}$$

where $\mathbf{f}_i$ is the feature vector retrieved from the i-th hash table. The final feature representation $\mathbf{F}(\mathbf{x})$ for the coordinate $\mathbf{x}$ is then obtained by concatenating the feature vectors across all resolution levels:

$$\mathbf{F}(\mathbf{x}) = \bigoplus_{i=1}^{L} \mathbf{f}_i, \tag{2}$$

where $\bigoplus$ denotes the concatenation operation. The feature representation $\mathbf{F}(\mathbf{x})$ is then passed through the neural network to predict the displacement vector $\mathbf{v}$:

$$g_\phi(\mathbf{F}(\mathbf{x})) \rightarrow \mathbf{v} = (\delta x, \delta y, \delta z). \tag{3}$$

Through this multi-level representation, NGP can manage different scales of spatial features, which is particularly beneficial for DIR where capturing both global anatomical structure and local tissue details is crucial.

2.2.3 Implementation details

For fair comparisons, we have re-implemented both IDIR and ccIDIR in a shared code base with NGPDIR, which will be made public. For training, our batch size was set at $10,000$ points, randomly sampled for each iteration. Specifically, we only used points within the lung for landmark evaluation and the body for MAE calculation. For IDIR and ccIDIR, we keep their original learning rate (LR) as $1e-4$, while for our method, we employed a higher LR of $1e-2$. Our training regimen included a warmup phase for the learning rate, followed by a cosine schedule that gradually reduced it to zero towards the end of training. To analyze the relationship between training time and accuracy, we trained each method for a range of steps: [60, 125, 250, 500, 1000, 2000, 4000]. All experiments were conducted on an RTX 4090 GPU.

2.3 Evaluation

Within the DIR-Lab dataset, the physician's labeled landmarks were provided as the ground truth. We further adopted Totalsegmentor [1] to get masks for sliding boundary regions like vertebras and ribs. Evaluations were done upon landmarks using the target registration error (TRE), and also on image difference between the fixed image and the warped one using mean absolute error (MAE). Sliding boundary areas are evaluated by Dice Coefficients and MAE. Furthermore, we visually compared the predicted motion, DVF as well as warped images from different methods.

[1] https://totalsegmentator.com/

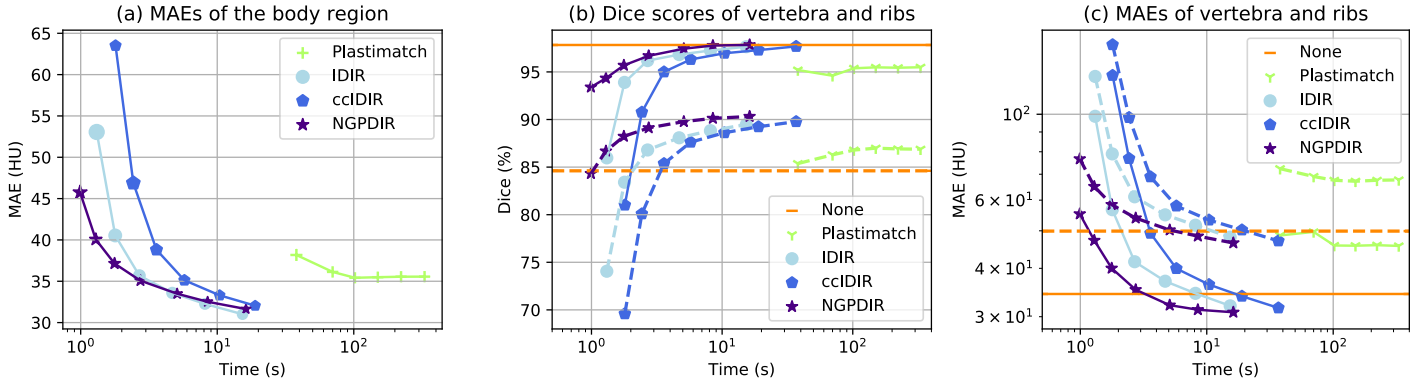
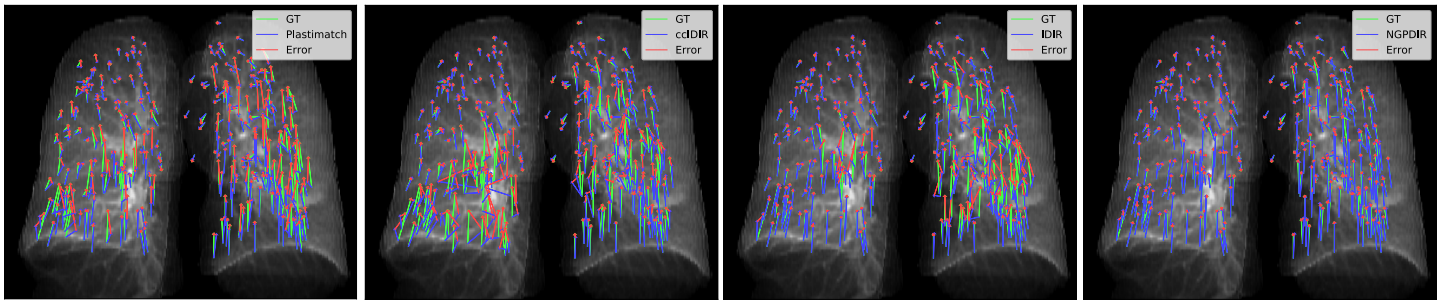


Figure 3: MAE and Dice scores for the different DIR methods averaged over all cases. Orange lines in sub-figures (b-c) represent the absence of registration as a baseline. Solid lines indicate results for the vertebra, while dashed lines correspond to ribs.



(a) Motion from Plastimatch (b) Motion from ccIDIR (c) Motion from IDIR (d) Motion from NGPDIR

Figure 4: Visualization of landmark movement for case 8. Ground-truth motions are denoted by green arrows, predictions by blue, and errors by red. The magnitude of errors corresponds to the length of the red arrows. These projections are in the coronal plane for clarity.

3 Results

Landmark prediction errors averaged over all 10 cases of DIR-Lab are depicted in Figure 1. Both IDIR, ccIDIR, NGPDIR, and Plastimatch [11] trained on the lung region with different steps are plotted as curves. Besides, we also incorporate conventional methods (isoPTV [12] and DIS-CO [13]) and DL-based methods (VoxelMorph [5] and MJ-CNN [14]) for comparisons. From the figure, conventional methods have the lowest landmark errors, while DL-based methods exceed in speed. Our NGPDIR achieved the best trade-off by far between speed and accuracy.

Other metrics are plotted in Figure 3, where the same trend can be observed. Notably, the conventional B-spline-based method (Plastimatch) cannot handle the sliding boundary regions well, achieving even worse results than the baseline (no registration), while NGPDIR quickly exceeds the baseline after training for only 1.2 seconds. Besides, ccIDIR displayed slower convergence compared to IDIR when trained on the entire body, contravening the trend in Figure 1.

The landmark motion visualizations in Figure 4 further depict NGPDIR's precision, with the shortest error vectors. Moreover, NGPDIR's DVF is notably smoother and more accurate, particularly in the most challenging case (Case No. 8), highlighted in Figure 5. Meanwhile, it surpasses all the others in the error map, outperforming Plastimatch significantly around the ribs (pointed by green arrows).

4 Discussion

In our pursuit of advancing DIR, we draw inspiration from the dynamic realm of 3D reconstruction techniques. Notably, IDIR has leveraged the foundational principles of INR and we harness the computational efficiency derived from Instant-NGP. Given that both INR and NGP are per-point query methodologies, their natural extension into DIR estimation involves querying all voxels within a volume. Further, this can be extended volume grids, bestowing our approach with the valuable property of resolution independence. This unique trait empowers our approach with consistent performance across images of varying sizes, distinguishing it from conventional DL-based methods, which can be sensitive to input resolution and size.

Our comparative analysis reveals the remarkable versatility and efficiency of this approach when compared against both Deep Learning (DL)-based and conventional DIR methods. By effectively eliminating the need for extensive datasets while excelling in per-case optimization, our approach distinguishes itself from DL-based methods. Furthermore, our utilization of neural networks and hash encoding techniques expedites convergence, presenting an appealing alternative to conventional DIR methods.

Turning our attention to the broader implications of our work, we recognize that the current limitation of implementing INR/NGP for DIR lies in their need for case-by-case fitting. As we look forward, there is exciting potential for further

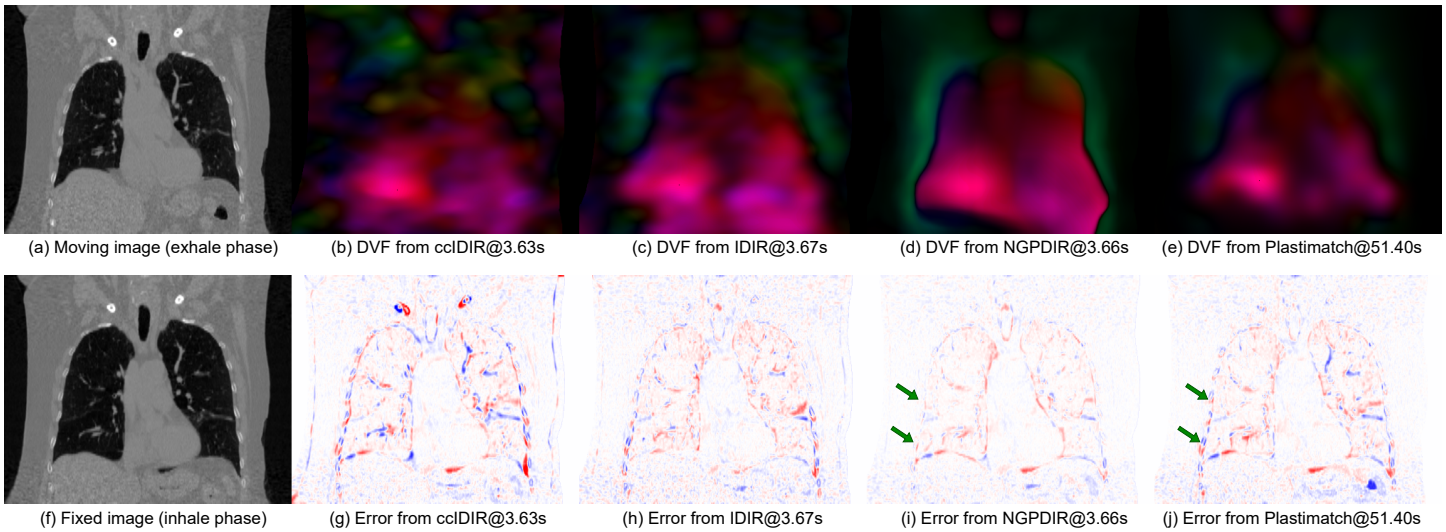


Figure 5: Example DVFs and error maps for case 8. ccIDIR, IDIR, and NGPDIR are compared with similar training times (around 3.65s). DVFs are color-coded for directionality, with the intensity reflecting the magnitude. Green arrows point to ribs (sliding boundaries).

generalization through meta-learning. This approach has the capacity to learn a general prior for specific human regions, followed by case-specific information fitting.

While we propose a fast and efficient DIR approach, it's important to acknowledge the potential bottlenecks that may hinder its real-time application. Challenges persist in the realms of image acquisition hardware and image reconstruction algorithms. Improved acquisition hardware and faster reconstruction algorithms are needed to fully realize the potential of our accelerated DIR method. Additionally, the evolving landscape of 3D reconstruction technologies holds promise, offering solutions to address the later challenges. Beyond real-time image guidance, the rapid speed of our DIR approach opens the door to a range of applications in radiotherapy, including adaptive treatment planning, organ motion tracking, and dose optimization to name a few.

5 Conclusion

In this study, we have successfully integrated NGP into DIR, a novel contribution that significantly enhances the accuracy and efficiency of medical image alignment as demonstrated on the DIR-lab dataset. The NGPDIR framework exhibits robust performance across various metrics, particularly in landmark alignment precision and the accommodation of anatomical sliding boundaries. This advancement not only propels the DIR field forward but also opens new avenues for real-time clinical applications, potentially transforming patient care with its rapid, reliable imaging capabilities.

References

[1] I. J. Chetty and M. Rosu-Bubulac. "Deformable registration for dose accumulation". *Semin. Radiat. Oncol*. Vol. 29. 3. Elsevier. 2019, pp. 198–208.

[2] R. Bajcsy and S. Kovačič. "Multiresolution elastic matching". *Computer vision, graphics, and image processing* 46.1 (1989), pp. 1–21.

[3] T. Vercauteren, X. Pennec, A. Perchant, et al. "Diffeomorphic demons: Efficient non-parametric image registration". *NeuroImage* 45.1 (2009), S61–S72.

[4] D. Rueckert, P. Aljabar, R. A. Heckemann, et al. "Diffeomorphic registration using B-splines". *MICCAI*. Springer. 2006, pp. 702–709.

[5] G. Balakrishnan, A. Zhao, M. R. Sabuncu, et al. "VoxelMorph: a learning framework for deformable medical image registration". *IEEE Trans. Med. Imaging* 38.8 (2019), pp. 1788–1800.

[6] V. Sitzmann, J. Martel, A. Bergman, et al. "Implicit neural representations with periodic activation functions". *NeurIPS* 33 (2020), pp. 7462–7473.

[7] J. M. Wolterink, J. C. Zwienenberg, and C. Brune. "Implicit neural representations for deformable image registration". *MIDL*. PMLR. 2022, pp. 1349–1359.

[8] L. D. Van Harten, J. Stoker, and I. Išgum. "Robust deformable image registration using cycle-consistent implicit representations". *IEEE Trans. Med. Imaging* (2023).

[9] T. Müller, A. Evans, C. Schied, et al. "Instant neural graphics primitives with a multiresolution hash encoding". *ACM Trans. Graph* 41.4 (2022), pp. 1–15.

[10] R. Castillo, E. Castillo, R. Guerra, et al. "A framework for evaluation of deformable image registration spatial accuracy using large landmark point sets". *Phys. Med. Biol.* 54.7 (2009), p. 1849.

[11] G. C. Sharp, R. Li, J. Wolfgang, et al. "Plastimatch: an open source software suite for radiotherapy image processing". *ICCR*. 2010.

[12] M. Tancik, P. Srinivasan, B. Mildenhall, et al. "Fourier features let networks learn high frequency functions in low dimensional domains". *NeurIPS* 33 (2020), pp. 7537–7547.

[13] J. Rühaak, T. Polzin, S. Heldmann, et al. "Estimation of large motion in lung CT by integrating regularized keypoint correspondences into dense deformable registration". *IEEE Trans. Med. Imaging* 36.8 (2017), pp. 1746–1757.

[14] Z. Jiang, F.-F. Yin, Y. Ge, et al. "A multi-scale framework with unsupervised joint training of convolutional neural networks for pulmonary deformable image registration". *Phys. Med. Biol.* 65.1 (2020), p. 015011.

Robust optimization for contouring uncertainties in adaptive radiation therapy

I Bengtsson[†1,2], A Smolders[†3,4], A Forsgren[1], A Lomax[3,4], DC Weber[3,5,6], A Fredriksson[2], and F Albertini[3]

[1]Department of Mathematics, KTH Royal Institute of Technology, Stockholm, Sweden
[2]RaySearch Laboratories AB, Stockholm, Sweden
[3]Paul Scherrer Institute, Center for Proton Therapy, Switzerland
[4]Department of Physics, ETH Zurich, Switzerland
[5]Department of Radiation Oncology, University Hospital Zurich, Switzerland
[6]Department of Radiation Oncology, Inselspital, Bern University Hospital, University of Bern, Switzerland

Abstract Online adaptive radiation therapy requires fast and automated contouring of daily scans for treatment plan re-optimization. However, automated contouring is imperfect and introduces contour uncertainties. This work aims at developing and comparing robust optimization strategies accounting for such uncertainties. A deep-learning method was used to predict the uncertainty of deformable image registration, and to generate a finite set of daily contour samples. Eight optimization strategies were compared: two baseline methods, three methods that convert contour samples into voxel-wise probabilities, and three methods accounting explicitly for contour samples as scenarios in robust optimization. Target coverage and organ-at-risk (OAR) sparing were evaluated robustly for simplified proton therapy plans for five head-and-neck cancer patients. We found that the probability-based methods covered the target with better OAR sparing than the baseline methods, without increasing the optimization time. The scenario-based methods spared the OARs even more, but increased integral dose and required more computation time.

1 Introduction

Radiation therapy has progressively evolved toward greater dose conformality with the tumor, resulting in better healthy tissue sparing. However, uncertainties arising from daily setup variations and anatomical changes limit further conformality improvements as they could lead to substantial target underdosage. Online adaptive radiation therapy aims to reduce such uncertainties by the acquisition of new image data shortly before delivery and adaptation of the treatment to the current anatomy [1, 2]. Adaptation requires acquisition of up-to-date contours of the *clinical target volumes* (CTVs) and the *organs at risk* (OARs) on the new image, as well as re-optimization of the treatment plan. Both processes need to be fast, to maximize the correspondence between the new image and the treated anatomy. This time limitation makes manual contouring infeasible and automated techniques necessary.
One promising technique is *deformable image registration* (DIR) [3], which can be used to propagate the contours from the planning CT to the daily image. However, DIR is inherently uncertain, which results in uncertainty in the propagated contours. To use DIR, it is therefore necessary to quantify and account for its uncertainty. Recent research has introduced a model to predict the uncertainty associated with a DIR algorithm [4], which can efficiently generate samples of propagated contours.

[*]These authors contributed equally to this work

The resulting contour uncertainty may be explicitly included in the re-optimization. Planning using uncertain target volumes has recently gained interest, and several optimization techniques have been proposed for use with a *clinical target distribution* (CTD) rather than with a CTV [5–7]. In the context of adaptive therapy, managing uncertainties by considering contours propagated by multiple DIRs was recently proposed [8]. However, increased target robustness comes at a price. Efficient mitigation of contour uncertainty therefore requires optimization algorithms that favorably navigate the trade-offs between target coverage and healthy tissue sparing. In this work, we develop and compare different strategies for robust optimization against contour uncertainty, in a statistically consistent framework. Specifically, we consider online adaptive re-optimization for head-and-neck proton therapy and employ the DIR uncertainty prediction model from [4].

2 Materials and methodology

2.1 Contour uncertainty model

In this work, DIR is used to propagate the contours from the planning CT to daily CTs. Contour uncertainty is modeled with a previously developed deep-learning DIR uncertainty model [4]. The model predicts uncertainty based on image similarity and contrast. Specifically, it predicts the standard deviation of a Gaussian *deformable vector field* (DVF), with mean given by the output of a known DIR algorithm. The resulting Gaussian DVF can be sampled, and each sample can warp contours to the daily image, yielding a finite set of daily contours. It was shown in [4] that the resulting contour distribution represents well the contouring uncertainty for head and neck patients by comparison with manual contours.

2.2 Robust optimization techniques

In clinical practice, plans are made robust by use of margins or by optimization of the worst case from a variety of range and set-up scenarios. This implicitly mitigates some of the contour uncertainty and we therefore employ two commonly used techniques to benchmark our results. Range uncertainty is treated consistently in all methods by including three range scenarios $(0\%, \pm 3\%)$ in the robust optimization.

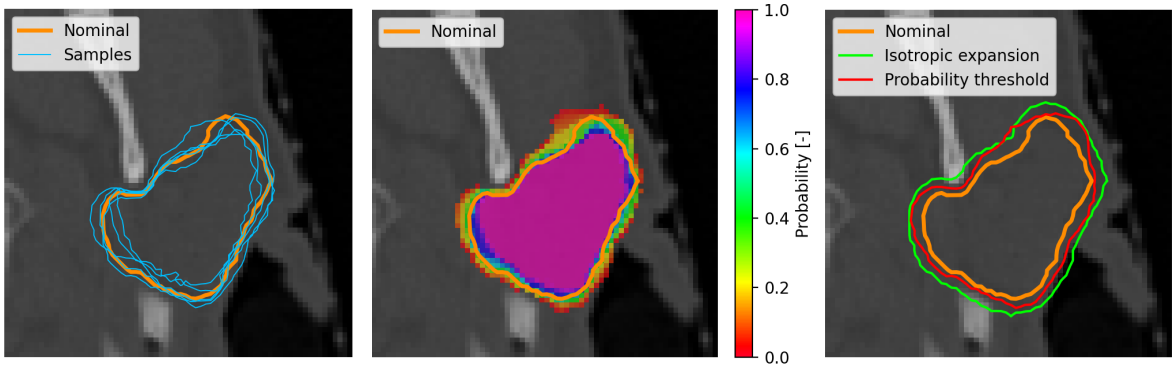


Figure 1: Visualization of contour samples with related marginal probabilities and derived non-isotropic margins. (Left) Nominal contour together with 5 contour samples; (middle) nominal contour overlayed on the voxel-wise marginal probabilties derived from the 5 contour samples; (right) nominal contour together with a 3 mm isotropic expansion and the contour related to the 20% probability threshold.

1. *Set-up robust*: These plans are optimized with isotropic set-up uncertainty (1, 3, or 5 mm) on the nominal contours, i.e., the contours warped with the mean DVF.

2. *Isotropic margin*: These plans are optimized on isotropically expanded (1, 2, or 3 mm) nominal contours.

The next set of methods all rely on converting the contour samples into voxel-wise marginal probabilities, by counting for each voxel the frequency with which it belongs to the contour (Fig. 1). No set-up uncertainty is considered explicitly by any of the following methods.

3. *Probability weighted*: These plans are optimized by scaling the objective weights with the voxel-wise probabilities, resulting in a higher importance for voxels that are more likely to be part of the contour. Additionally, we define the *expected volume* of a structure as the sum of its voxels' marginal probabilities, and use it when imposing dose-volume objectives, e.g., a D30 objective neglects 30% of the expected contour volume. This allows the optimizer to either neglect many voxels with low probability, or fewer with higher probability.

4. *Probability threshold*: These plans are optimized by only including voxels in the contour that have a probability above a certain threshold (1%, 5%, 15%, or 25%). This is similar to adding a margin around the nominal contour, but this margin is not isotropic and depends on where the contour is uncertain (Fig. 1).

5. *Expected volume objectives*: These plans replace any total-volume objectives – such as the minimum-dose objective on the target – with dose-volume objectives considering the expected volume. This allows the optimizer to neglect part of structures (99.4%, 99.6% or 99.8 %) corresponding to the replaced objectives based on their expected volumes.

Our final set of methods extends the scenario definition to also include the contours, by considering each contour sample as a plausible realization of the contour set. Each combination of a contour sample and a range error then constitutes a scenario that can be used in optimization.

6. *Worst-case scenario*: These plans are optimized by minimizing the maximum value of the objective function over the set of scenarios.

7. *Expected scenario*: These plans are optimized by minimizing the expected value of the objective function over the set of scenarios.

8. *Near-worst-case scenario*: These plans are optimized by, at every iteration, ignoring the fraction of scenarios (5%, 15%, or 25%) with the worst objective values. The maximum is then taken over the remaining scenarios.

2.3 Data

The dataset used in this work contains 5 patients with head and neck cancer. It was previously described in [3] and has also been used to validate the DIR uncertainty model [4]. Each patient had one planning CT and 4–7 repeated CTs, considered as samples of daily CTs for an adaptive treatment. Each repeated CT was registered to their respective planning CT using RayStation Anaconda [9] (RaySearch Laboratories, Stockholm, Sweden) and the DIR uncertainty model was used to predict the standard deviations of these DVFs. The resulting Gaussian DVFs were subsequently sampled 200 times yielding 200 contour samples. Of those, 100 were used in the optimization, and 100 for evaluation.
We created a simple planning template with a robust objective on the CTV and non-robust objectives on six OARs: optic nerves, brainstem, parotids and thyroid. These OARs were chosen because of their proximity to the CTV, to cover serial and parallel OAR types, and to cover both regions with high and low image contrast. All treatment plans used IMPT with the same 3 beams at gantry angles $\pm$65 and 180 with 1 cm range shifter. The plans aimed to deliver 54 GyRBE to the CTV in 33 fractions. Maximum dose objectives were used

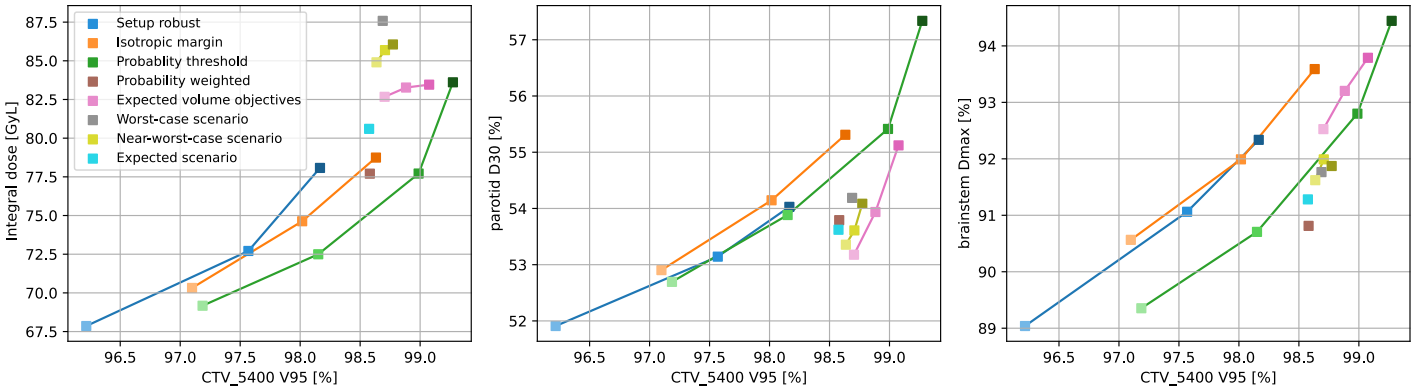


Figure 2: Mean nominal integral dose, parotid D30 and brainstem Dmax as a function of mean near-worst-case CTV coverage for the different robust optimization strategies. For each method, the markers are more transparent for the lower robustness settings (e.g. 1 mm margin) and more opaque for higher robustness (e.g. 3 mm margin).

for the optic nerves and the brainstem, whilst D30 objectives were used for the parotids and the thyroid. The dose limit in each objective function was chosen in such a way that it competed with the target objective, to highlight the effect in the optimization. In addition, a dose-fall-off objective was used to limit the dose anywhere outside the target. All plans were optimized for 50 iterations using a research version of RayStation.

2.4 Robust plan evaluation

To compare the methods, we evaluate the *near-worst-case* target coverage, *on average*: we take for each patient-scan pair the 5th percentile worst-case CTV V95% over the contour and range scenarios and compute the mean of these values over the patients and scans. Instead, the OAR metrics and integral dose were evaluated in the nominal scenario, since their objectives were not treated robustly in the optimization.

3 Results

To compare the methods, we inspect the trade-off between target coverage and three of the dose-limiting metrics (Fig. 2). Increasing the robustness parameter for any method always improved the near-worst-case target coverage. However, this robustness came at the cost of increased integral and OAR dose. In any of the plots, a position toward the lower right indicates efficient target coverage in relation to the shown dose-limiting metric. In this regard, the probability threshold method accounts for most of the efficient set of plans regarding integral dose. In addition, it also more efficiently spares the brainstem when compared with the isotropic expansion method and the setup-uncertainty method. These results suggest that an informed selection of the margin based on the probabilities outperforms the benchmark methods.

The probability-weighted method resulted in a slight increase in integral dose but less dose to OARs, when compared to the frontier formed by the probability-threshold method. This observation, that the use of probability weights prioritizes OAR-sparing over limiting integral dose when compared to the use of a margin, is consistent with previous results [6, 7]. In comparison, the plans using expected volume objectives further improved the trade-off between target coverage and parotid sparing, at the expense of brainstem sparing and integral dose.

The (near-) worst-case scenario methods spared the OARs more efficiently than the probability-threshold method but led to substantially higher integral doses. This is presumably because scenarios that conflict with OARs are likely the worst in terms of the objective value, and that trade-offs with the target coverage in affected regions will be prioritized during optimization. Although this leads to better OAR sparing, it also results in lower importance elsewhere leading to a higher integral dose. The effect is illustrated by considering dose differences between instances of the probability-threshold plan and the near-worst-case plan (Fig. 3). In contrast, the same reasoning does not apply when optimizing the expected scenario, as each scenario is consistently prioritized equally. Rather, it is theoretically similar to probability weights, which is reflected in the dosimetrically similar trade-offs between the methods.

Notably, whilst the other methods require approximately the same computation time ($\sim$100 s for 50 iterations), scenario-based methods require substantially more. The increase from 3 range scenarios to 300 composite scenarios resulted in an increase of a factor 10 in computation time.

4 Discussion

We integrated DIR-induced contour uncertainty with robust optimization in an adaptive re-planning framework for radiation therapy. It is evident from the results that accounting for the uncertainty even in simple ways, such as basing margin thresholds or voxel-wise weights on the probabilities, improves the trade-offs between robust target coverage and nominal dose-limiting objectives. Further, the trade-offs

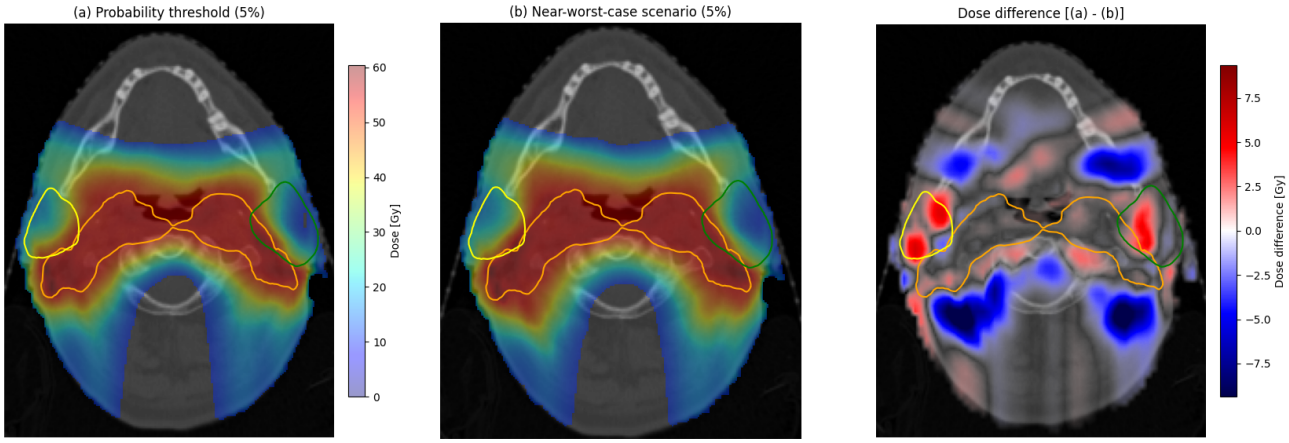


Figure 3: Doses for CT 2 of patient 3. Left: Probability threshold (5%). Center: Near-worst-case scenario (5%). Right: Dose difference. The 5% probability-threshold contours are shown for the CTV (orange), left parotid (green), and right parotid (yellow).

with OARs can be made efficient by employing scenario- or expected volume-based methods, that implicitly prioritize certain subvolumes of the target. Neglecting a fraction corresponding to the worst scenarios is much different from ignoring the same fraction of least probable voxels.
Similarly to as in [6, 7], the methods were evaluated with samples drawn from the same distribution as those used in optimization, rather than evaluating on a set of ground-truth manual contours. This was chosen because the main goal of this work is to compare the optimization algorithms, rather than to test the accuracy of the DIR uncertainty model.
Lastly, target robustness comes at the cost of increased integral and OAR dose. In this work, we have not accounted for the uncertainty in the propagated OAR contours. With the DIR uncertainty model, it is straightforward to generate the corresponding uncertainty of the OARs. It is then possible to address whether the methods can simultaneously increase the robustness of both the target and the OARs, which is the focus of an ongoing investigation.

5 Conclusion

Robust optimization using the contour uncertainty information more efficiently provides target coverage robustness compared to benchmark methods. In addition, different methods prioritize OAR sparing and limiting integral dose differently. Given the increased computational effort required for scenario-based methods, methods using thresholds or objective weights based on voxel-wise probabilities are a practical alternative in online adaptive therapy.

Acknowledgements

This project has received funding from the European Union's Horizon 2020 Marie Skłodowska-Curie Actions under Grant Agreement No. 955956.

References

[1] S. Lim-Reinders, B. M. Keller, S. Al-Ward, et al. "Online Adaptive Radiation Therapy". *International Journal of Radiation Oncology*Biology*Physics* 99.4 (2017), pp. 994–1003. DOI: https://doi.org/10.1016/j.ijrobp.2017.04.023.

[2] F. Albertini, M. Matter, L. Nenoff, et al. "Online daily adaptive proton therapy". *The British journal of radiology* 93.1107 (2020), p. 20190594.

[3] A. Smolders, E. Choulilitsa, K. Czerska, et al. "Dosimetric comparison of autocontouring techniques for online adaptive proton therapy". *Physics in Medicine & Biology* 68.17 (Sept. 2023), p. 175006. DOI: 10.1088/1361-6560/ace307.

[4] A. J. Smolders, A. J. Lomax, D. C. Weber, et al. "Deep learning based uncertainty prediction of deformable image registration for contour propagation and dose accumulation in online adaptive radiotherapy". *Physics in Medicine & Biology* (Oct. 2023). DOI: 10.1088/1361-6560/ad0282.

[5] N. Shusharina, D. Craft, Y.-L. Chen, et al. "The clinical target distribution: a probabilistic alternative to the clinical target volume". *Physics in Medicine and Biology* 63.15 (July 24, 2018), p. 155001. DOI: 10.1088/1361-6560/aacfb4.

[6] G. Buti, K. Souris, A. Maria Barragán Montero, et al. "Introducing a probabilistic definition of the target in a robust treatment planning framework". *Physics in Medicine and Biology* 66.15 (July 22, 2021). DOI: 10.1088/1361-6560/ac1265.

[7] I. Bengtsson, A. Forsgren, and A. Fredriksson. "Implications of using the clinical target distribution as voxel-weights in radiation therapy optimization". *Physics in Medicine & Biology* 68.9 (May 2023), p. 095005. DOI: 10.1088/1361-6560/acc77b.

[8] L. Nenoff, G. Buti, M. Bobić, et al. "Integrating Structure Propagation Uncertainties in the Optimization of Online Adaptive Proton Therapy Plans". *Cancers* 14.16 (2022). DOI: 10.3390/cancers14163926.

[9] O. Weistrand and S. Svensson. "The ANACONDA algorithm for deformable image registration in radiotherapy". *Medical Physics* 42.1 (2015). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1118/1.4894702, pp. 40–53. DOI: 10.1118/1.4894702.

Short-term mortality prediction in SIII NSCLC: What do clinicians want and can models provide it?

Matthew Craddock[1,2], Azadeh Abravan[1,2], Alan McWilliam[1,2], Corinne Faivre-Finn[1,2] and Gareth Price[1,2]

[1] University of Manchester, Division of Cancer Sciences, Manchester, United Kingdom
[2] The Christie NHS Foundation Trust, Department of Clinical Oncology, Manchester, United Kingdom

Abstract Selecting the "right" treatment in the diverse SIII NSCLC patient cohort is a challenging process which is poorly supported by evidence. Currently, ~20% of SIII NSCLC patients receiving curative intent radiotherapy in Manchester die within 6 months, a timescale in which arguably they may derive no benefit from the treatment and which, we consider could be reduced by model informed clinical decision making.

We present a clinician informed modelling study in which we survey oncologists to determine their views of the utility and specification of a short term survival model in SIII NSCLC. Conventional time to event modelling, modern machine learning based methods and discrete time modelling are applied to demonstrate the inherent difficulties of deriving clinically useful predictions from low event datasets with significant class overlap.

1 Introduction

Stage III disease accounts for ~20% of all non-small cell lung cancer (NSCLC) cases and presents in a highly diverse cohort including both the young and fit, as well as the elderly, frail and co-morbid [1]. As such, a wide range of treatment options are available to SIII patients ranging from simple, single fraction palliative radiotherapy to combined curative intent chemoradiotherapy and immunotherapy.

Choosing the "right" treatment regimen from several plausible options is a challenging multi-factorial process shared between clinicians and patients. Currently, there is a dearth of evidence to support optimal treatment selection and this has led to diversity in practice as well as the potential over use of intensive radical radiotherapy treatments [1,2].

Currently, ~20% of SIII NSCLC patients receiving curative intent radiotherapy in Manchester die within 6 months, a timescale in which arguably they may derive no benefit from the treatment. While there will always be a number of patients with unexpectedly short survival, we consider 20% to be a figure which could be reduced by model informed clinical decision making.

We present a clinician informed modelling study in which we conduct a survey of oncologists to determine clinicians' views of the utility and specification of a short term survival model in SIII NSCLC. Conventional time to event modelling, modern machine learning based methods and discrete time modelling are applied to demonstrate the inherent difficulties of deriving clinically useful predictions from low event datasets with significant class overlap.

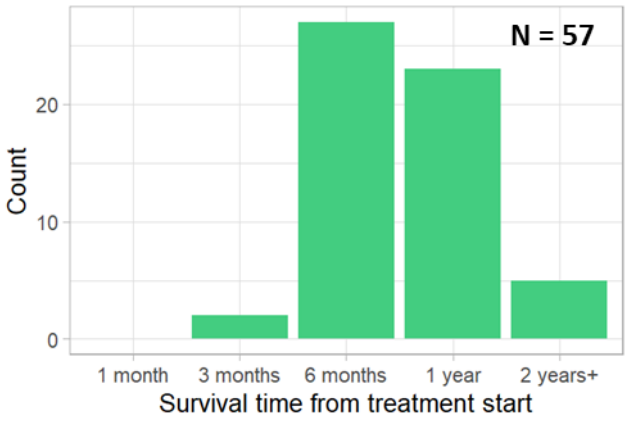


Figure 1. Responses from the survey of oncologists to: "What is the minimum meaningful survival time from the first fraction which in your opinion justifies a radical treatment?".

2 Materials and Methods

Oncologists specialising in lung cancer from multiple institutions were invited to participate in the short survey (14 questions) by email. The survey was hosted on the Qualtrics platform and 60 responses were recorded between August and October 2023 [3].

Models were developed using a dataset of 783 SIII NSCLC patients receiving a radical treatment of either 50-55 Gy in 20 fractions or 60-66 Gy in 30-33 fractions between 2012 and 2022 at The Christie Hospital, Manchester. Bloods data acquired within 30 days of the first fraction were available for a subset of 481 patients, including measurements of albumin, alkaline phosphate, calcium, glucose, lactate dehydrogenase, potassium, sodium, bilirubin, protein and urea. The publicly available dataset from Oberije et al was used for external validation of the models [4].

Variable selection was performed based on domain expert knowledge and using the analytical approach of random forest based variable importance measures with bootstrapping [5].

Three distinct modelling methods were applied: time to event, machine learning classification algorithms and discrete time modelling. For time to event modelling, the semi-parametric Cox model was used, and classification algorithms were built using the random forest method [6,7]. Discrete time modelling (DTM) is effectively a hybrid of the time to event and classification approaches in that it permits the use of classification algorithms whilst dealing with censoring [8]. It splits survival time into discrete boxes

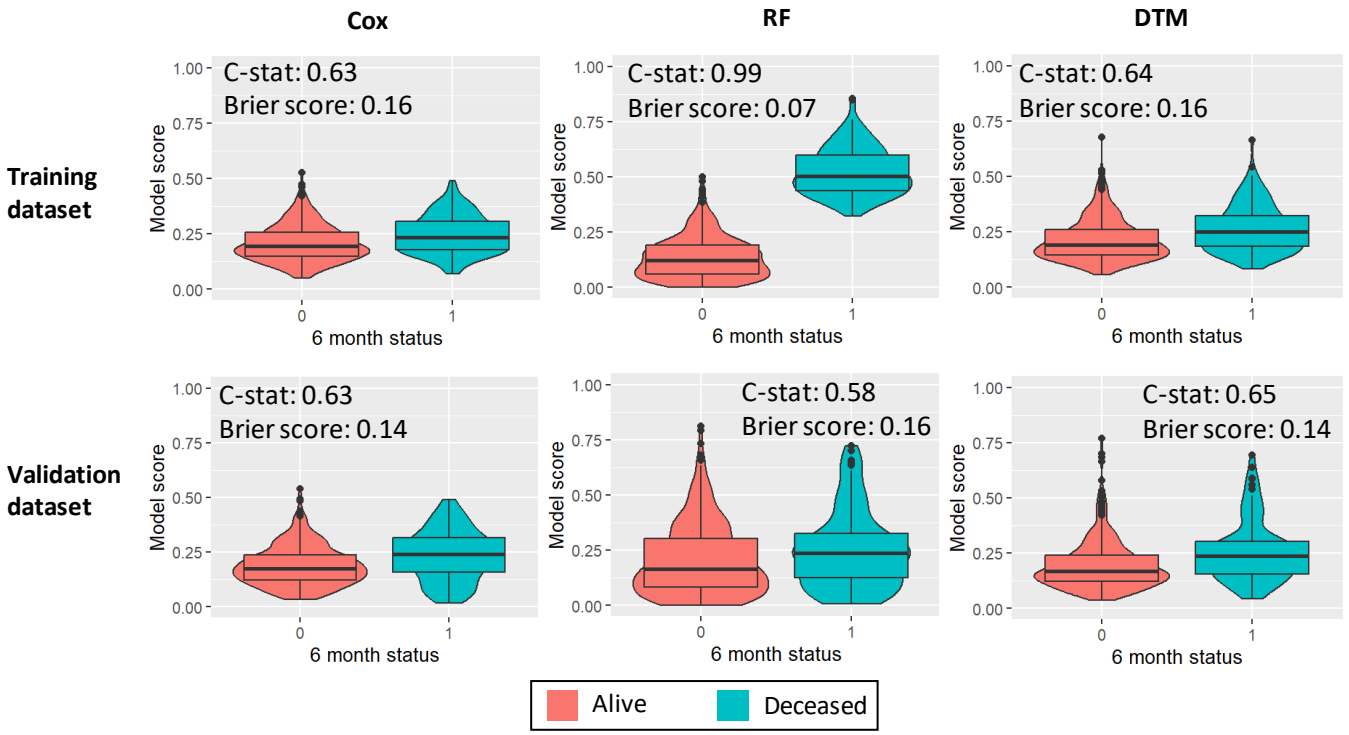


Figure 2. Prediction distributions of the Cox, random forest and neural net based discrete time model with base clinical variables (age, gender, performance status, log PTV) in the training and validation datasets.

and fits a different model in each box. The R package autoSurv (V0.1.0) was used to perform discrete time modelling using random survival forests, generalized linear models (GLMs), regularized GLMs, support vector machines, and neural network algorithms [9].

Model performance was assessed graphically by plotting the distribution of predictions by patient survival status at 6 months and quantitatively, with calculation of the discrimination C-statistic and Brier score. The Brier score is a measure of predictive accuracy which takes values between 0, representing perfect accuracy, and 1, representing inverted predictions.

Decision curve analysis (DCA) is a method of contextualising model performance with clinical decision thresholds[10]. Decision thresholds here refer to the probability values at which clinicians opt to change their treatment decision; in this work this refers to the decision to opt for a radical treatment based on the patients 6 month survival probability. As the aim of this study is to reduce the futile use of radical radiotherapy, DCA curves of interventions avoided were constructed to understand the potential impact of model predictions on current decision making.

3 Results

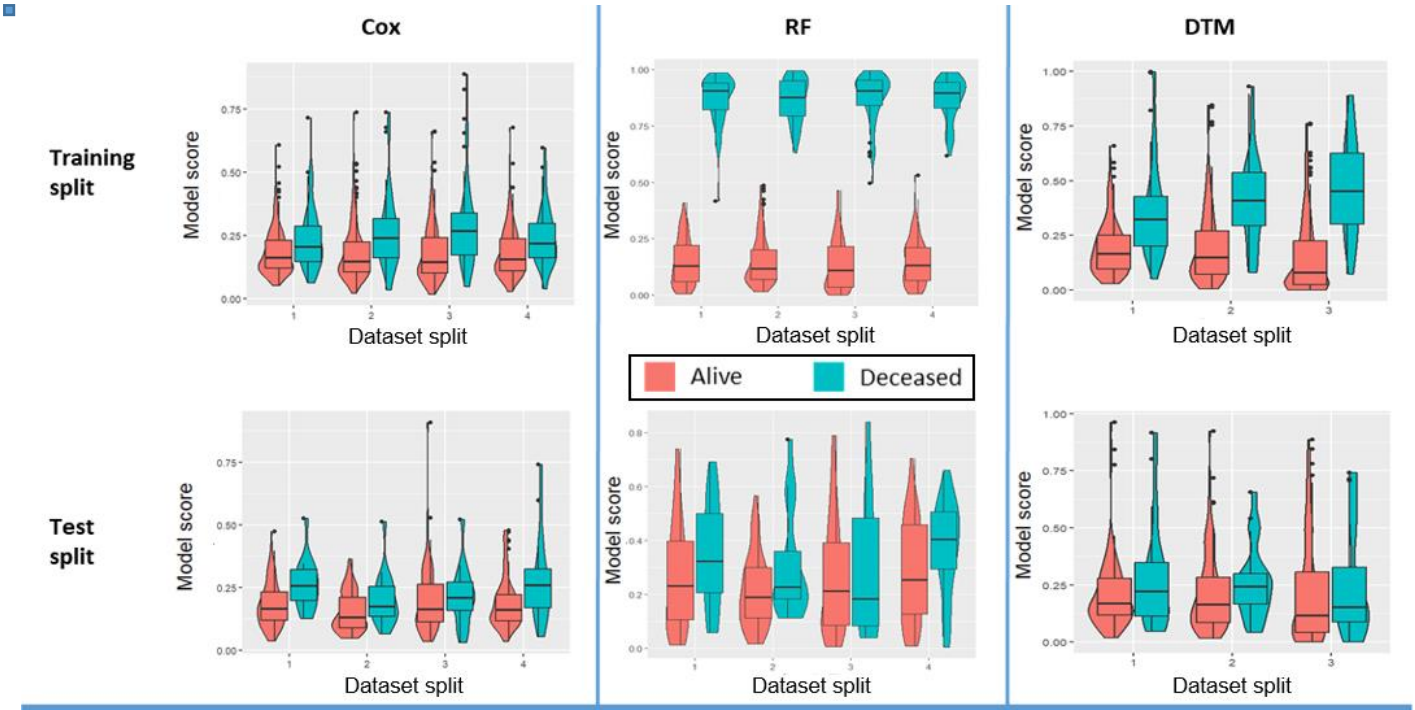
3.1 Survey

89% of respondents selected either 6 months (49%) or 1 year (40%) as the minimum meaningful survival time which they felt justified a radical treatment, with 6 months the most common response and the weighted mean over all responses equating to ~9.5 months. A short term mortality model was deemed likely to change clinical decision making by 41 out of 54 respondents and a strong preference was expressed for model output in the form of risk groupings rather than a continuous outcome or binary classification. Respondents were also asked at what threshold of probability of a patient surviving for 6 months or more would their treatment selection change from palliative to radical. The response to this question was the most mixed of all, potentially indicative of the diversity of clinical decision making in this area. The majority vote was for 50%, with an overall mean of 49%.

3.2 Modelling

Inspection of the pre-existing survival models for SIII NSCLC reveals that performance status is common to all of them, with age, gender and a tumour related volume (GTV/PTV etc) selected in 3 out of 4 models. Accordingly, these 4 variables were selected as a base set of clinical variables with which to train models to provide a performance baseline to assess the added value of any other predictors. Analytical variable reduction supported the selection of this base variable set and in the subset of patients with bloods data, lactate dehydrogenase, albumin and urea were also highly selected. As a result, models were trained with the base clinical variables alone and with the addition of the bloods metrics.

Figure 2 shows the distribution of model predictions by patient survival status at 6 months in the Christie training dataset and the MAASTRO validation dataset with base clinical variables. The random forest model is the only approach which produces meaningful separation in the training dataset, however, this performance does not translate to the validation dataset, with the model's



Performance metrics				
C-stat	Training	0.69 (0.03)	0.99 (0.00)	0.77 (0.07)
	Test	0.66 (0.08)	0.60 (0.07)	0.55 (0.08)
Brier score	Training	0.14 (0.00)	0.03 (0.00)	0.13 (0.02)
	Test	0.15 (0.01)	0.18 (0.02)	0.18 (0.03)

Figure 3. Prediction distributions of the Cox, random forest and neural net based discrete time model with base variables and blood metrics in 4-fold cross validation of the Christie dataset.

discrimination (C-statistic) falling from 0.99 to 0.58. Due to the lack of an external validation dataset for the bloods data, the generalizability of model performance with blood metrics was tested by 4-fold cross validation. As shown in Figure 3, the inclusion of blood metrics leads to higher discrimination and Brier score values in the training folds of the Christie dataset. In the test folds, the models again exhibit marked degradation in performance. The Cox model exhibits the most stable performance with discrimination (C-statistic) of 0.66 and a brier score of 0.15, but with a very meagre prediction range with an average difference in the median prediction in the dead and alive groups across all test folds of less than 10%.

3.3 Clinical Utility

Figure 4 depicts the impact of model predictions on the number of futile radical treatments avoided for a survival endpoint of 6 months, using the Cox model with base variables and blood metrics. Here, futility is defined as a radically treated patient dying within 6 months of treatment initiation. The DCA plot shows that the model only offers benefit at thresholds of between 10% and 25% which falls well below the mean decision threshold of 50% that was established by our survey of clinicians. As such the information provided by the model would offer negligible benefit to clinical decision making.

4 Discussion

In this study, we demonstrate the principle of clinician-informed modelling. By surveying the intended end user group prior to model development, we ensure selection of an end point of clinical relevance and avoid contributing to the preponderance of models which predict events of no clinical relevance [11]. Clinician informed model design also permits model evaluation which goes beyond the conventional metrics of calibration and discrimination, which are not linked to clinical utility [12]. By establishing clinical decision thresholds, we can begin to contextualize model performance. Although simplistic, methods such as decision curve analysis allow us to gain some understanding of whether the models offer benefit over current decision making practices.

The paucity of models for short term survival in SIII NSCLC may be reflective of the difficulty of achieving adequate performance at endpoints of less than 12 months. A fundamental cause of this is the lack of events – reported rates of early mortality at time points of 3-6 months range from 6% to 20%, meaning that even large datasets suffer from a relative lack of information in the patient group of most interest [13,14]. A nationwide analysis of risk factors for early mortality in lung cancer patients in France by Gossault highlighted emergency presentation and hospitalisation lasting for more than 30 days as key drivers, potentially suggesting broader datasets encompassing primary care

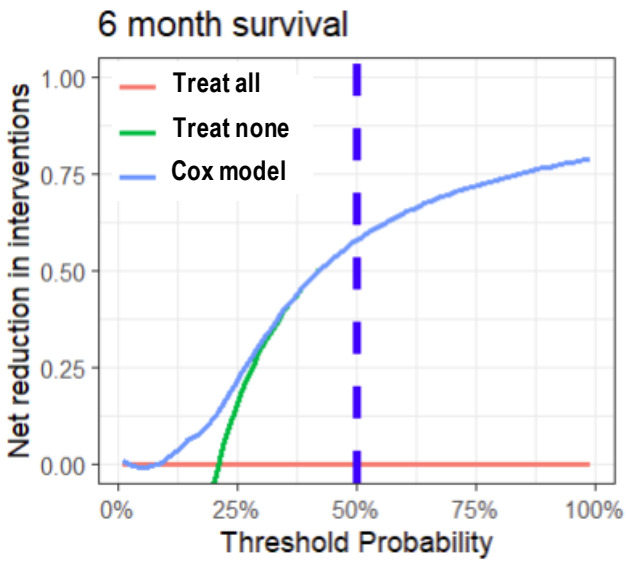


Figure 4. Decision curve analysis plot showing the net reduction in futile radical treatments as a function of decision threshold probability. The blue line indicates decisions made in concordance with the Cox model predictions and the green/red lines indicates the scenarios of none/all of the patients receiving a radical treatment.

information may be required to achieve adequate performance[14].

Our survey of oncologists indicated a strong preference for model output in the form of risk groupings. Quantitative definition of these groups is vital in correctly interpreting the meaning of a model and indeed is a requirement of the TRIPOD statement (item 11)[15]. Despite this, 47% of existing modelling studies fail to appropriately define their risk groupings and there is no consensus on the methodology used to do this [16]. This poses a risk as poor model performance can be obfuscated by misleading group names [13]. For survival models a simple approach would be to present the median survival in each group as a proportion of that in the whole cohort, such that values close to 1 represent poor group separation and values further away represent more meaningful separation.

While this study represents a rare example of clinician informed model design, it is important to recognize that decision making is a process shared between oncologists and patients. Patients may well have different requirements for model output as well as different variables driving their decision making, with treatment experience and side effects potentially being considered as important as survival [17,18]. As such a survey of patients would be a useful extension of this work and a necessity to understand model suitability and performance for all parties involved in clinical decision making.

5 Conclusion

Clinicians recognise that short term (6-12 months) survival prediction in SIII NSCLC is likely to change their decision making. Current methods are unable to generate performative models with clinically relevant endpoints in low event datasets with significant class overlap. New methods and/or novel variables are required to achieve meaningful separation of patients by their survival status at 6 months in the cohort of SIII NSCLC.

6 References

1. Royal College of Physicians. National Lung Cancer Audit Annual Report (for the Audit 2019 England, Wales and Guernsey and 2020 England Only).; 2022. www.hqip.org.uk/national-programmes

2. Brada M, Forbes H, Ashley S, Fenwick J. Improving Outcomes in NSCLC: Optimum Dose Fractionation in Radical Radiotherapy Matters. J Thorac Oncol. 2022;17(4):532-543. DOI:10.1016/j.jtho.2022.01.006

3. Qualtrics survey platform, https://www.qualtrics.com/uk/ (Accessed: 22.01.24)

4. Oberije C, De Ruysscher D, Houben R, et al. A validated prediction model for overall survival from stage III non-small cell lung cancer: Toward survival prediction for individual patients. Int J Radiat Oncol Biol Phys. 2015;92(4):935-944. DOI:10.1016/j.ijrobp.2015.02.048

5. Archer KJ, Kimes R V. Empirical characterization of random forest variable importance measures. Comput Stat Data Anal. 2008;52(4):2249-2260. DOI:10.1016/j.csda.2007.08.015

6. Kleinbaum DG, Klein M. The Cox Proportional Hazards Model and Its Characteristics.; 2012. DOI:10.1007/978-1-4419-6646-9_3

7. Wongvibulsin S, Wu KC, Zeger SL. Clinical risk prediction with random forests for survival, longitudinal, and multivariate (RF-SLAM) data analysis. BMC Med Res Methodol. 2019;20(1):1-14. DOI:10.1186/s12874-019-0863-0

8. Suresh K, Severn C, Ghosh D. Survival prediction models: an introduction to discrete-time modeling. BMC Med Res Methodol. 2022;22(1):5-6. DOI:10.1186/s12874-022-01679-6

9. autoSurv github repository. https://github.com/ksuresh17/autoSurv

10. Vickers AJ, Elkin EB. Decision curve analysis: A novel method for evaluating prediction models. Med Decis Mak. 2006;26(6):565-574. DOI:10.1177/0272989X06295361

11. Wyatt JC, Altman DG. Commentary: Prognostic models: clinically useful or quickly forgotten? BMJ. 1995;311(7019):1539-1591. DOI:10.1136/bmj.311.7019.1539

12. Craddock M, Crockett C, McWilliam A, et al. Evaluation of Prognostic and Predictive Models in the Oncology Clinic. Clin Oncol. 2022;34(2):102-113. DOI:10.1016/j.clon.2021.11.022

13. Jochems A, El-Naqa I, Kessler M, et al. A prediction model for early death in non-small cell lung cancer patients following curative-intent chemoradiotherapy. Acta Oncol (Madr). 2018;57(2):226-230. DOI:10.1080/0284186X.2017.1385842

14. Goussault H, Gendarme S, Assié JB, et al. Risk factors for early mortality of lung cancer patients in France: A nationwide analysis. Cancer Med. 2022;11(24):5025-5034. DOI:10.1002/cam4.4821

15. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD Statement. BMC Med. 2015;13(1):1-10. DOI:10.1186/s12916-014-0241-z

16. Mallett S, Royston P, Dutton S, Waters R, Altman DG. Reporting methods in studies developing prognostic models in cancer: A review. BMC Med. 2010;8. DOI:10.1186/1741-7015-8-20

17. Josfeld L, Keinki C, Pammer C, Zomorodbakhsch B, Hübner J. Cancer patients' perspective on shared decision-making and decision aids in oncology. J Cancer Res Clin Oncol. 2021;147(6):1725-1732. DOI:10.1007/s00432-021-03579-6

18. Farmer GD, Pearson M, Skylark WJ, Freeman ALJ, Spiegelhalter DJ. Redevelopment of the Predict: Breast Cancer website and recommendations for developing interfaces to support decision-making. Cancer Med. 2021;(November 2020):1-13. DOI:10.1002/cam4.4072

Deep learning autosegmentation performance on Mid-Position and time-averaged lung CT scans

Joren Brunekreef[1] and Jan-Jakob Sonke[1]

[1]Department of Radiotherapy, Netherlands Cancer Institute, Amsterdam, the Netherlands

Abstract 4D CT scans are routinely used in the treatment preparation process of radiation therapy (for targets in the upper abdomen and thorax to reduce geometric uncertainties. Online adaptive RT (OART) is an emerging technique to increase the accuracy in radiation therapy but in room 4D imaging techniques are currently scarce. To increase the efficiency of OART, deep learning auto-segmentation is an active field of research. In this work, we investigate the difference in auto-segmentation quality in motion averaged and motion compensated scans. We found that MidP scans led to increased segmentation performance for GTVs and organs-at-risk compared to mean-intensity projections. These results indicate that future workflows for online adaptive radiotherapy may benefit from integrating motion-compensation techniques when using deep learning auto-segmentation algorithms for delineating anatomical structures.

1 Introduction

The thoracic and upper-abdominal region are subject to respiratory motion which induces geometric uncertainties during radiation therapy. A widely adopted strategy to reduce such geometric uncertainties is to acquire 4D CTs for treatment planning [1]. Delineations of targets and organs at risk and treatment plan optimization are then typically performed on individual phases, mean/maximum intensity projections [2] or the time averaged mean position (mid-Position (MidP)) [3]. An emerging technology to reduce geometric uncertainties in radiation therapy is online adaptive radiotherapy (OART) [4]. In OART, the plan is adapted daily using in room imaging modalities such as CT, CBCT or MRI. Currently, OART is time consuming with segmentation of the target and organs at risk (OAR) being one of the bottlenecks. Therefore, deep-learning based auto-segmentation is being studied to increase the efficiency of OART [5–7]. In-room 4D imaging modalities are currently scarce. Consequently, 3D imaging techniques are being used as an alternative which are subject to artifacts such as blur. It is unknown whether such artifacts impact the performance of auto-segmentation algorithms. The aim of this study was therefore to investigate the difference in auto-segmentation quality in motion averaged and motion compensated scans.

2 Materials and Methods

Following IRB approval, we collected the 4D planning CT (pCT) scans and corresponding structure sets of 466 early stage lung cancer patients treated with stereotactic body radiotherapy (SBRT). From the 4D scan sequences, both the midPosition (MidP) and mean-intensity projection 3D scans were computed. The gross target volume (GTV) were manually delineated by the treating radiation oncologists and the organs at risk (OARs) by the dosimetrist on the 3D MidP scans. Voxel sizes ranged from 0.98x0.98x3.00mm^3 to 1.17x1.17x3.00mm^3.

The potential difference in utility between MidP and time-averaged scans for autosegmentation was analysed from two distinct perspectives. On the one hand, we investigated the effectiveness of the datasets resulting from both post-processing methods for training a neural network to segment GTVs (**Task 1**). On the other hand, we evaluated the performance of an existing state-of-the-art deep learning segmentation algorithm for anatomical structures by comparing the results to the OAR-annotations provided by the dosimetrist (**Task 2**).

More specifically, in **Task 1** we split the dataset into 400 patients for training/validation and 66 patients for testing. We subsequently train an nnUNet [8] to segment GTVs, based on the manual delineations. We distinguish two subtasks in this context: for **Task 1a** the model is given the full chest scan as input, and for **Task 1b** we simplify the objective by cropping a 15x15x15cm^3 box around the tumor's center of mass and providing the cropped scan as input to the model.

The best model is obtained through a five-fold crossvalidation procedure, and we subsequently evaluate it on the test split. We perform this process twice: once for the MidP scans, and once for the time-averaged 3D scans. Note that the dataset split into training/validation and test data happens only once, and on the patient level — as a result, both models are evaluated on the same set of patients in order to make an unbiased comparison.

In order to evaluate the autosegmentation performance, we compute the Dice similarity coefficient (DSC) and the symmetric 95% Hausdorff distance (s95HD) between the trained nnUNet's predictions and the ground truth delineations provided by the radiation oncologists. Subsequently, we compare these performance metrics between the models trained and evaluated on MidP scans on the one hand, and time-averaged scans on the other hand.

For **Task 2**, we use TotalSegmentator [9] to segment five organs-at-risk on all 466 patients in our dataset, for both the MidP and time-averaged 3D scans. The chosen OARs are the heart, spinal cord, esophagus, and left and right lungs. We

then compare these autosegmentations to the dosimetrists' ground truth annotations by again computing the Dice similarity coefficient and the 95% Hausdorff distance. However, since the dosimetrists' annotations are sometimes incomplete in regions far removed from the GTV area in the craniocaudal direction, we restrict computation of these metrics to only the axial slices where the relevant OAR annotations are not everywhere zero. Also, since the annotation conventions of the dosimetrist and TotalSegmentator are not identical,[1] we report the *asymmetric* 95% Hausdorff distance (a95HD) from the ground truth dosimetrist delineation to TotalSegmentator's predicted segmentations (as opposed to the s95HD).

3 Results

3.1 Task 1

The results of the GTV segmentation task are shown in Figure 1 and Table 1, where "GTV (full)" refers to Task 1a and "GTV (crop)" to Task 1b. When the objective is to segment the tumor in the full scan (Task 1a), many tumors are missed completely in both the midposition- and time-averaged scans. For the simpler Task 1b of segmenting the GTV in a smaller cube-shaped region centered on the tumor, all GTVs are at least partially found for both scan post-processing methods. The models trained (and evaluated) on MidP scans attain the best performance in both scenarios, reaching a mean Dice similarity coefficient of 0.85 and a (symmetric) 95% Hausdorff distance of 4.2mm in the second subtask.
We present an example slice of a test patient's scan for both the midposition- and time-averaging methods in Figure 2. This Figure demonstrates the blurring effect of the breathing cycle motion, and the mitigation of this effect due to the midposition-averaging. The ground truth GTV delineation is shown in green in both of the images. In the leftmost image, we show the predicted segmentation of the model trained on MidP scans, whereas the rightmost image shows in red the prediction of the model trained on time-averaged scans.

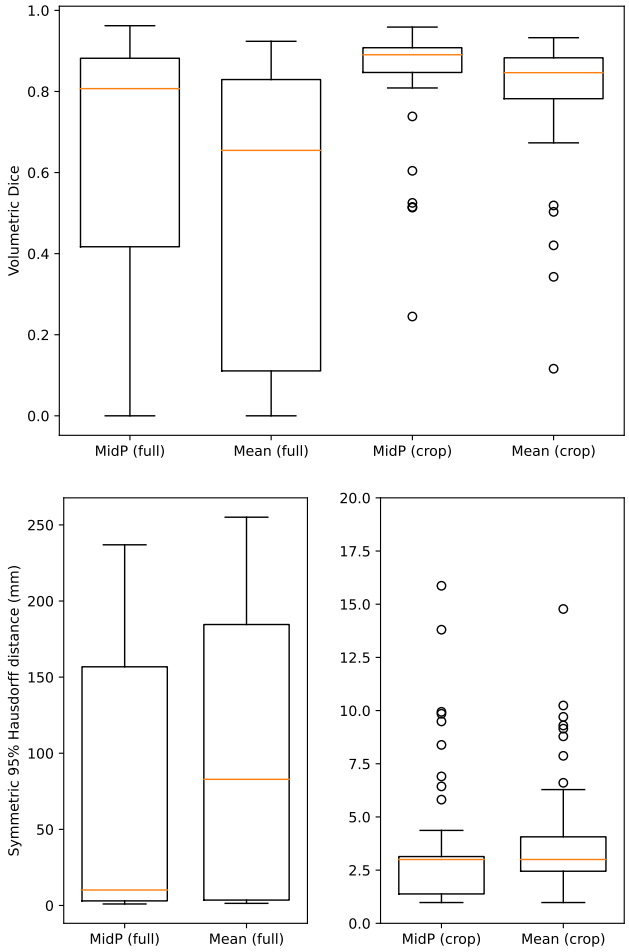


Figure 1: Performance metrics for GTV segmentations on MidP and time-averaged 3D scans, as evaluated on the test set consisting of 66 patients (Tasks 1a, 1b. Three outliers (one for MidP-averaged, two for time-averaged) are not shown in the 95% Hausdorff distance plot so that the bulk of the results can be inspected in more detail.

3.2 Task 2

Table 1 reports the performance metrics of TotalSegmentator's predicted organs-at-risk segmentations as compared to the ground truth annotations. Dice similarity coefficients are consistently better on the MidP scans than on the time-averaged scans, but the difference is minimal. The (asymmetric) 95% Hausdorff distance is marginally lower on the MidP scans for both lungs and the spinal cord, pointing towards slightly better performance for these OARs. For the lung segmentations, this a95HD is close to the voxel spacing for both scan-averaging methods. The a95HD for heart and

[1] For instance, the pulmonary artery is considered part of the heart by TotalSegmentator, whereas the dosimetrists did not include this substructure in the annotations (following the standard [10]).

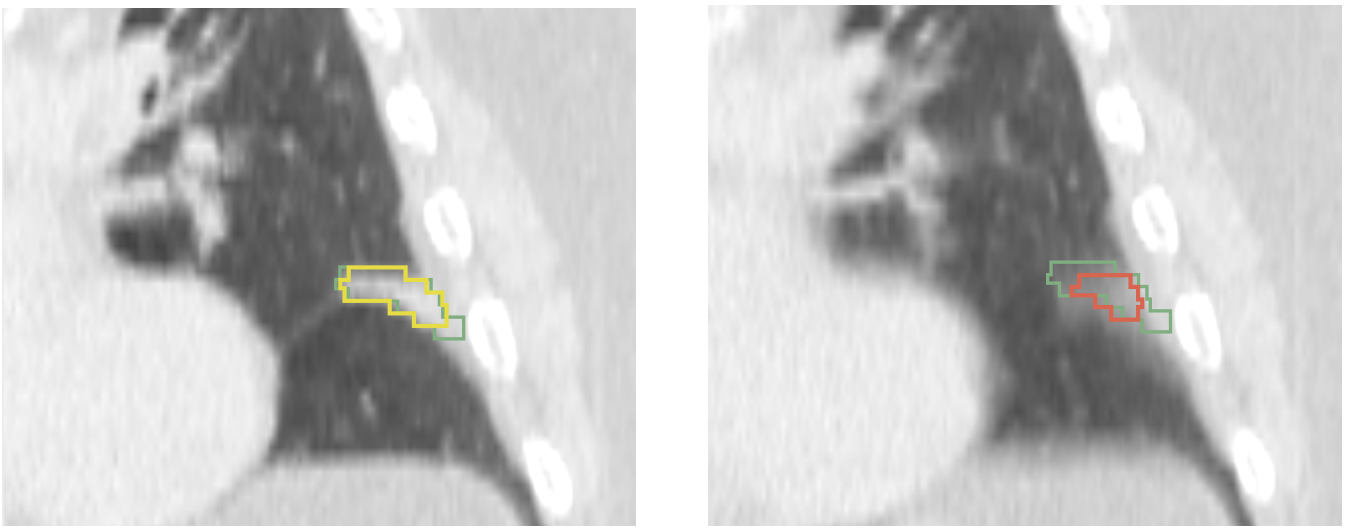


Figure 2: Example of paired MidP (left) and time-averaged (right) scan slices in the sagittal plane, shown with ground truth GTV annotation (green) and the predicted segmentation obtained from the corresponding trained nnUNet (yellow for the model trained on MidP scans, red for the model trained on time-averaged scans).

esophagus is comparable between both averaging methods, as the means overlap within one standard error.

4 Discussion and Conclusion

We have evaluated deep learning autosegmentation on MidP and time averaged CT scans of lung cancer patients and found improved performance for MidP scans for both the GTV and OARs. These results indicate that in-room motion compensated imaging techniques may be beneficial for online adaptive radiotherapy for anatomical regions subject to respiratory motion. The manual delineations in this work were performed on the MidP scans. This potentially induces a bias compared to the time average scans. However, as the MidP scans are designed to reconstruct the anatomy in the time averaged mean position, the segmentations should also represent the underlying anatomy for the time averaged scans. Manual delineations on time averaged scans will be challenging due to the respiratory motion induced blur and induced increased inter-observer variation. In the work we evaluated the auto-segmentation performance for both targets and organs at risk. It is often assumed that the segmentation of targets is more challenging than segmentation of OARs. In this work the performance of autosegmentation targets did not seem to be inferior to OARs although there were considerable differences between different OARs. This may be partly explained by the low interobserver target delineation variation in lung SBRT [11]. As our hospital doesn't use internal target volumes to obtain a GTV-to-PTV expansions, these structures were not available for evaluation. Maximum intensity projection scans may be more suitable for this task. In this work we used the DICE and 95HD for performance evaluation as these are widely used in this context. These metrics, however, do not strongly correlate with the amount of time involved in correcting auto-segmentations or the margin that is needed to correct for auto-segmentation inaccuracies. Added path length and root-mean-square error may be more suitable for these tasks. In this work, we used scans derived from 4D CT scans acquired for treatment planning to evaluate the performance of autosegmentation for which we had hundreds of annotated datasets available. Further research should be performed to confirm the impact of in-room 4D or motion compensated imaging techniques for online adaptive. In this work we approximated 3D imaging by time averaging of the 4D CT. While this approach induces respiratory induced blur, actual 3D imaging techniques may have additional artifacts such as streaks or shifts.

Acknowledgments

This work was supported by an institutional grant from the Dutch Cancer Society and the Dutch Ministry of Health, Welfare and Sport. We would like to thank Vineet Mohan, Lennert Ploeger, and Bas Schipaanboord for their assistance in preprocessing the data.

References

[1] E. C. Ford, G. Mageras, E. Yorke, et al. "Respiration-correlated spiral CT: a method of measuring respiratory-induced anatomic motion for radiation treatment planning". *Medical physics* 30.1 (2003), pp. 88–97.

[2] R. W. Underberg, F. J. Lagerwaard, B. J. Slotman, et al. "Use of maximum intensity projections (MIP) for target volume generation in 4DCT scans for lung cancer". *International Journal of Radiation Oncology* Biology* Physics* 63.1 (2005), pp. 253–260.

[3] J. W. Wolthaus, J.-J. Sonke, M. Van Herk, et al. "Reconstruction of a time-averaged midposition CT scan for radiotherapy planning of lung cancer patients using deformable registration a". *Medical physics* 35.9 (2008), pp. 3998–4011.

		Dice (↑)		95HD (mm) (↓)	
		MidP	Mean	MidP	Mean
Task 1	GTV (full)	**0.635** (0.040)	0.513 (0.042)	**89.97** (3.41)	98.08 (3.39)
	GTV (crop)	**0.854** (0.015)	0.802 (0.017)	**4.21** (0.98)	4.91 (0.98)
Task 2	Left lung	**0.957** (0.001)	0.953 (0.001)	**5.72** (0.41)	7.48 (0.41)
	Right lung	**0.956** (0.002)	0.950 (0.002)	**6.85** (0.44)	9.35 (0.42)
	Heart	**0.807** (0.003)	0.794 (0.003)	12.03 (0.11)	**11.95** (0.12)
	Esophagus	**0.819** (0.003)	0.803 (0.003)	**6.19** (0.20)	6.41 (0.22)
	Spinal cord	**0.846** (0.002)	0.843 (0.002)	**2.74** (0.04)	2.78 (0.04)

Table 1: Mean performance metrics (with standard errors) for GTV segmentation (Tasks 1a, 1b) and organs-at-risk segmentation (Task 2). The reported 95HD is the *symmetric* 95% Hausdorff distance for GTVs and its *asymmetric* counterpart for the organs-at-risk. Results for Task 1 were computed on the 66 patients in the held-back test split, whereas results for Task 2 were computed on the set of all 466 patients.

[4] S. Lim-Reinders, B. M. Keller, S. Al-Ward, et al. "Online adaptive radiation therapy". *International Journal of Radiation Oncology* Biology* Physics* 99.4 (2017), pp. 994–1003.

[5] P. Sibolt, L. M. Andersson, L. Calmels, et al. "Clinical implementation of artificial intelligence-driven cone-beam computed tomography-guided online adaptive radiotherapy in the pelvic region". *Physics and imaging in radiation oncology* 17 (2021), pp. 1–7.

[6] R. R. Savjani, M. Lauria, S. Bose, et al. "Automated tumor segmentation in radiotherapy". *Seminars in radiation oncology*. Vol. 32. 4. Elsevier. 2022, pp. 319–329.

[7] C. M. Benitez, M. D. Chuong, L. A. Künzel, et al. "MRI-Guided Adaptive Radiation Therapy". *Seminars in Radiation Oncology*. Vol. 34. 1. Elsevier. 2024, pp. 84–91.

[8] F. Isensee, P. F. Jaeger, S. A. Kohl, et al. "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation". *Nature methods* 18.2 (2021), pp. 203–211.

[9] J. Wasserthal, H.-C. Breit, M. T. Meyer, et al. "Totalsegmentator: Robust segmentation of 104 anatomic structures in ct images". *Radiology: Artificial Intelligence* 5.5 (2023).

[10] M. Feng, J. M. Moran, T. Koelling, et al. "Development and validation of a heart atlas to study cardiac exposure to radiation following treatment for breast cancer". *International Journal of Radiation Oncology* Biology* Physics* 79.1 (2011), pp. 10–18.

[11] H. Peulen, J. Belderbos, M. Guckenberger, et al. "Target delineation variability and corresponding margins of peripheral early stage NSCLC treated with stereotactic body radiotherapy". *Radiotherapy and Oncology* 114.3 (2015), pp. 361–366.

Surrogate-Optimized Respiration Motion Model to Estimate Motion for Each Projection of 3D CBCT Scan

Yuliang Huang[1,2], Kris Thielemans[1,3], and Jamie R McClelland[1,2]

[1]Centre for Medical Image Computing, Department of Medical Physics and Biomedical Engineering,University College London, London, UK
[2]Wellcome and EPSRC centre for Interventional and Surgical and Sciences, Department of Medical Physics and Biomedical Engineering,University College London, London, UK
[3]Institute of Nuclear Medicine, University College London, London, UK

Abstract Cone-beam CT (CBCT) images are commonly acquired during radiotherapy to help setup the patient and to monitor changes of their anatomy. However, standard 3D CBCT images do not estimate breathing motion and are susceptible to motion artifacts when it occurs. State-of-the-art surrogate driven motion models can estimate the motion for each CBCT projection using correspondence between the internal motion and one or more respiration surrogate signals. However, it can be challenging to measure or extract surrogate signals that are strongly related to the internal motion. Therefore this study proposes a novel surrogate-optimized respiratory motion model and investigates the feasibility of applying it on CBCT projection data. This method optimises the surrogate signal(s) as well as the correspondence model, eliminating the need for the surrogate signals to have strong relationship with the internal motion. Our method is evaluated on both simulated and real patient data and compared with a surrogate-driven motion model, showing promising improvements across all metrics for the simulated data, and improved image quality for the real data.

1 Introduction

3D Cone-beam CT (CBCT) is commonly used in radiotherapy (RT) to help setup the patients and to monitor their anatomical changes during the treatment course. However, the acquisition time of 1-2 minutes for 3D CBCT means that the standard reconstructions cannot estimate patients' breathing motion, and the blurring artifact caused by the motion [1] can make it difficult to accurately assess anatomical changes.

A common approach to address the above problem is 4D CBCT, which sorts CBCT projections into multiple phases/bins according to respiratory signals. Projections within each phase are used to reconstruct a single CBCT image with less motion artifacts than the 3D CBCT. However, this method assumes reproducible periodic respiratory motion, so cannot account for inter-cycle (i.e. breath to breath) variations in the motion [2]. Moreover, 4D CBCT acquisitions require more projections, leading to a longer acquisition times (4-6 minutes), and the uneven distribution of projection angles within each respiration phase induces streak artifacts.

Recently we proposed a surrogate-driven motion model [3] to estimate the motion for each projection of a 3D CBCT scan and reconstruct a motion-free image. It adapted the general motion modelling framework in [4], which can model inter-cycle variations, so that it can be fitted to CBCT projection data, and hence does not require phase sorting or extra acquisition time. This method models the relationship between 3D motion fields and low-dimensional respiration surrogates signals which can be measured by external devices [5] or extracted directly from projection images [6]. Since the surrogate signals are synchronised with the CBCT acquisition, the motion field for each projection can be inferred from the surrogate signals taken at the same time. However, such methods assume there is a strong relationship between the surrogate signals and the internal motion, which is not always true.

This study aims to use a data-driven approach to optimize the surrogate signals at the same time as fitting the motion model directly to the CBCT projection data, which is denoted as a surrogate-optimized motion model. Both simulated and real clinical 3D CBCT scans are used to demonstrate the advantage of the surrogate-optimized motion model over the surrogate-driven motion model.

2 Methodology

2.1 General Motion Modelling Framework

The input data consists of n measured CBCT projections $(\mathbf{P_t}, t = 1, ..., n)$. The aim is to estimate transformations for each projection that can warp the reference state image $\mathbf{I_0}$ to dynamic image $\mathbf{I_t}$ at any time t, as shown below:

$$\mathbf{I_t} = T(\mathbf{I_0}, \mathbf{M_t}) \tag{1}$$

where T is a function that uses the spatial transformation parameterised by $\mathbf{M_t}$ to resample $\mathbf{I_0}$. This study adopts the free-form-deformation transformation model, in which case $\mathbf{M_t}$ refers to control point grid displacements at time t. Using the estimated transformations, the reference state image $\mathbf{I_0}$ can be obtained by motion compensated reconstruction [7] to remove the motion artifacts.

The transformation parameters over time can be arranged into a $N_m \times n$ matrix $\mathbf{M} = [\mathbf{M_1}, ..., \mathbf{M_n}]$, where N_m is number of parameters that define the transformation $\mathbf{M_t}$ at a single timepoint t. Assuming the low-rankness of the respiration motion [8], $\mathbf{M}$ can be approximated to

$$\mathbf{M} = \mathbf{C} \cdot \mathbf{S} = \sum_{i=1}^{r} \mathbf{C}^{(i)} \cdot \mathbf{S}^{(i)} \tag{2}$$

where r is the rank number, $\mathbf{C}$ and $\mathbf{S}$ are $N_m \times r$ and $r \times n$ matrices denoting r spatial bases and r temporal bases respectively. For surrogate driven models, $\mathbf{C}$ is the motion model which is optimised during the fitting, but $\mathbf{S}$ are the measured surrogate signals which are fixed during fitting. In this work we **extend the surrogate-driven method by also optimising the surrogate signals**.

To determine the model parameters, we minimize the following loss function

$$f = -\sum_{t=1}^{n} L(\mathbf{P}'_\mathbf{t}, \mathbf{P}_\mathbf{t}) \tag{3}$$

where L represents local normalized cross correlation [9], $\mathbf{P}'_\mathbf{t}$ is the estimated projection at time t and is given by:

$$\mathbf{P}'_\mathbf{t} = \mathbf{A}_\mathbf{t} \cdot \mathbf{I}_\mathbf{t} = \mathbf{A}_\mathbf{t} \cdot T(\mathbf{I_0}, \mathbf{M_t}) \tag{4}$$

in which $\mathbf{A_t}$ is the CBCT projection acquisition matrix at time t.

2.2 Optimization Algorithm

How many temporal basis functions are needed is an open question, although it has been shown that two principal components are sufficient to approximate intra-fraction lung motion [10]. In this study, one spatial and temporal basis were optimized first and more degrees of freedom were included and optimized sequentially to account for more complicated motion. More specifically, we started with optimizing just one spatial and temporal basis until objective function value did not improve, after which another pair of spatial and temporal basis was incorporated and optimized with the previous bases frozen. The same process could continue until no improvement of image quality was observed by manual check. The first temporal basis was initialised with a surrogate signal extracted from the CBCT projections using the Intensity Analysis method [6] and the other temporal bases were initialised with the derivative of the previous optimized temporal basis.

Stochastic gradient descent is used to optimize $\mathbf{C}^{(i)} (i = 1, ..., r)$. More specifically, ten subsets of evenly spaced projections are created and one of them is randomly selected with uniform probability to update the gradient at each iteration:

$$\begin{aligned} \frac{\partial f_s}{\partial \mathbf{C}^{(i)}} &= \sum_{t \in \Omega_s} \frac{\partial \mathbf{M_t}}{\partial \mathbf{C}^{(i)}} \cdot \frac{\partial \mathbf{I_t}}{\partial \mathbf{M_t}} \cdot \frac{\partial \mathbf{P}'_\mathbf{t}}{\partial \mathbf{I_t}} \cdot \frac{f}{\partial \mathbf{P}'_\mathbf{t}} \\ &= -\sum_{t \in \Omega_s} \mathbf{S}_t^{(i)} \cdot \frac{\partial T(\mathbf{I_0}, \mathbf{M_t})}{\partial \mathbf{M_t}} \cdot \mathbf{A_t}^* \cdot \frac{L(\mathbf{P}'_\mathbf{t}, \mathbf{P_t})}{\partial \mathbf{P}'_\mathbf{t}} \end{aligned} \tag{5}$$

where f_s is the loss function as in Eq. 3 but now summed over the selected projection subset, $\mathbf{A}^*$ is the back-projection matrix and Ω_s is the set of indices in the selected subset.

$\mathbf{S}^{(i)}$ is optimized with gradient descent for each time point separately. For time t, the gradient is similar to Eq. 5:

$$\frac{\partial f}{\partial \mathbf{S}_t^{(i)}} = -\sum_{v=1}^{N_m} \mathbf{C}_v^{(i)} \cdot \frac{\partial T(\mathbf{I_0}, \mathbf{M_t})}{\partial \mathbf{M}_{\mathbf{t},v}} \cdot \mathbf{A_t}^* \cdot \frac{L(\mathbf{P}'_\mathbf{t}, \mathbf{P_t})}{\partial \mathbf{P}'_\mathbf{t}} \tag{6}$$

Note that the outermost summation is over space rather than over time.

For the ith basis function, we performed 5 iterations of optimising $\mathbf{C}^{(i)}$ and then 5 iterations of optimising $\mathbf{S}^{(i)}$. Then $\mathbf{S}^{(i)}$ was normalized to have mean of 0 and standard deviation of 1, and the corresponding spatial basis was scaled accordingly. Then motion compensated FDK reconstruction [7] was used to update $\mathbf{I_0}$, which should match the mid respiration position since the surrogate signals have been normalized. All the steps were repeated up to a maximum of 5 times or until no more improvement.

This study adopted a multi-resolution approach with three spatial resolution levels. The optimisation was only performed at the first two levels as it was found that optimising at the final (original) resolution level only provided marginal improvement but greatly increased the computation time. Our method has been implemented by extending an open source software SuPReMo, and will itself be released in the future.

2.3 Test Data

Phantom simulation and clinical 3D CBCT scans from eight patients were used to evaluate and compare the surrogate-optimized and surrogate-driven motion model.

The 4DXCAT software [11] was used to generate a reference state image and the motion fields for each projection of a 3D CBCT scan, according to breathing signals measured from cine sagittal MR images of real patients. A python library for motion field post-processing [12] was utilized to make the simulated motion fields more physically plausible. The OpenRTK library [13] was then used to generate the CBCT projection images from the dynamic phantom images according to projection acquisition geometry of a 1-minute CBCT scan on Elekta Synergy (Elekta AB, Stockholm, Sweden). The number, dimension size and resolution of the projection images were 310, 512×512 and 0.8×0.8 mm. For the phantom simulation, the repspiration was driven by two input breathing signals. Therefore, two surrogate signals / temporal basis were used. For the surrogate-driven motion model, we extracted the first surrogate signal from projections using Intensity Analysis and used its derivative as the second surrogate signal.

The Clinical 3D CBCT scans were from the ROSS-LC clinical trial (REC ref. 14/NW/0037). The data were acquired using an Elekta XVI system under standard 3D CBCT settings, i.e. ∼600 projections during a 2-minute scan over a full rotation. The SID and SDD were 1000 mm and 1536 mm respectively. Resolution and dimensional size of the projection images were 0.8×0.8 mm and 504×504 respectively. Eight patients with medium or large field-of-view scans were used.

For the phantom simulation data, the ground-truth motion fields for each projection were known, which could be used to generate tumor masks for each projection and a gold standard

motion compensated reconstruction. Four evaluation metrics were calculated: the dice score between estimated and true tumor masks (DSC), the euclidean distance between estimated and true tumor centroid positions (E_{center}), the magnitude of motion field differences (E_D), and the Peak-Signal-To-Noise-Ratio between the motion compensated reconstructions using estimated and true motion fields (PSNR). All the metrics above refer to the average value over all projections.

3 Results

Table 1 displays the evaluation metrics for the phantom simulation experiment. By applying a motion model all the metrics were substantially improved. Compared with the surrogate-driven method, the proposed surrogate-optimized motion model further improved all the metrics, although the improvements were relatively small except for DSC and E_{center}.

Scenarios	DSC	E_{center}	E_D	PSNR
no motion model	0.627	4.188	1.701	24.806
surrogate driven	0.835	1.728	1.300	25.88
surrogate optimized	0.883	1.181	1.260	26.00

Table 1: Evaluation metrics on phantom experiments under three scenarios: no motion model, surrogate-driven motion model, or surrogate-optimized motion model. Definition of metrics are in section 2.3.

Figure 1 shows the surrogate signal extracted by Intensity Analysis and the optimized temporal basis for patient 1 from the ROSS-LC clinical trial, respectively. For this patient, manual check found that the reconstructions out of one and two bases looked almost the same, so only one temporal basis was used. It can be seen that the surrogate-optimized approach produces an optimised temporal basis that is different to the input surrogate signal extracted by Intensity Analysis, but whether the optimised temporal basis can better estimate the patients' internal motion is impossible to determine without knowing the ground-truth motion.

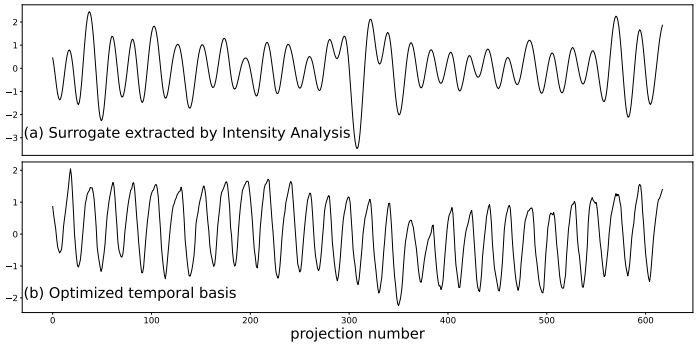


Figure 1: Input surrogate signal extracted by Intensity Analysis(a) and optimized temporal basis out of surrogate-optimized motion model(b) for one 3D CBCT scan from ROSS-LC trial.

However, the image quality from the different methods could be compared visually. Figure 2 displays the standard (non-motion compensated) FDK reconstruction, and the motion compensated FDK reconstruction using the surrogate-driven and surrogate-optimized motion models, at slices intersecting tumor for three patients from ROSS-LC trial. It can be seen that the surrogate-driven motion model already greatly reduced the motion artifacts in the scans, but the surrogate-optimized motion model further improved the sharpness of the diaphragm and visibility of details of the tumor and lung tissues.

For all the other patients from the ROSS-LC trial, the surrogate-optimized motion model produced motion compensated FDK reconstructions that were better or as good as those from the surrogate-driven motion model. The three patients shown in Figure 2 were selected to show the cases with the largest improvements of the surrogate-optimized method.

4 Discussion and Conclusion

Surrogate-driven motion models rely on being able to obtain surrogate signals that have a strong relationship with the patients' internal motion. However, in clinical practice it is often not easy or even possible to acquire suitable signals. This study has used phantom simulation data and clinical 3D CBCT scans to show that motion estimation accuracy and image quality of motion compensated reconstructions can be improved using the surrogate-optimized motion model compared with surrogate-driven motion model. It should be noted that we have also tested other surrogate extraction method such as the well known Amsterdam Shroud method [14], but the Intensity Analysis method was found to be the most robust one.

The output of our models include the estimated motion for each projection during the CBCT scan, as well as the motion compensated CBCT reconstruction image with reduced motion artifacts. The improved CBCT images could facilitate better identification of anatomical changes that occur throughout the course of radiotherapy. The estimated motion can be used to perform dose calculations and to define targets that account for the motion.

Our code was run on an Intel Core i7-10700K CPU of 3.8GHz and 8 cores. The surrogate-driven approach took on average half an hour to fit the model. However, the time to fit the surrogate-optimised models was considerably longer (8 hours), which is expected as optimising the temporal basis cannot take advantage of the stochastic subset approach. Future work will develop more advanced parallel algorithms and implementations that can be run on a GPU to reduce the computation time.

In conclusion, this study has presented a method for applying a surrogate-optimized motion model to CBCT projection data. This method has been evaluated using both simulated

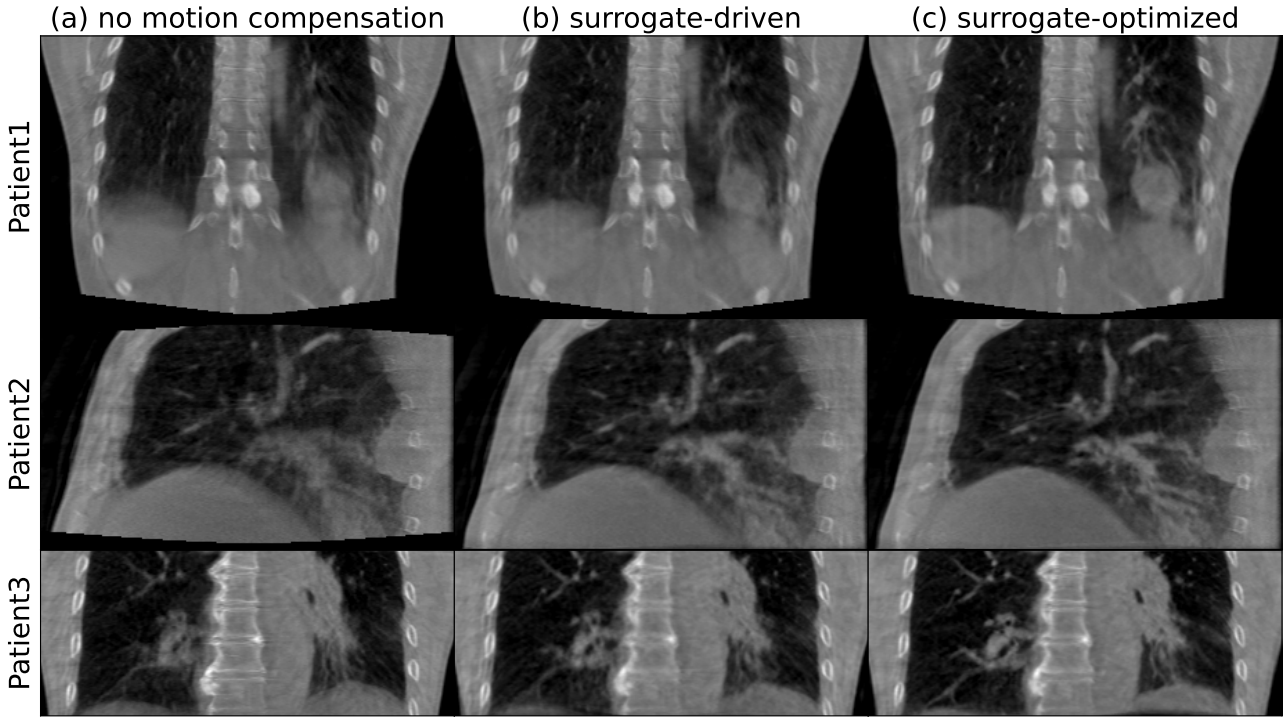


Figure 2: Comparison between (a) standard FDK reconstruction, and motion compensated FDK reconstructions using (b) surrogate-driven and (c) surrogate-optimized motion models at slices intersecting tumor for three patients in the ROSS-LC trial.

and real patient data. In the future such an approach could be of great benefit for adaptive radiotherapy treatments to lung cancer patients.

Acknowledgments

This work is supported by Elekta Ltd. Crawley and the EPSRC-funded UCL Centre for Doctoral Training in Intelligent, Integrated Imaging in Healthcare (i4health) (EP/S021930/1).

References

[1] J. H. Lewis, R. Li, X. Jia, et al. "Mitigation of motion artifacts in CBCT of lung tumors based on tracked tumor motion during CBCT acquisition". *Physics in Medicine & Biology* 56.17 (2011), p. 5485.

[2] J. Dhont, J. Vandemeulebroucke, M. Burghelea, et al. "The long- and short-term variability of breathing induced tumor motion in lung and liver over the course of a radiotherapy treatment". *Radiotherapy and Oncology* 126.2 (2018), pp. 339–346.

[3] Y. Huang, K. Thielemans, G. J. Price, et al. "Surrogate-driven respiratory motion model for projection-resolved motion estimation and motion compensated Cone-Beam CT reconstruction from unsorted projection data". *Physics in Medicine and Biology* (2023).

[4] J. R. McClelland, M. Modat, S. Arridge, et al. "A generalized framework unifying image registration and respiratory motion models and incorporating image reconstruction, for partial image data or full images". *Physics in Medicine & Biology* 62.11 (2017), p. 4273.

[5] J. R. McClelland, D. J. Hawkes, T. Schaeffter, et al. "Respiratory motion models: a review". *Medical image analysis* 17.1 (2013), pp. 19–42.

[6] A. Kavanagh, P. M. Evans, V. N. Hansen, et al. "Obtaining breathing patterns from any sequential thoracic x-ray image set". *Physics in Medicine & Biology* 54.16 (2009), p. 4879.

[7] S. Rit, J. W. Wolthaus, M. van Herk, et al. "On-the-fly motion-compensated cone-beam CT using an a priori model of the respiratory motion". *Medical physics* 36.6Part1 (2009), pp. 2283–2296.

[8] N. R. Huttinga, T. Bruijnen, C. A. van den Berg, et al. "Nonrigid 3D motion estimation at high temporal resolution from prospectively undersampled k-space data using low-rank MR-MOTUS". *Magnetic resonance in medicine* 85.4 (2021), pp. 2309–2326.

[9] P. Cachier and X. Pennec. "3D non-rigid registration by gradient descent on a Gaussian-windowed similarity measure using convolutions". *Proceedings IEEE Workshop on Mathematical Methods in Biomedical Image Analysis. MMBIA-2000 (Cat. No. PR00737)*. IEEE. 2000, pp. 182–189.

[10] R. Li, J. H. Lewis, X. Jia, et al. "On a PCA-based lung motion model". *Physics in Medicine & Biology* 56.18 (2011), p. 6009.

[11] W. P. Segars, G. Sturgeon, S. Mendonca, et al. "4D XCAT phantom for multimodality imaging research". *Medical physics* 37.9 (2010), pp. 4902–4915.

[12] B. Eiben, J. Bertholet, M. J. Menten, et al. "Consistent and invertible deformation vector fields for a breathing anthropomorphic phantom: a post-processing framework for the XCAT phantom". *Physics in Medicine & Biology* 65.16 (2020), p. 165005.

[13] S. Rit, M. V. Oliva, S. Brousmiche, et al. "The Reconstruction Toolkit (RTK), an open-source cone-beam CT reconstruction toolkit based on the Insight Toolkit (ITK)". *Journal of Physics: Conference Series*. Vol. 489. 1. IOP Publishing. 2014, p. 012079.

[14] L. Zijp, J.-J. Sonke, and M. van Herk. "Extraction of the respiratory signal from sequential thorax cone-beam X-ray images". *International conference on the use of computers in radiation therapy*. 2004, pp. 507–509.

Monte Carlo robustness analysis of breath-hold and free-breathing photon and proton treatments of small lung tumors using Poisson displacement fields

Nils Olovsson[1,2], Kenneth Wikström[1,3], Anna Flejmer[1,2,4], Anders Ahnesjö[1], and Alexandru Dasu[1,2]

[1]Department of Immunology, Genetics and Pathology, Uppsala University, Uppsala, Sweden
[2]The Skandion Clinic, Uppsala, Sweden
[3]Department of Medical Physics, Uppsala University Hospital, Uppsala, Sweden
[4]Department of Oncology, Uppsala University Hospital, Uppsala, Sweden

Abstract The adoption of proton therapy for the treatment of small lung tumors has been cautious due to the sensitivity of proton range to geometric and radiological uncertainties. There is however a potential of sparing dose to organs at risk if such treatments could be performed confidently. In this study a comprehensive Monte Carlo based method of evaluating the impact of these uncertainties for treatments in breath-hold and free-breathing is proposed together with a method for deforming images based on Laplace and Poisson editing. Tabulated values was used for the uncertainty parameters and free-breathing motion was simulated using tumor positions from multiple cine-CT.
A breath-hold proton plan was compared to breath-hold and free-breathing photon plans. With robust treatment planning adequate target coverage was achieved for the proton plan with the benefit of lower mean heart dose but at the cost of higher dose to the esophagus.

1 Introduction

Hypofractionated, stereotactic body radiotherapy (SBRT) for small peripheral lung tumors has improved the treatments for these patients. However, the uncertainties caused by patient setup, breathing motion and stopping power ratio (SPR) conversion from computed tomography (CT) image values have caused a cautious adoption of analogous proton therapy strategies [1].

At the same time, since the survival of these patients with early-stage disease have seen such improvements they are the ones that could benefit the most from the dose sparing of proton therapy. Robust optimization that considers the uncertainties can ensure dose coverage of the target and the use of Monte Carlo (MC) to thoroughly evaluate treatment plans has been demonstrated [2]. Given current capabilities of most proton therapy centers these patients could be treated either in breath-hold (BH) or in free-breathing (FB). The choice of treatment modality would depend on the extent of the breathing motion, breath-hold capacity, location of the tumor in the lung and estimated benefits of choosing proton compared to photon therapy.

We propose a unified framework based on Laplace and Poisson editing for creating deformed evaluation images for both BH and FB treatments were the FB motion was simulated using tumor positions from cine-CT. Further, we propose a MC framework for sampling pre-calculated perturbed dose distributions computed on the evaluation images. A BH proton plan and a BH and a FB photon plan were evaluated and compared. The results of the proposed robustness analysis for one patient is presented.

Table 1: Standard deviations for the parameters used in the robust planning and evaluation. Σ refers to systematic errors that persist during the whole treatment and σ to random errors that are different at each fraction or breath-hold.

	LR	AP	SI
Patient setup [mm] [4]			
Σ_{ps}	1.1	1.5	1.4
σ_{ps}	1.4	1.7	1.7
Breath-hold tumor position [mm] [5]			
Σ_{bh}	1.3	1.2	1.1
σ_{bh}	0.9	1.0	1.0
Stopping power ratio [%] [6]			
Σ_{ρ}		3.8	

2 Materials and Methods

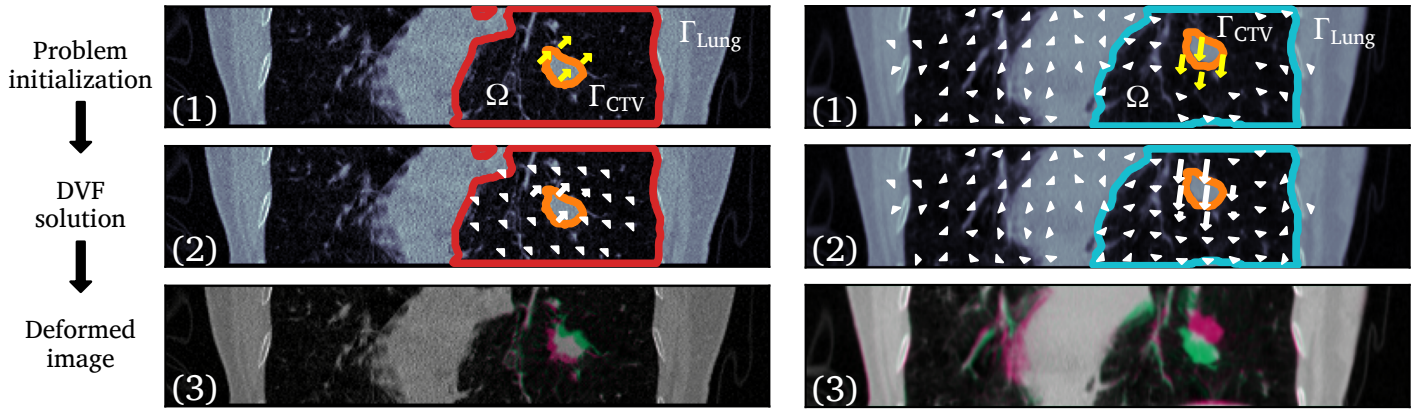
2.1 Patient image data and uncertainty values

The patient data set consisted of one breath-hold CT (BHCT), a 4DCT with ten respiratory phase images and a set of 300 cine-CT acquired at three occasions of approximately eight minutes each. Deformable image registration was used to map the 30% phase image to the other images in the 4DCT and rigid image registration was used to find the cine-CT tumor position in the 4DCT frame of reference [3].

Tabulated values were used to simulate uncertainties and are summarized in Table 1. Patient setup uncertainty was taken from Li *et al.* [4], BH tumor position reproducibility uncertainty from Josipovic *et al.* [5] and proton SPR uncertainty from Li *et al.* [6].

2.2 Robust treatment planning

Robust planning was performed for both the proton and the photon plans to the 90% confidence interval [7]. Perturba-



(a) Breath-hold (Laplace) (b) Free-breathing (Poisson)

Figure 1: Image deformation modeling using (a) Laplace and (b) Poisson editing. The DVF arrows have been scaled for the illustration.

Table 2: Robust optimization perturbations calculated using the values in Table 1 to the 90th confidence interval using the method by Buti *et al.* [7]. The SPR perturbations were only used for proton therapy planning.

	Isotropic shift [mm]	SPR [%]	4DCT
Breath-hold	7.0	4.8	No
Free-breathing	4.0	4.8*	Yes

*Proton therapy was not evaluated in free-breathing.

tions used in the planning were calculated from the standard deviations in Table 1 and can be seen in Table 2. For the FB treatment 4DCT was used to accommodate the respiratory motion.

All three treatment plans had the intention of delivering a dose in three fractions such that CTV $D_{50\%}$ = 54 Gy. The treatment planning and evaluation dose computations were performed with the RayStation treatment planning system (RaySearch Laboratories AB, Stockholm, Sweden). The photon dose computations were performed using collapsed cone and the proton doses using MC with constant relative biological effectiveness factor of 1.1.

2.3 Deformed evaluation images

A simple method of generating smooth displacement vector fields (DVF) for rigid tumor shifts inspired by the classical works on Poisson image [8] and Laplacian surface [9] editing is proposed.

Breath-hold evaluation images were constructed by solving the Laplace equation

$$\nabla^2 u = 0, \quad \text{on } \Omega \tag{1}$$

subject to a zero displacement at the surface of the lung, Γ_{Lung}, and a fixed displacement, a, set for the tumor, Γ_{CTV}

$$\begin{cases} u = 0, & \text{on } \Gamma_{\text{Lung}} \\ u = a, & \text{on } \Gamma_{\text{CTV}} \end{cases} \tag{2}$$

where the geometry and a solution is shown in Figure 1 (a). This problem was solved once for each cardinal direction and the vector field was constructed from each component scalar field, u, and applied to the BHCT.

Values for the tumor displacement were sampled at 15 positions according to a systematic error and for each such offset at 15 positions according to a random error in the BH reproducibility according to the values in Table 1 [6]. In total 225 breath-hold evaluation images were created and used.

Free-breathing evaluation images were constructed by solving the Poisson equation

$$\nabla^2 u = f, \quad \text{on } \Omega \text{ with } f = \nabla^2 v \tag{3}$$

where the source term f was computed from the background displacement with component in one direction, v. This background DVF was computed with deformable image registration that transforms the reference image to the phase-image in the 4DCT with the tumor closest to the evaluation position. This evaluation tumor position is set to one of the positions found in the cine-CT image sequences. The boundary conditions were set to

$$\begin{cases} u = v, & \text{on } \Gamma_{\text{Lung}} \\ u = a, & \text{on } \Gamma_{\text{CTV}} \end{cases} \tag{4}$$

such that the DVF at the lung boundary and outside is set to the original deformable transform, Figure 1 (b). The DVF was applied to the 30% phase image in the 4DCT.

Both the Laplace and the Poisson problems were discretized using finite differences on a coarsely sampled image grid and solved using a sparse solver. The DVFs were inverted to allow dose accumulation by applying the transform to the image element positions and resampling these values on the original grid using nearest-neighbor.

2.4 Evaluation dose and probability computations

A large set of dose distributions were computed for each plan. Systematic patient setup-error was sampled at 15 positions, one with zero shift, six positions on the cardinal axes

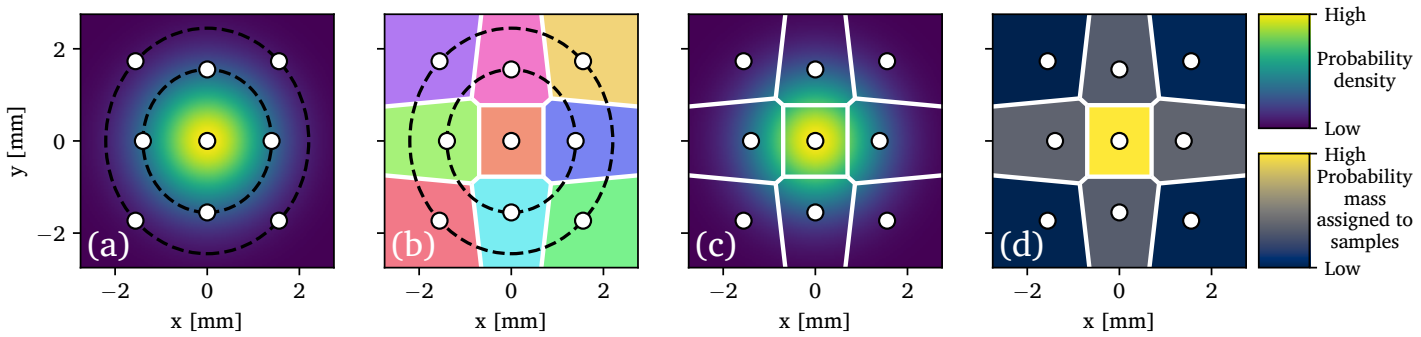


Figure 2: Positioning uncertainty and perturbations. The probability density was modeled using a normal distribution and the probability mass associated with a sample was found by MC integration of the probability density of all points in space closest to the sample.

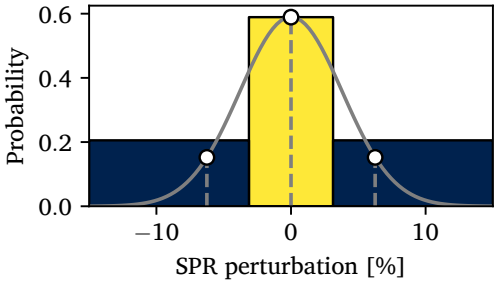


Figure 3: SPR uncertainty and perturbations.

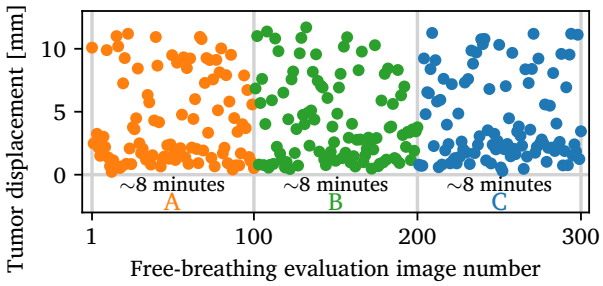


Figure 4: Vector lengths of the cine-CT tumor displacements.

at the 70% confidence interval and eight positions on the axis-aligned box at the 95% confidence interval. The same sampling pattern was used for the random setup error and both the systematic and random breath-hold tumor position error. This sample pattern is illustrated for 2D in Figure 2 where the discretized probability associated with each sample is also shown. These probabilities were calculated using MC integration of the probability density of all points lying closest to the sample as indicated by the Voronoi cells.

This sampling pattern resulted in $15^4 = 50\,625$ dose distributions for the BH photon evaluation and, with three sample positions for SPR uncertainty as seen in Figure 3, 151 875 dose distributions for the proton evaluation.

The free-breathing treatment was simulated using a set of 300 cine-CT derived tumor positions [3] divided into three according to the three imaging occasions, A, B and C in Figure 2. These were used to simulate the dose in one treatment fraction where either of the three sequences were considered as equally probable. This resulted in a total of $15^2 \times 300 = 67\,500$ photon dose distributions that were further accumulated per respiratory sequence into 675 possible fraction doses.

2.5 Monte Carlo scenario simulations and analysis

By MC sampling and accumulation of pre-computed dose distributions treatment scenarios consisting of three fractions were simulated.

For the two BH treatments it was assumed that one treatment field could be delivered in a single BH such that the proton plan was delivered in three BHs per fraction and the photon plan in seven as indicated by the number of beams in Figure 5 (a_1) and (b_1). Thus, one BH treatment scenario was simulated by sampling one SPR position, one systematic patient position, one systematic BH tumor position, then for each of the three fractions a random patient position was sampled and then one random BH tumor position was sampled for each beam direction.

One FB treatment was simulated by sampling one systematic patient position, then for each of the three fractions a random patient position and one of the three respiratory sequences seen in Figure 4 were sampled.

1000 scenarios were simulated per plan. The plans were evaluated at the 90% probability, $\mathbb{P}_{90\%}$, to achieve a certain dose metric for the CTV. The analogous analysis was performed for the organs at risk but $\mathbb{P}_{90\%}$ represents the probability of protecting the structure from higher doses.

3 Results

All three treatment plans achieved a $\mathbb{P}_{90\%}$(CTV $D_{50\%}$) in line with the intended 54 Gy, with the range of values being higher for the proton plan. The two photon plans provided similar results. The proton plan had lower mean heart dose and smaller lung sub-volume at 11.4 Gy compared to the photon plans. The dose to the esophagus was however higher for the proton plan.

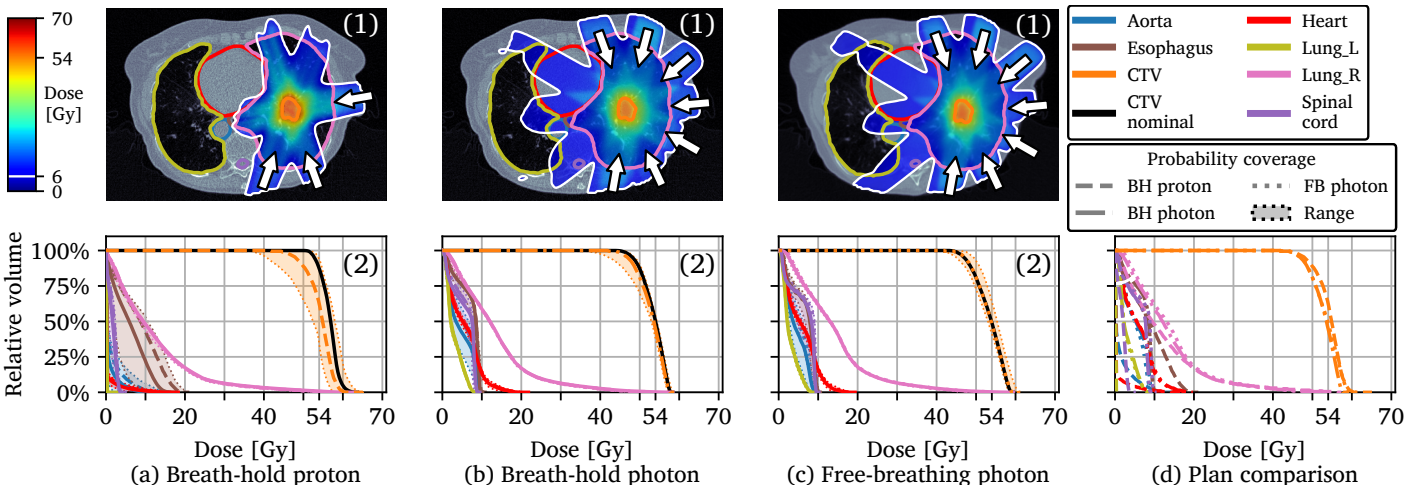


Figure 5: Nominal plan dose distributions, truncated at 6 Gy, and DVH bands. Nominal DVHs are shown as solid lines, while the $\mathbb{P}_{90\%}$ to achieve a coverage for the CTV and protection for any of the organs at risk are shown as dashed, dash-dotted and dotted lines for the three plans respectively in (a), (b) and (c) where all are shown together in (d). The range of the DVHs are shown as brighter bands.

Table 3: Resulting dose metrics.

	BH proton Nominal [min, max] ($\mathbb{P}_{90\%}$)	BH photon Nominal [min, max] ($\mathbb{P}_{90\%}$)	FB photon Nominal [min, max] ($\mathbb{P}_{90\%}$)
CTV $D_{50\%}$ [Gy]	56.9 [52.5, 58.0] (54.9)	54.0 [53.0, 54.4] (53.6)	53.9 [53.0, 55.4] (53.9)
Heart D_{mean} [Gy]	0.5 [0.2, 0.8] (0.6)	5.7 [5.3, 6.1] (5.8)	5.8 [5.5, 6.2] (6.0)
Lung $V_{11.4\,\text{Gy}}$ [cm^3]	385 [348, 390] (386)	475 [464, 479] (477)	468 [457, 480] (473)
Esophagus D_{max} [Gy]	17.5 [12.2, 21.2] (19.8)	9.6 [9.2, 12.0] (10.2)	9.9 [9.5, 10.2] (10.0)

4 Discussion

Although the proton plan did not have consistently lower doses for all organs at risk the reduced heart dose could be of clinical benefit. The quality of the proton plan will however depend on the beam angles, here at the cost of high dose to the esophagus. The DVF creation method was not validated against images with known displacements. The proposed discrete sampling schema imposed long computation times and not all dose distributions were used in the simulations.

5 Conclusion

In this study a MC based framework for the evaluation of proton and photon therapy was proposed. BH and FB plans were compared and proton BH plan demonstrated a reduced dose to the heart and lung compared to the two photon plans at the cost of higher dose to the esophagus.

References

[1] E. D. Brooks, M. S. Ning, V. Verma, et al. “Proton therapy for non-small cell lung cancer: The road ahead”. *Translational Lung Cancer Research* 8.Suppl 2 (2019), S202–S212. DOI: 10.21037/tlcr.2019.07.08.

[2] K. Souris, A. Barragan Montero, G. Janssens, et al. “Technical Note: Monte Carlo methods to comprehensively evaluate the robustness of 4D treatments in proton therapy”. *Medical Physics* 46.10 (2019), pp. 4676–4684. DOI: 10.1002/mp.13749.

[3] N. Olofsson, K. Wikström, A. Flejmer, et al. “Dosimetric robustness of lung tumor photon radiotherapy evaluated from multiple event CT imaging”. en. *Physica Medica* 103 (Nov. 2022), pp. 1–10. DOI: 10.1016/j.ejmp.2022.09.007.

[4] W. Li, T. G. Purdie, M. Taremi, et al. “Effect of Immobilization and Performance Status on Intrafraction Motion for Stereotactic Lung Radiotherapy: Analysis of 133 Patients”. en. *International Journal of Radiation Oncology Biology Physics* 81.5 (Dec. 2011), pp. 1568–1575. DOI: 10.1016/j.ijrobp.2010.09.035.

[5] M. Josipovic, M. C. Aznar, J. B. Thomsen, et al. “Deep inspiration breath hold in locally advanced lung cancer radiotherapy: validation of intrafractional geometric uncertainties in the INHALE trial”. en. *The British Journal of Radiology* 92.1104 (Dec. 2019), p. 20190569. DOI: 10.1259/bjr.20190569.

[6] B. Li, H. C. Lee, X. Duan, et al. “Comprehensive analysis of proton range uncertainties related to stopping-power-ratio estimation using dual-energy CT imaging”. en. *Physics in Medicine & Biology* 62.17 (Aug. 2017), pp. 7056–7074. DOI: 10.1088/1361-6560/aa7dc9.

[7] G. Buti, K. Souris, A. M. B. Montero, et al. “Towards fast and robust 4D optimization for moving tumors with scanned proton therapy”. *Medical Physics* 46.12 (2019), pp. 5434–5443. DOI: 10.1002/mp.13850.

[8] P. Perez, M. Gangnet, and A. Blake. “Poisson image editing”. en. *ACM Transactions on Graphics* 22.3 (2003), pp. 313–318. DOI: 10.1145/882262.882269.

[9] O. Sorkine, D. Cohen-Or, Y. Lipman, et al. “Laplacian surface editing”. en. *Proceedings of the 2004 Eurographics/ACM SIGGRAPH symposium on Geometry processing*. Nice France: ACM, July 2004, pp. 175–184. DOI: 10.1145/1057432.1057456.

Real-Time CBCT Imaging via a Single Arbitrarily-Angled X-ray Image using a Joint Dynamic Reconstruction and Motion Estimation (DREME) Framework

Hua-Chieh Shao[1], Tielige Mengke[1], Tinsu Pan[2], and You Zhang[1]

[1] Department of Radiation Oncology, UT Southwestern Medical Center, Dallas, TX, 75390, USA
[2] Department of Imaging Physics, UT MD Anderson Cancer Center, Houston, TX, 77030, USA

Abstract Real-time cone-beam CT (CBCT) imaging provides instantaneous patient anatomy for image guidance and online treatment adaptation in radiotherapy. However, real-time CBCT imaging faces a stringent temporal constraint (<500 ms) under respiration-induced anatomical motion, leading to extreme under-sampling (only a few x-ray projections can be acquired). We proposed a joint dynamic CBCT reconstruction and motion estimation (DREME) framework to address this challenge. DREME reconstructs a dynamic CBCT from a pre-treatment conventional CBCT scan, while simultaneously deriving a deep learning-based, angle-agnostic model for real-time motion and CBCT estimation based on a single X-ray projection. Specifically, DREME uses a prior model-free spatiotemporal implicit neural representation (PMF-STINR) framework to reconstruct dynamic CBCTs and derive a data-driven motion model. The motion model, which captures the up-to-date and patient-specific motion patterns, is used by DREME to train a deep learning model to estimate real-time motion and CBCT from an arbitrarily-angled X-ray projection. DREME was evaluated using digital phantom simulations and a dataset of lung patients. The results showed that DREME can accurately solve 3D motion with high accuracy (phantom: 1.0±0.4 mm, and patients: 1.3±1.2 mm) in real time (~1.5 ms inference time for each X-ray projection).

1 Introduction

The technical advancements of radiotherapy have significantly improved its accuracy over the past few decades. However, intra-treatment motion, particularly respiration-induced motion, remains a major source of uncertainty during the radiation delivery. Real-time adaptive radiotherapy that tracks and adapts to the volumetric motion in real time can ultimately solve the challenge. However, for respiratory motion (~3-5 s per cycle), the temporal latency of real-time imaging should be < 500 ms. Within such a short duration, only several X-ray projections can be acquired for CBCT imaging and motion tracking, leading to extreme under-sampling.

Under the extreme under-sampling, conventional methods fell short in reconstructing CBCT images for volumetric motion tracking. Recently, deep learning (DL)-based approaches have been proposed for real-time imaging and motion estimation, due to their imaging processing power and fast inference time. The DL-based methods include reconstruction-based [1] and registration-based [2] approaches. Reconstruction-based methods used neural networks to directly reconstruct a 3D image from extremely limited projections (a single projection as in [1]). However, direct 2D-to-3D reconstruction without access to additional information makes a highly ill-posed inverse problem, rendering such models susceptible to robustness and generalization issues. Additional registration/segmentation is also needed after the reconstruction for motion tracking, which results in additional latency. Alternatively, registration-based methods directly solved motion fields on existing prior images for tracking. For example, principal component analysis-based motion model from a prior 4D-CT was used for real-time motion estimation [2]. While the prior information significantly alleviates the ill-posed condition, it nevertheless introduces additional biases as the prior image or the motion model can be outdated due to inter-fractional anatomical and motion variations.

To address the above challenges, we proposed to combine a recently-developed dynamic CBCT reconstruction method [3] with a DL-based, real-time motion estimation model into a joint dynamic reconstruction and motion estimation (DREME) framework for real-time motion and CBCT estimation. DREME uses a prior model-free spatiotemporal implicit neural representation (PMF-STINR) method for dynamic CBCT reconstruction. PMF-STINR reconstructs the latest patient-specific dynamic CBCTs and extracts an up-to-date motion model directly from pre-treatment onboard projections to capture both regular and irregular motion. Based on the PMF-STINR motion model, DREME further trains a CNN-based DL model to extract motion-related image features from an arbitrarily-angled X-ray projection to predict the 3D motion field and estimate real-time CBCTs. DREME was evaluated by the extended cardiac torso (XCAT) digital phantom and a dataset of lung patients, to assess its real-time CBCT imaging and motion tracking accuracy.

2 Materials and Methods

2.1 DREME workflow overview

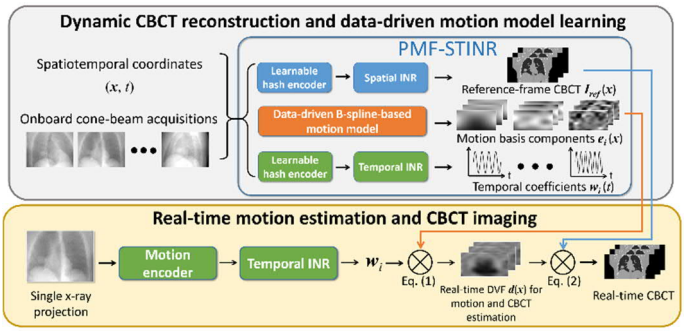


Fig. 1. Workflow of the DREME framework. (*Top panel*) In the first step of DREME, a dynamic CBCT sequence and a motion model is derived from a standard CBCT acquisition by PMF-STINR. PMF-STINR takes spatial and temporal coordinates ($\boldsymbol{x}$, t) and cone-beam projections as inputs, and reconstructs a reference-frame CBCT $\boldsymbol{I}_{ref}(\boldsymbol{x})$, a motion model of motion basis components $\boldsymbol{e}_i(\boldsymbol{x})$ (i=1-3), and corresponding temporal coefficients $\boldsymbol{w}_i(t)$ (i =1-3). The motion model, scaled by the temporal coefficients, will generate DVFs to deform $\boldsymbol{I}_{ref}(\boldsymbol{x})$ to dynamic CBCTs. (*Bottom panel*) In the second step, the learned reference-frame CBCT and motion model are used to train a motion encoder on arbitrarily-angled X-ray projections to estimate real-time DVFs for motion and CBCT estimation.

Figure 1 illustrates the two-step workflow of DREME. In the first step, PMF-STINR [3] was employed to reconstruct a dynamic CBCT sequence and a patient-specific motion model from a standard free-breathing CBCT scan. The dynamic CBCTs were generated by reconstructing a reference-frame CBCT $\boldsymbol{I}_{ref}(\boldsymbol{x})$ and solving time-varying deformation vector fields (DVFs) $\boldsymbol{d}(\boldsymbol{x}, t)$ with respect to $\boldsymbol{I}_{ref}(\boldsymbol{x})$. Here, $\boldsymbol{x}$ and t represent the spatial voxel coordinates and temporal frame index of the dynamic CBCTs, respectively. A low-rank representation of $\boldsymbol{d}(\boldsymbol{x}, t)$ was employed for dimension reduction:

$$\boldsymbol{d}(\boldsymbol{x}, t) = \sum_{i=1}^{L} \boldsymbol{w}_i(t) \times \boldsymbol{e}_i(\boldsymbol{x}), \quad (1)$$

where $\{\boldsymbol{w}_i(t)\}_{i=1}^{L}$ and $\{\boldsymbol{e}_i(\boldsymbol{x})\}_{i=1}^{L}$ respectively denote the temporal coefficients and spatial components of $\boldsymbol{d}(\boldsymbol{x}, t)$, and L denotes the number of spatial components in the low-rank representation. Here L=3 was used, as it has been demonstrated sufficient to accurately describe respiratory motion in previous works. $\{\boldsymbol{e}_i(\boldsymbol{x})\}_{i=1}^{3}$ made a set of basis components spanning a functional space of motion, rendering a motion model. The whole dynamic CBCT sequence was obtained by deforming $\boldsymbol{I}_{ref}(\boldsymbol{x})$ via:

$$\boldsymbol{I}(\boldsymbol{x}, t) = \boldsymbol{I}_{ref}(\boldsymbol{x} + \boldsymbol{d}(\boldsymbol{x}, t)). \quad (2)$$

In the second step, DREME leveraged the learned onboard motion model $\{\boldsymbol{e}_i(\boldsymbol{x})\}_{i=1}^{3}$ and the reference-frame image $\boldsymbol{I}_{ref}(\boldsymbol{x})$ for real-time motion and CBCT estimation. Specifically, it trained a convolution-based motion encoder to extract motion-related features from an arbitrarily-angled X-ray projection. The motion encoder output a motion feature vector which was then fed into the inherited temporal INR from PMF-STINR to predict the corresponding temporal coefficients $\{\boldsymbol{w}_i\}_{i=1}^{3}$. The predicted $\{\boldsymbol{w}_i\}_{i=1}^{3}$, combined with the motion model $\{\boldsymbol{e}_i(\boldsymbol{x})\}_{i=1}^{3}$, yielded real-time DVFs $\boldsymbol{d}(\boldsymbol{x})$ for motion tracking and CBCT estimation. In the following, we elaborated the details of PMF-STINR and the 2nd step motion encoder, along with the evaluation schemes of DREME.

2.2 Dynamic CBCT reconstruction (PMF-STINR)

As illustrated in Fig. 1, PMF-STINR consisted of three modules, separately and simultaneously solving $\boldsymbol{I}_{ref}(\boldsymbol{x})$, $\{\boldsymbol{e}_i(\boldsymbol{x})\}_{i=1}^{3}$, and $\{\boldsymbol{w}_i(t)\}_{i=1}^{3}$. The overall reconstruction scheme was based on an emerging machine learning technique named implicit neural representation (INR) [4]. Acting as a non-parametrized universal approximator, the INRs can be used to map complex spatial images (spatial INR) and temporal variations (temporal INR) with lightweight multi-layer perceptrons to promote sparsity and under-sampled learning. PMF-STINR used a spatial INR to represent and map the reference-frame CBCT $\boldsymbol{I}_{ref}(\boldsymbol{x})$. A learnable hash encoder [5] was employed together with the spatial INR to improve the learning capacity of high-frequency image features. In parallel, PMF-STINR also used a temporal INR and a B-spline-based data-driven model to capture the motion dynamics $\{\boldsymbol{w}_i(t)\}_{i=1}^{3}$ and $\{\boldsymbol{e}_i(\boldsymbol{x})\}_{i=1}^{3}$ from the cone-beam projections, respectively. The combined temporal INR and the data-driven motion model solved the time-varying DVFs $\boldsymbol{d}(\boldsymbol{x}, t)$ with respect to $\boldsymbol{I}_{ref}(\boldsymbol{x})$ to yield the dynamic CBCTs (Eq. 2). The temporal INR shared an architecture similar to the spatial INR. The data-driven motion model was learned on-the-fly during PMF-STINR reconstruction, without relying on any patient-specific prior information. To regularize the motion learning, $\{\boldsymbol{e}_i(\boldsymbol{x})\}_{i=1}^{3}$ was parametrized by a cubic B-spline interpolant, and the spatial resolution of the B-spline control points was doubled for each higher level. At each level, the B-spline control point values are learnable parameters.

The training of PMF-STINR was driven by a projection-domain similarity loss (mean squared error) and a total variation-based image-domain regularization loss, via a one-shot learning scheme. The spatial and temporal INRs and the data-driven motion model were trained simultaneously to improve the consistency and coherence between $\boldsymbol{I}_{ref}(\boldsymbol{x})$, $\boldsymbol{d}(\boldsymbol{x}, t)$, and the cone-beam projections. The one-shot learning does not require any additional training dataset and avoids the generalization issues encountered in most DL-based approaches. PMF-STINR had been demonstrated to accurately capture regular and irregular respiratory motion with high spatial (<=2 mm^3) and temporal (~0.1 s) resolutions. Further details of PMF-STINR can be found in our recent manuscript [3].

2.3 Real-time motion and CBCT estimation

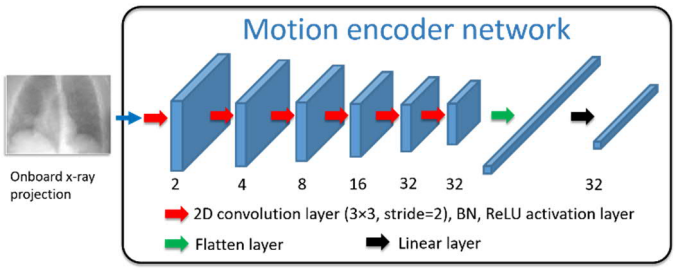


Fig. 2. Network architecture of the angle-agnostic motion encoder. The input of the motion encoder was an arbitrarily-angled X-ray projection, and the motion-related image features were extracted by a series of 2D convolution layers to yield a 32-element feature vector.

The dynamic reconstruction by PMF-STINR yielded an up-to-date, patient-specific motion model; and a reference-frame CBCT reflecting the latest patient anatomy. To use the same motion model and reference CBCT for real-time motion and CBCT prediction, we replaced the hash encoder of the temporal INR by a 2D convolution-based motion encoder (Fig. 2) in the 2nd step. By PMF-STINR, the hash encoder encoded temporal indices for the temporal INR to derive corresponding temporal motion coefficients (Fig. 1). These coefficients were then used to weight the motion model to derive dynamic DVFs and deform the reference CBCT to dynamic CBCTs. Similarly, by the real-time motion estimation framework (Fig. 2), the motion encoder directly encoded an arbitrarily-angled X-ray projection for the temporal INR to derive real-time motion coefficients (Fig. 1). The motion coefficients, combined with the motion model and the reference CBCT, yielded real-time DVFs to generate real-time CBCTs.

To train an angle-agnostic motion encoder, we simulated digitally reconstructed radiographs (DRRs) at random angles from different dynamic CBCTs, which were generated by deforming the reference CBCT with various dynamic DVFs (with known motion coefficients). With the known 'ground-truth' motion coefficients as training labels, we used a supervised learning scheme to train the motion encoder to extract motion-related features from the DRRs for the temporal INR to predict corresponding motion coefficients. X-ray photon Poisson noise and electronic Gaussian noise were simulated and added to the DRRs to improve the model robustness.

2.4 Evaluation datasets and schemes

The overall DREME framework was evaluated by the XCAT phantom and a dataset of 4 lung patients.

2.4.1 XCAT simulation study

We used XCAT to simulate a lung patient. A spherical tumor 30-mm in diameter was inserted in the lower right lung to evaluate the real-time motion estimation accuracy. Six free-breathing motion scenarios (X1-X6) featuring various irregular motion patterns (different breathing amplitudes, breathing periods, baseline drifts/shifts, and their combinations) were simulated. For each motion scenario, dynamic/real-time cone-beam projections were simulated based on a 360° scan with 60 s scanning time and an 11 frames/s projection acquisition rate. We trained DREME on one scenario and then tested it on the remaining scenarios (for instance, trained on X1 and tested on X2-6), yielding six groups of evaluations.

The reconstructed real-time CBCTs were quantitatively evaluated using relative error (RE) and structural similarity index (SSIM), measured against the known 'ground-truth' CBCTs. The real-time motion estimation accuracy was evaluated by comparing the solved motion of the lung tumor against the known 'ground-truth'. Specifically, the lung contour in the reference-frame CBCT was propagated to the real-time frames using the solved real-time DVFs $\boldsymbol{d}(\boldsymbol{x})$. Based on the propagated and 'ground-truth' contours, metrics including Dice similarity coefficient (DSC) and center-of-mass error (COME) were evaluated.

2.4.2 Patient study

DREME was further evaluated on a dataset of four lung patients to assess its performance. Four lung patients from two sources were curated for evaluation (Table I). To evaluate PMF-STINR, a subset of the projections were sparsely-sampled from the full acquisition for PMF-STINR reconstruction, and the remaining projections were used for real-time motion estimation and evaluation. For the MDACC patients (P1-P3), we sparsely-sampled the original projections by selecting every third projection for PMF-STINR reconstruction. For the SPARE patient (P4), both sparse and full sampling datasets were available from the original source [6]. The sparsely-sampled acquisition (340 projections) was used by PMF-STINR, and the rest projections (675 projections) were used to evaluate the accuracy of real-time motion estimation.

TABLE I. Summary of imaging parameters of the patient study.

Patient ID[a]	Source[a]	Projection size[b]	Pixel size (mm^2)	kVp/mA/ms	SAD[c] (mm)/ SDD (mm)
P1	MDACC	512×384×1983	0.776×0.776	120/80/25	1000/1500
P2	MDACC	512×384×2729	0.776×0.776	120/80/25	1000/1500
P3	MDACC	512×384×1653	0.776×0.776	120/80/25	1000/1500
P4	SPARE	512×512×1015	0.8×0.8	125/20/20	1000/1536

[a]MDACC: MD Anderson Cancer Center. SPARE: SPARE Challenge [6].

[b]width (in pixel number) × height (in pixel number) × N_p (number of projections).

[c]SAD: source-to-axis distance. SDD: source-to-detector distance.

Unlike the XCAT study, the patient dataset had no 'ground-truth' real-time motion/CBCT for evaluation. Instead, we evaluated the accuracy of estimated motion at each projection by re-projecting the solved real-time CBCTs to DRRs and comparing with the corresponding X-ray projections. The Amsterdam Shroud (AS) method [7], an intensity-based, projection-domain motion extraction technique, was employed to quantify the motion differences between DRRs and the X-ray projections. By calculating the intensity gradients in the superior-inferior (SI) direction and integrating the resulting images horizontally, we can convert each projection/DRR into a line profile and then consolidate the profiles from different views to track anatomical motion. Prominent anatomies with sharp contrasts, including the diaphragms and the lung nodules, can be tracked from the AS images to compare their motion trajectories between the DRRs and the X-ray projections. We compared the trajectories using the Pearson correlation coefficient and the localization error (average point-wise distance between the two trajectories).

3 Results

3.1 The XCAT results

Table II quantitatively summarizes the mean RE and SSIM of the estimated real-time CBCTs, and mean COME and DSC of the tracked tumors. The small RE and large SSIM show that the estimated real-time CBCTs matched well with the 'ground-truth'. For real-time motion estimation, the tumor can be tracked with ≤1 mm accuracy.

TABLE II. Summary of the mean RE and SSIM of the real-time CBCTs; and the mean COME and DSC of the tracked real-time tumor motion for the XCAT study. The results were presented in terms of mean±SD.

Training motion scenario	Real-time CBCT accuracy		Real-time tumor tracking accuracy	
	RE	SSIM	COME (mm)	DSC
X1	0.143±0.008	0.985±0.002	0.9±0.4	0.904±0.036
X2	0.143±0.015	0.985±0.003	1.0±0.4	0.920±0.031
X3	0.144±0.019	0.984±0.004	1.0±0.4	0.915±0.028
X4	0.152±0.018	0.982±0.005	1.1±0.5	0.917±0.029
X5	0.166±0.006	0.980±0.002	1.0±0.5	0.916±0.022
X6	0.141±0.009	0.985±0.002	0.8±0.4	0.913±0.030

Figure 3 shows the solved real-time XCAT tumor motion trajectories in the SI and anterior-posterior (AP) directions. Due to the space limit, only the DREME results of training scenarios X3-X5 were presented. Regardless of the training scenarios, the solved real-time tumor motion in each testing scenario matched accurately with the 'ground-truth', even if the motion irregularities in the testing scenarios were not seen in the training scenarios.

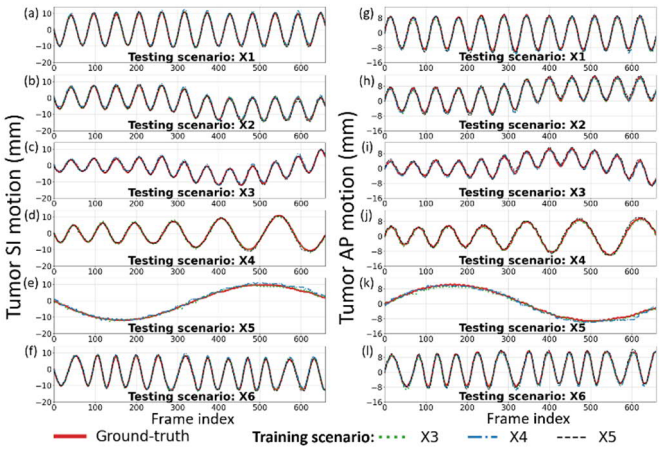


Fig. 3. Real-time XCAT tumor tracking results. The two columns show the predicted real-time tumor trajectories in the SI and AP directions, respectively.

3.2 The patient results

Figure 4 presents example AS images (a-c), and the correspondingly tracked DRR and X-ray projection motion trajectories for the patient study (d-g). Table III summarizes the corresponding localization errors and the Pearson correlation coefficients for the tracked real-time motion. The solved real-time motion trajectories achieved 1.5 to 2.4 mm average accuracy for different patients. Considering the magnification factor (~1.5) of cone-beam projections, the average tracking error reduced to 1.3±1.2 mm in the Cartesian space.

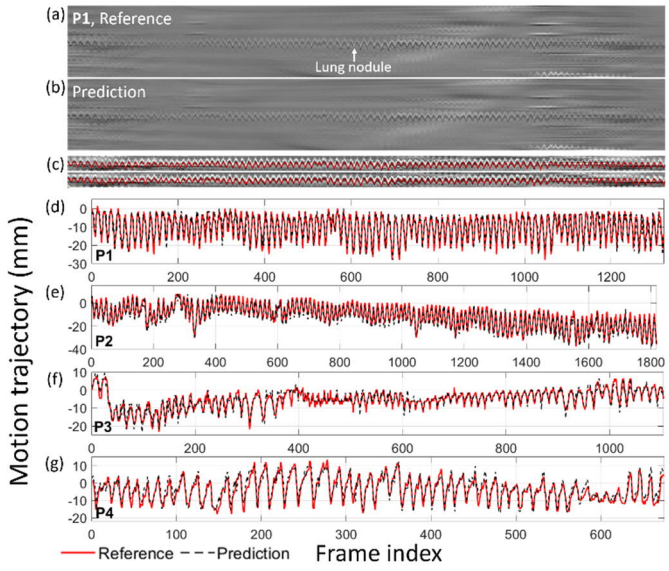


Fig. 4. **(a-b)** Amsterdam Shroud (AS) images of the P1 patient calculated from (a) the onboard X-ray projections and (b) re-projected DRRs. **(c)** Cropped AS images of the lung nodule region of (a) and (b), based on which the motion trajectories were extracted. **(d-g)** The extracted motion trajectories of the four lung patients (P1-P4).

TABLE III. Summary of the real-time motion estimation accuracy of the patient study. The results were presented in terms of mean±SD.

Patient ID	Pearson correlation coefficient	Localization error in projection space (mm)	Localization error in Cartesian space (mm)
P1	0.92	1.9±1.8	1.3±1.2
P2	0.96	2.4±2.0	1.6±1.5
P3	0.91	1.5±1.5	1.0±1.0
P4	0.93	2.1±1.7	1.4±1.1

Figure 5 presents the solved lung nodule motion curves and real-time CBCT images of P1 in three views. The mean inference time is ~1.5 ms per projection.

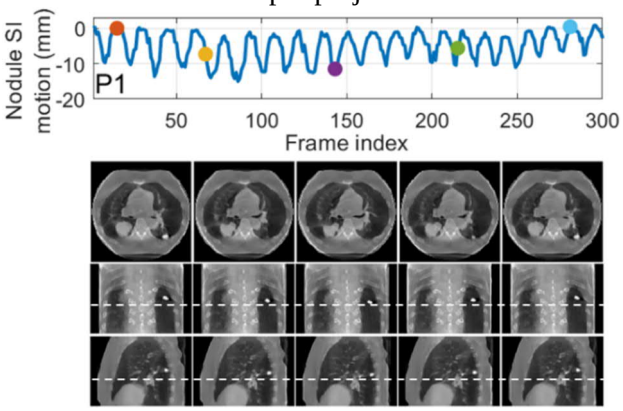


Fig. 5. Example real-time CBCTs solved for P1. The top panel shows the solved lung nodule motion in the first 300 frames, and the dots indicate the time points used to plot the real-time CBCTs in 3 views.

4 Conclusion

We developed a joint dynamic reconstruction and motion estimation framework, DREME, for real-time volumetric motion and CBCT estimation. DREME uses the latest onboard CBCT and motion model information, extracted from dynamic CBCT reconstructions, to solve the following real-time motion. The XCAT and patient results showed that DREME solves accurate real-time motion and CBCTs from arbitrarily-angled X-ray projections, paving the way for future real-time adaptive radiotherapy.

References

[1] L. Y. Shen, W. Zhao, and L. Xing, "Patient-specific reconstruction of volumetric computed tomography images from a single projection view via deep learning," (in English), *Nat Biomed Eng*, vol. 3, no. 11, pp. 880-888, Nov 2019, doi: 10.1038/s41551-019-0466-4.

[2] R. Wei *et al.*, "Real-time tumor localization with single x-ray projection at arbitrary gantry angles using a convolutional neural network (CNN)," (in English), *Physics in Medicine and Biology*, vol. 65, no. 6, p. 065012, Mar 2020, doi: ARTN 065012 10.1088/1361-6560/ab66e4.

[3] H.-C. Shao, M. Tielige, T. Pan, and Y. Zhang, "Dynamic CBCT Imaging using Prior Model-Free Spatiotemporal Implicit Neural Representation (PMF-STINR)," 2023, doi: https://arxiv.org/abs/2311.10036.

[4] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis," (in English), *Commun Acm*, vol. 65, no. 1, pp. 99-106, Jan 2022. [Online]. Available: <Go to ISI>://WOS:000731456200022.

[5] T. Muller, A. Evans, C. Schied, and A. Keller, "Instant Neural Graphics Primitives with a Multiresolution Hash Encoding," (in English), *Acm T Graphic*, vol. 41, no. 4, Jul 2022, doi: Artn 102 10.1145/3528223.3530127.

[6] C. C. Shieh *et al.*, "SPARE: Sparse-view reconstruction challenge for 4D cone-beam CT from a 1-min scan," *Med Phys*, vol. 46, no. 9, pp. 3799-3811, Sep 2019, doi: 10.1002/mp.13687.

[7] L. Zijp, J. J. Sonke, and M. van Herk, "Extraction of the respiratory signal from sequential thorax cone-beam X-ray images," *International conference on the use of computers in radiation therapy*, pp. 507-509, 2004.

Enhancing Clinical Utility of 0.55T 4DMRI in Radiotherapy Planning: Exploring Compressed Sensing, Denoising, and Auto-segmentation

Austen Curcuru[1], Young Hun Yoon[1, 2, 3], Shanti Marasini[1], H Michael Gach[1, 4, 5], Jianing Pang[6], Xinzhou Li[6], Xin Miao[6], Jin Sung Kim[2,3], and Taeho Kim[1]

[1] Department of Radiation Oncology, Washington University School of Medicine, St. Louis, MO, USA
[2] Department of Radiation Oncology, Yonsei Cancer Center, Seoul, Korea
[3] Medical Physics and Biomedical Engineering Lab (MPBEL), Yonsei University College of Medicine, Seoul, South Korea
[4] Department of Radiology, Washington University School of Medicine, St. Louis, MO, USA
[5] Department of Biomedical Engineering, Washington University in St. Louis, St. Louis, MO, USA
[6] Siemens Medical Solutions USA Inc., Malvern, PA

Abstract Respiratory self-gating 4DMRI offers superior soft tissue contrast compared with 4DCT without imparting any ionizing radiation or requiring an external surrogate device. Recent advancements in MR hardware and sequence design have made 4DMRI simulation an increasingly attractive option for patients with tumors in the thoracoabdominal area that may experience several centimeters of motion due to respiration. This work investigates different respiratory trace binning options and compressed sensing reconstruction in a vendor designed 4DMRI sequence. Five volunteers were scanned on a 0.55 T MRI. The k-space data was retrospectively reconstructed using both phase and amplitude binning, with 5 and 10 total bins, and both with and without compressed sensing. A denoising algorithm was applied to the reconstructed images to improve image quality. A deep learning auto-segmentation network was used to determine whether the images were of sufficient quality to produce stable segmentation results.

1 Introduction

Magnetic resonance imaging (MRI) offers superior soft tissue contrast to computed tomography (CT) imaging without imparting any ionizing radiation to the imaging subject. MRI additionally provides more variety in image quality and options compared with CT.[1] As a result, the use of MRI has increased substantially in radiation therapy planning where accurate tumor and organ-at-risk (OAR) delineation is critical to safe and effective treatment.

In radiotherapy, 4DCTs are commonly used to generate planning images throughout the breathing cycle to assess tumor and OAR motion. Respiratory correlated 4DMRI is a recent technological advancement that can also provide fundamental motion information for tumors in the abdomen or thorax that may experience several centimeters of motion due to patient respiration.[2] 4DMRI imaging data is generated without ionizing radiation, offers superior soft tissue contrast to 4DCT, and allows for imaging in any orientation whereas 4DCT is limited to transverse imaging.

As 4DMR imaging becomes commercially available and more widely used in radiation therapy planning, it is imperative that efficient workflows and tools be developed to aid in clinical utility. This includes investigating adequate binning parameters and ensuring adequate image quality. Additionally, time spent treatment planning can be substantially reduced using deep learning auto-segmentation (DLAS) algorithms. In this study, we are examining a prototype 4DMRI sequence using a 0.55 T MRI scanner. Our investigation involves exploring various binning strategies, reconstructing the MRIs both with and without compressed sensing, and implementing an additional denoising filter. The resulting images are additionally evaluated using an in-house developed DLAS network to assess the feasibility of auto segmentation.

2 Materials and Methods

Five healthy volunteers were recruited for 4DMR imaging on a 0.55-T MRI system (MAGNETOM Free.Max, Siemens Healthineers, Erlangen, Germany) in accordance with an Institutional Review Board approved protocol. Images were acquired during free breathing using a prototype T_1-weighted gradient echo sequence using a phased array torso coil. The sequence utilizes a 3-D "stack of stars" k-space trajectory allowing for self-navigation.[3,4] Because the sequence uses the signal from the center of k-space to generate a respiratory motion trace, no external surrogate signal is required to sort the 4D data into the appropriate respiratory bins. Furthermore, the golden-angle radial k-space trajectory results in incoherent under-sampling artifacts and is sparse in both the spatial domain and along the direction of the motion state.[5] These properties make the results suitable for compressed sensing (CS) reconstruction. Scans were acquired using 2000 radial views to be reconstructed without CS and using 1000 radial views to be reconstructed with CS. Both types of reconstructions were performed retrospectively. The CS and non-CS k-space data were both reconstructed four times using two different bin numbers (5, 10) and binning types (amplitude, phase). This resulted in eight 4D image sets for each volunteer. Additionally, exhalation breath-hold 3D balanced steady state free procession (bSSFP) images were acquired for each volunteer to use as reference.

To improve the quality of the reconstructed 4D datasets, a block matching 3D filtering (BM3D) method was used for denoising. This model divides images into overlapping blocks and collaboratively filters similar blocks using a Wiener filter in the 3D transform domain.[6] BM3D preserves edges while reducing image noise. Denoised images were evaluated against images without denoising using a natural image quality evaluator (NIQE) score.[7]

To evaluate clinical feasibility, a DLAS network was used to segment the liver on both the original and denoised 4DMRI data sets. The DLAS network was based on MONAI (NVIDIA, Santa Clara, CA, USA) DynUnet, which is a modified nnU-Net.[8] Model training utilized an AdamW optimizer and a combined Dice Similarity Coefficient/Cross Entropy loss function. The network was previously trained on images acquired from 27 thoracoabdominal patients on a 1.5 T MRI system (Ingenia, Philips, Amsterdam, Netherlands) using a respiratory-gated T_1-weighted 3D-Axial magnetization-prepared rapid gradient-echo (MP-RAGE) sequence. Test time augmentation (TTA) was utilized to ensure stable segmentation results. Preprocessing inferences for TTA included foreground cropping, Gaussian sharpening with three different configurations, and random contrast adjustment. The five segmentation results from each preprocessing inference were averaged to form the DLAS output.

To further refine the quality of segmentation, a multi-step post-processing strategy was adopted following TTA. First, any segmentation results outside of the body were discarded. Next, Connected Component 3D (CC3D) was performed to isolate and retain only the largest segmented portion of the image.[9] Finally, binary erosion/dilation was used to detach any spurious segments that were connected to correctly segmented areas before CC3D and small object removal were repeated. Volumetric Dice similarity coefficients (DSC3D) were then calculated on the resulting segmentation results with segmentation from the 3D breath-hold bSSFP images serving as the reference.

3 Results

4D images reconstructed with CS displayed significantly ($p << 0.01$) improved NIQE scores (mean NIQE score: 12.83 ± 0.86) compared with images reconstructed without CS (mean NIQE score: 27.45 ± 16.08). Additionally, CS results displayed less visual noise and produced more stable images than the reconstructed images without CS. Amplitude binning and phase binning resulted in similar NIQE scores for images reconstructed with CS (amplitude: 12.59 ± 0.81, phase: 12.98 ± 0.85). For the images reconstructed without CS, amplitude binning produced much better results than phase binning (amplitude: 13.97 ± 0.44, phase: 31.04 ± 16.32). In all cases, the 4D MRI with 5 bins outperformed the one with 10 bins in terms of NIQE score, primarily because it had a larger amount of data available per respiratory bin.

BM3D denoising modestly improved mean NIQE scores for 4D images reconstructed with CS (BM3D: 12.17 ± 0.98, no denoising: 12.83 ± 0.86). Images reconstructed without CS saw a dramatic improvement in mean NIQE score following BM3D denoising (BM3D: 12.94 ± 1.44, no denoising: 27.44 ± 16.08). This difference in BM3D denoising efficacy is expected due to the increased noise present in the 4DMRI images that were reconstructed without CS. Images pre and post denoising are shown for amplitude binning (Figure 1) and for phase binning (Figure 2).

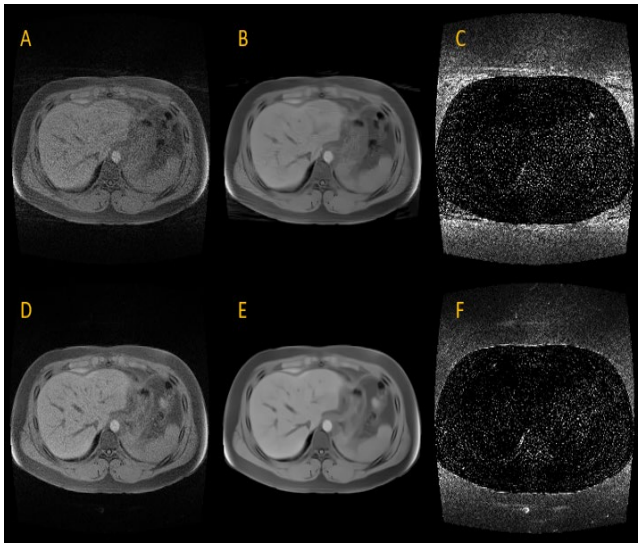


***Figure 1**: 45% **amplitude** binned images reconstructed both without CS (top row) and with CS (bottom row). Images in the leftmost column (A, D) are shown prior to BM3D denoising. Images in the center column (B, E) are shown post BM3D denoising. The rightmost column shows the difference image.*

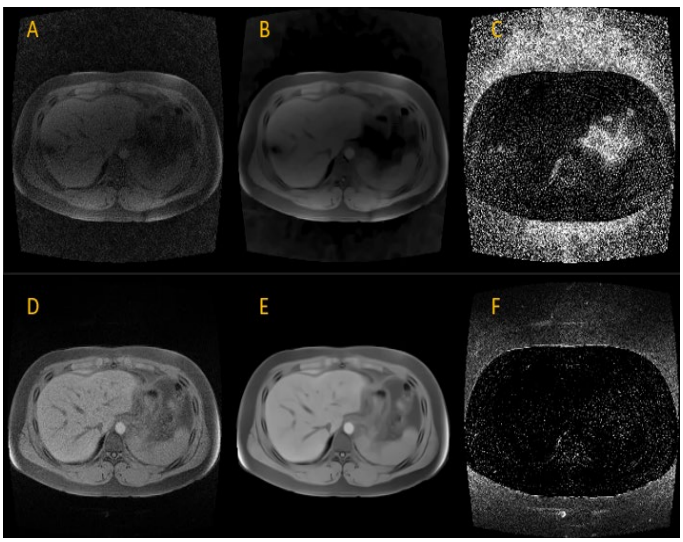


***Figure 2:** 45% **phase** binned images reconstructed both without CS (top row) and with CS (bottom row). Images in the leftmost column (A, D) are shown prior to BM3D denoising. Images in the center column (B, E) are shown post BM3D denoising. The rightmost column shows the difference image.*

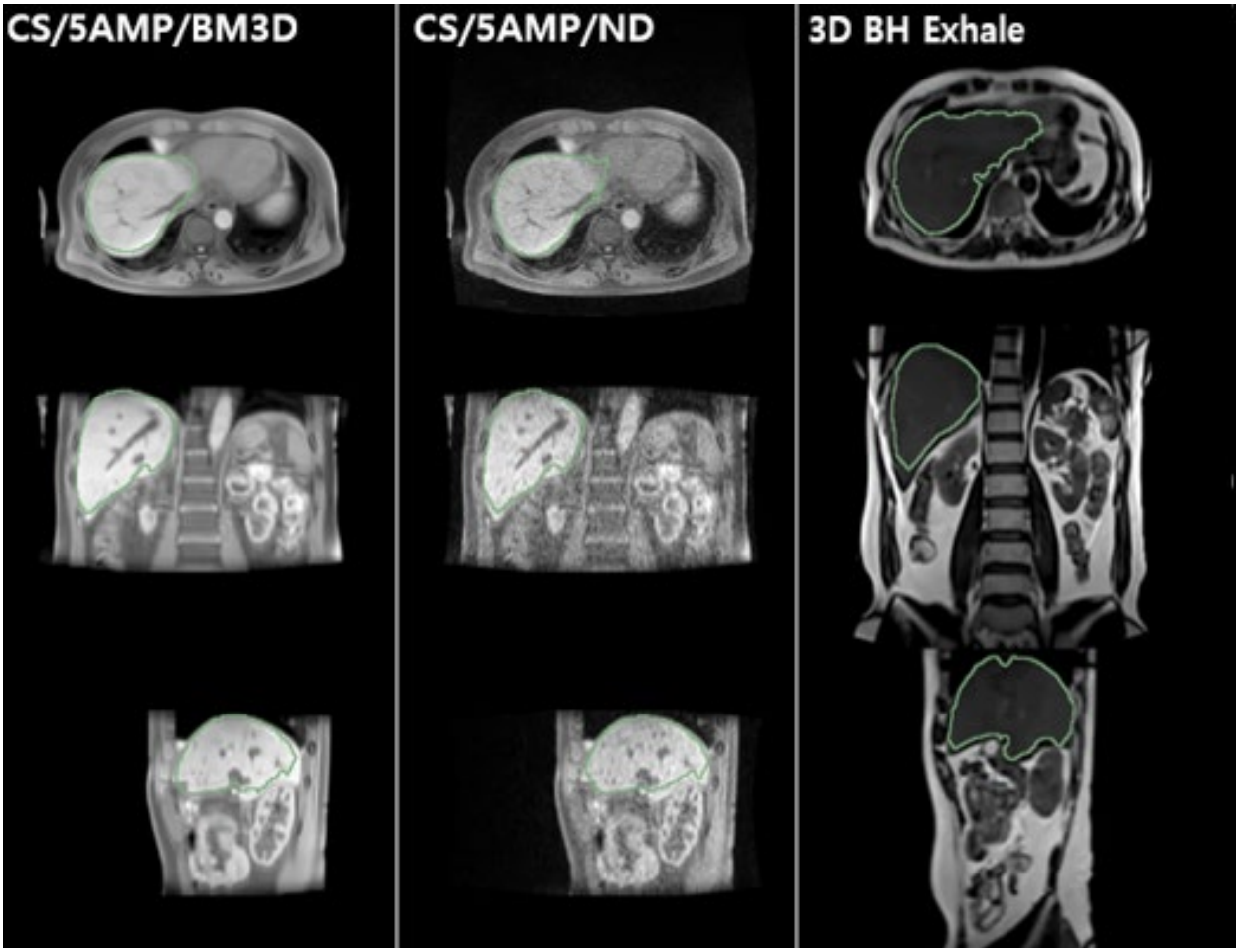


Figure 3: CS 5 amplitude bin reconstructed images with BM3D denoising (first column) and without denoising (second column). The 3D exhalation breath-hold image used as a reference image is shown in the third column. The DLAS generated liver contour is shown in green on all figures.

Segmentation was improved following denoising (DSC3D = 0.83 ± 0.11 with BM3D vs 0.80 ± 0.18 without BM3D) as shown in Figure 3. This result held for both images reconstructed with and without CS. Likewise, images reconstructed with CS showed significantly better ($p < 0.05$) DSC3D results compared to images reconstructed without CS (DSC3D = 0.87 ± 0.06 with CS vs 0.74 ± 0.14 without CS). DLAS model architecture is shown in Figure 4.

4 Discussion

Respiratory self-gating 4DMRI offers the potential for more precise motion management strategies and improved thoracoabdominal target delineation. However, these benefits are contingent upon efficient workflows that produce high quality images. While the CS reconstructed results produced satisfactory results, several phase binned 4DMRI datasets reconstructed without CS did not. Utilizing denoising algorithms such as BM3D can significantly improve non-CS phase binned reconstructions, although denoising does not always ensure segmentation performance. Subject variation additionally plays a critical role in both image and segmentation quality particularly for phase binned imaging studies.

The preliminary results in this work are promising. However, this study can be expanded upon in several ways. For example, alternative binning strategies such as dual amplitude and phase binning may become beneficial in the presence of motion hysteresis.[10] Additionally, with more volunteers the DLAS model could be trained on 0.55T images directly which should yield even better segmentation results. Future work will also focus on a wider variety of OARs. Finally, the relatively regular breathing traces seen in healthy volunteers may not be possible for patients undergoing MR simulation. Dynamic MRI phantom studies that can utilize respiratory traces acquired on surrogate systems during 4DCT simulation are warranted.

5 Conclusion

CS reconstruction produced lower NIQE scores and displayed better segmentation results than non-CS reconstruction despite being highly undersampled. Amplitude binning resulted in higher quality images and more stable results, especially for non-CS reconstructions. BM3D denoising improved image quality and aided in

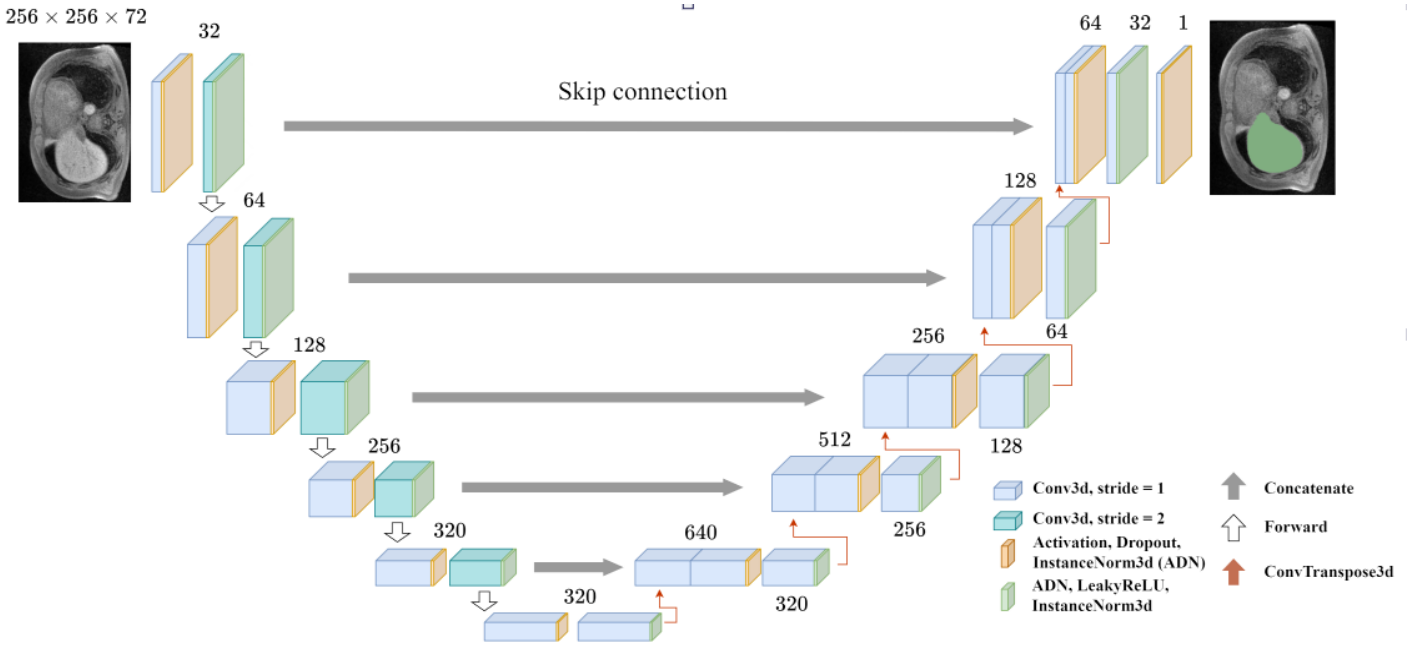


Figure 4: Deep Learning Auto-Segmentation (DLAS) network architecture. The model is based on MONAI DynUnet.

segmentation. Future work is warranted to further improve image quality and segmentation results.

Acknowledgements

The research was sponsored by the Washington University in St. Louis Department of Radiation Oncology. Washington University in St. Louis has Master Research Agreements (MRAs) with Siemens who supplied the 4D MRI work in progress (WIP) sequence and Philips.

References

1 Stemkens, B., Paulson, E. S. & Tijssen, R. H. N. Nuts and bolts of 4D-MRI for radiotherapy. *Phys Med Biol* **63**, 21TR01 (2018). https://doi.org/10.1088/1361-6560/aae56d

2 Zhang, J. *et al.* Clinical evaluation of 4D MRI in the delineation of gross and internal tumor volumes in comparison with 4DCT. *J Appl Clin Med Phys* **20**, 51-60 (2019). https://doi.org/10.1002/acm2.12699

3 Feng, L. *et al.* XD-GRASP: Golden-angle radial MRI with reconstruction of extra motion-state dimensions using compressed sensing. *Magn Reson Med* **75**, 775-788 (2016). https://doi.org/10.1002/mrm.25665

4 Yang, W. *et al.* Four-Dimensional Magnetic Resonance Imaging With 3-Dimensional Radial Sampling and Self-Gating-Based K-Space Sorting: Early Clinical Experience on Pancreatic Cancer Patients. *Int J Radiat Oncol Biol Phys* **93**, 1136-1143 (2015). https://doi.org/10.1016/j.ijrobp.2015.08.028

5 Winkelmann, S., Schaeffter, T., Koehler, T., Eggers, H. & Doessel, O. An optimal radial profile order based on the Golden Ratio for time-resolved MRI. *IEEE Trans Med Imaging* **26**, 68-76 (2007). https://doi.org/10.1109/TMI.2006.885337

6 Lebrun, M. An Analysis and Implementation of the BM3D Image Denoising Method. *Image Processing On Line* **2**, 175-213 (2012). https://doi.org/10.5201/ipol.2012.l-bm3d

7 Mittal, A., Soundararajan, R. & Bovik, A. C. Making a "Completely Blind" Image Quality Analyzer. *IEEE Signal Processing Letters* **20**, 209-212 (2013). https://doi.org/10.1109/LSP.2012.2227726

8 Isensee, F., Jaeger, P. F., Kohl, S. A. A., Petersen, J. & Maier-Hein, K. H. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods* **18**, 203-211 (2021). https://doi.org/10.1038/s41592-020-01008-z

9 He, L. *et al.* The connected-component labeling problem: A review of state-of-the-art algorithms. *Pattern Recognition* **70**, 25-43 (2017). https://doi.org/10.1016/j.patcog.2017.04.018

10 Thomas HR, *et al.* Contrast-enhanced 4D-MRI for internal target volume generation in treatment planning for liver tumors. *Radiotherapy and Oncology*. **173**, 69–76. (2022) https://doi.org/10.1016/j.radonc.2022.05.037

Reconstruction of Time-Activity Curves in MRT Therapy Using Bayesian Unfolding

F. Nicolanti[1,2], L. Campana[3], F. Collamati[1], R. Faccini[1,2], R. Mirabelli[1,2], S. Morganti[1], E. Solfaroli-Camillocci[4], and C. Mancini-Terracciano[1,2]

[1]National Institute of Nuclear Physics, INFN Rome, Italy.
[2]Department of Physics, Sapienza University of Rome, Italy.
[3]Department of Basic and Applied Sciences for Engineering, Sapienza University of Rome, Italy.
[4]National Center for Radiation Protection and Computational Physics, Italian National Institute of Health, Italy.

Abstract Molecular Radiotherapy (MRT) holds the potential for personalized cancer treatment, yet achieving precise dosimetry remains a challenge. WIDMApp (Wearable Individual Dose Monitoring Apparatus) is a novel approach to address this by developing a method to accurately characterize the Time-Activity Curves (TACs) of organs using Time-Counts Curves (TCCs) acquired with wearable radiation detectors designed on purpose. The WIDMApp approach relies on the detectors, a Monte Carlo simulation, and an unfolding procedure. The MC simulation estimates the probabilities that the particles produced in each organ produce a signal in all the detectors. Such probabilities are then used to infer the organ activities (i.e. the TACs) given the detectors measurements (i.e. the TCCs) using a Bayesian approach. Thanks to the Bayesian formalism, this algorithm does not need to assume the TACs functional form. Utilizing the MIRD computational human phantom, we developed a Monte Carlo simulations to generate the dataset of an ^{131}I thyroid treatment. Assuming a mono-exponential washout of all the relevant organs, we simulated the measurement of TCCs and we used this dataset to test our algorithm. We assess the Bayesian algorithm's performances under varying initial conditions and data fluctuations, and compare results with those obtained by using a fitting procedure. This study indicates that the Bayesian unfolding can reconstruct organ TACs within 15% accuracy, even with significant noise and uncertainty in initial activity estimations. The obtained results pave the way for the exploration of more complex and realistic clinical scenarios.

1 Introduction

Enhancing personalization in Molecular Radiotherapy (MRT) offers significant benefits for improving treatment outcomes and minimizing toxicity. Achieving accurate dosimetry at the individual organ level is essential for understanding the unique biodistribution of radiopharmaceuticals in patients, allowing for precise calculation of absorbed doses to both normal organs and targeted tissues. Traditional dosimetric methods, which fit a limited number of Time Activity Curves (TACs) points from SPECT measurements to predefined models, lead to substantial uncertainties on absorbed dose (order of 30–100%) and poorly represent individual patient pharmacokinetics [1].

In this framework, WiDMApp (Wearable Individual Dose Monitoring Apparatus) aims at providing a precise characterization of the TACs of any organ of interest starting from Time-Counts Curves (TCCs) acquired with a limited number of radiation detectors worn by the patient during the treatment period. A deconvolution method is then used to deconvolve the detected signals and infer the organ-specific radiopharmaceutical kinetics. In a previous work [2], we demonstrated the feasibility of the WIDMApp's approach by employing a fitting process for the cumulative chi square function as deconvolution method. While the methodology exhibited significantly enhanced reconstruction accuracy compared to current state-of-the-art methods, it was constrained by the assumptions on the functional form of the radiopharmaceuticals biokinetics. This limitation prompted the exploration of a more flexible, model-independent approach through Bayesian unfolding, eliminating the need for predefined TAC's functional forms.

In the first section, we will introduce the case being examined, and provide a concise overview of the Monte Carlo simulation and the creation of the data set. In Section 2.2, we will describe the Bayesian deconvolution method. Section 2.3 will detail the stability studies conducted to evaluate the algorithm's performance. Subsequent sections will then present and discuss the findings of this study, by comparing them with he reconstruction accuracy obtained by the fit procedure used in our previous approach.

2 Materials and Methods

To compare the performance of the Bayesian algorithm with the fitting approach already published in [2], we analyzed the same dataset. Below we provide a brief summary of how this dataset was constructed.

2.1 Monte Carlo simulation and TCC generation

We used the Geant4 Monte Carlo toolkit [3, 4] to simulate a simplified ^{131}I thyroid treatment, employing the MIRD phantom for patient anatomy [5]. We focused on five organs as potential radiation sources: liver, kidneys, spleen, bladder, and thyroid, each uniformly emitting 364 keV photons resulting from the de-excitation of ^{131}Xe*, which is the final state of ^{131}I β^- decay. To detect radiation, seven sensors were strategically placed on the phantom's body: two on the thyroid and five on the other organs, all assuming 100% detection efficiency. The MC simulation is used to assess the

Organ	A_0 [MBq]	$t^{bio}_{1/2}$ [h]	$t^{eff}_{1/2}$ [h]
Thyroid	750.0	1920.0	174.9
Right Kidney	115.0	12.0	11.3
Left Kidney	105.0	10.0	9.5
Bladder	35.0	7.0	6.8
Liver	1115.0	24.0	21.3
Spleen	130.0	17.0	15.6

Figure 1: Initial activities A_0, and biological t^{bio} and effective t^{eff} decay times assigned to each organ.

probability of generating a signal in each detector, yielding a 6×7 probability matrix.
Realistic initial activities and effective half-lives for each organ [6] (reported in table 1) were used to generate Time Counts Curves over ten days, based on a mono-exponential decay model. Using the Monte Carlo output, and assuming the TACs parameters in table 1, the time-dependent signal generated by each detector, was constructed as:

$$S_i(t) = \sum_j p_{ij} A_{0j} \exp\left(-\frac{t}{t^{eff}_{1/2\,j}} \ln 2\right) \quad (1)$$

where p_{ij} are the probability matrix elements, A_{0j} the activity of the j organ at $t = 0$, and $t^{eff}_{1/2\,j}$ the effective half-life of the radiopharmaceutical in that organ.
Experimental TCCs were generated over a period of 240 hours post-administration, with a higher sampling rate for the initial 48 hours and a lower one for the remaining 8 days. This approach was strategically implemented in our previous work [2], motivated by the usage of a cumulative chi-square for fitting purposes. The variation in sampling rates was specifically designed to give equal statistical weight to data over the entire data-taking period. Contrary to our previous method, the Bayesian unfolding is done at each time step independently and so it's independent from the sampling rate. In principle, this dataset originally optimized for our previous algorithm, is not required for Bayesian unfolding. Nevertheless, to facilitate direct comparison, we are using the same dataset for evaluating both algorithms. We applied Poisson fluctuation to TCCs from eq. 1 to simulate real measurement variance. Generated signals for each detector are shown in figure 2.

2.2 Temporal Bayesian unfolding

To unfold the distribution over time from a number of counts measurements to an organ-activity distribution we use the Bayesian approach proposed in [7], which makes use of the concept of causes and effects. We define each organ's radiopharmaceutical activity as a cause C_i, and the counts detected by each sensor as effects E_j.
Using this formalism, the Eq. 1 could be written as:

$$n(E_j) = \sum_{i=1}^{n_C} n(C_i) P(E_j|C_i) \quad (2)$$

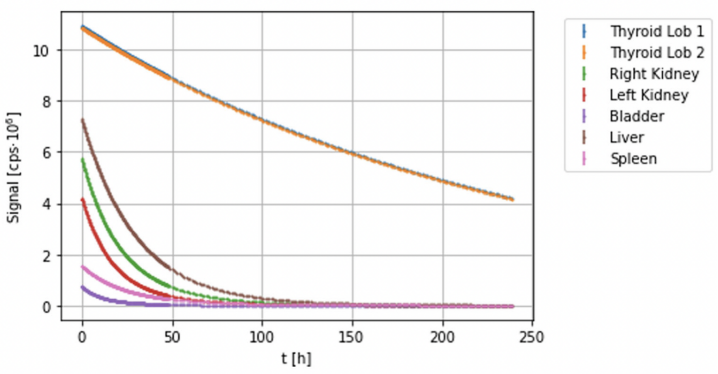


Figure 2: The simulated TCCs, one for each of the seven detectors. This dataset is the same of one of our previous works [2] to allow an easy comparison of this approach with the one already published.

which states that the number of counts in the jth detector $n(E_j)$ is a combination of the activities $n(C_i)$ of all the n_C possible sources (or causes) each of them multiplied with the probability $P(E_j|C_i)$ that the specific source (or cause) produces a signal in the considered detector.
Now the goal is to estimate the source:

$$n(C_i) = \frac{1}{\varepsilon_i} \sum_{j=1}^{n_E} n(E_j) P(C_i|E_j) \quad (3)$$

where n_E is the number of detectors (or effects), ε_i the efficiency of the $i-th$ cause to have an effect, defined as $\varepsilon_i = \sum_{j=1}^{n_E} P(E_j|C_i)$. Lastly, $P(C_i|E_j)$ is the probability that a signal in the j-th detector is due to the i-th source, and can be calculated using the Bayes' theorem:

$$P(C_i|E_j) = \frac{P(E_j|C_i)P(C_i)}{\sum_{l=1}^{n_c} P(E_j|C_l)P(C_l)} \quad (4)$$

assuming that we know the conditional probability $P(E_j|C_i)$, and the initial *a priori* probability of the causes $P(C_i)$.
Combining the Eq.s 3 and 4, we obtain the unfolding algorithm:

$$n_{t+dt}(C_i) = \frac{1}{\varepsilon_i} \sum_{j=1}^{n_E} n_t(E_j) \frac{P(E_j|C_i)P_t(C_i)}{\sum_{l=1}^{n_c} P(E_j|C_l)P_t(C_l)} \quad (5)$$

where we explicitate that the i-th cause at time $t + dt$, $n_{t+dt}(C_i)$, is estimated from its value at time t, $n_t(C_i)$. From the unfolded sources in eq. 5, we can estimate the posterior probabilities of the causes as:

$$P_{t+dt}(C_i) = \frac{n_{t+dt}(C_i)}{\sum_{l=1}^{n_c} n_{t+dt}(C_l)} \quad (6)$$

to be used as the *a priori* probability for the subsequent temporal step. Referring to our case under study, we observe that:

- $n_t(E_j)$ is the number of counts $S_j(t)$ on each detector j at a given fixed time t;
- $n_t(C_i)$ is the activity $A_i(t)$ in organ i at the same time t;

- $P(E_j|C_l)$ is the probability that a particle generated within the i-th organ produces, directly or indirectly, a signal in the j-th detector, called p_{ij} in our previous work. We estimate it with Monte Carlo simulations.

- $P_t(C_i)$ is the *a priori* probability at time t that a particle is produced in the i-th organ. It corresponds to the i-th organ activity, $n_t(C_i)$, normalised by the activity of all the body, i.e.: $\sum_{l=1}^{n_c} n_t(C_l)$.

The Bayesian unfolding, described in Eq.s 5 and 6, needs as a first step an estimation of the organs activities, $n_{t=0}(C_i)$, which could be achieved performing a SPECT scan at the beginning of the treatment. Subsequently, it iteratively estimates the activity of all organs through a sequential process over time.

2.3 Stability of the deconvolution algorithm

To evaluate the Bayesian unfolding's robustness and facilitate a direct comparison with the fitting methodology, we conducted stability studies similar to those performed in our previous work. Specifically, we evaluated the TACs reconstruction performances as a function of:

- the initial values A_{0j} used as starting point of the deconvolution algorithm (hereinafter referred to as *priors*), to test the impact of uncertain values derived from SPECT scan;

- the amplitude of random uniform fluctuation on signals, to evaluate the impact of external noise sources, such as detector mispositioning.

For the former, we executed the unfolding process 100 times, each with priors at $t = 0$ randomly chosen from a uniform distribution centered around their true values, varying the distribution's half-width from 10% to 90%.

The unfolding algorithm's resilience to noise was tested by dispersing data points with uniform noise ranging from 5% to 40%. We repeated the unfolding process 100 times for each level of noise, starting each unfolding with priors uniformly distributed around the true values, with a fixed half-width of 50%.

For each stability study, the TACs obtained from the 100 minimizations (100 for each distribution amplitude), were averaged and their variance calculated. The cumulative activity A^{tot} for each organ, has then been determined by integrating the TACs over the measurement campaign's duration with the WIDMApp sensors, i.e. from time $t = 0$ of radiopharmaceutical administration, to time $t_f = 240h$.

The obtained results regarding the cumulated activity are presented in the subsequent section.

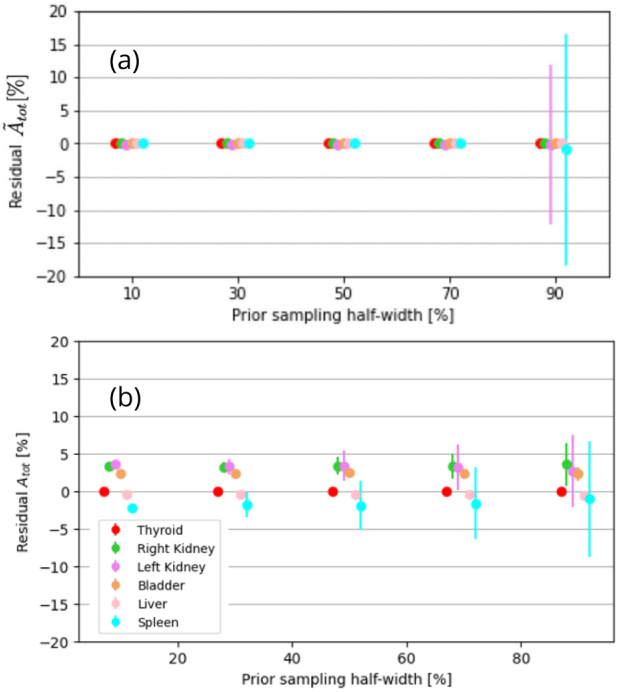


Figure 3: Percentage residuals between the organs cumulated activities A^{tot} resulted from the unfolding procedure and their true values obtained with the parameters reported in table 1, as a function of the width of the uniform distributions from which the priors are sampled. In plot (a), there are results obtained with the fit procedure, in plot (b) those obtained with Bayesian unfolding.

3 Results

3.1 Stability of cumulated activity estimate with respect of priors

Figure 3(b) illustrates the algorithm's stability in estimating the cumulated organ activity in the presence of only poissonian noise, showing how this accuracy varies with the uncertainty of the uncertainty on priors A_{0i}. Specifically, the percentage residuals between the cumulative activity A^{tot} estimated by the unfolding process and the true values calculated using parameters from table 1, are presented as a function of the half-width of the distribution from which the priors were sampled. These results are compared with the results obtained with the fit procedure in 3(a).

We can see that in the case of Bayesian unfolding, the estimated activities exhibit a slight bias (order of a few percent) even when starting with priors close to the true values. Nevertheless, these biases remain almost constant up to 80% higher/lower than the truth, with uncertainty always remaining below 15%.

3.2 Stability of cumulated activity estimate with respect of data dispersing

In Figure 4 we report the combined effect of the priors choice and data fluctuations on cumulated activity. The percentage residuals A_j^{tot} between the estimated and actual cumulated

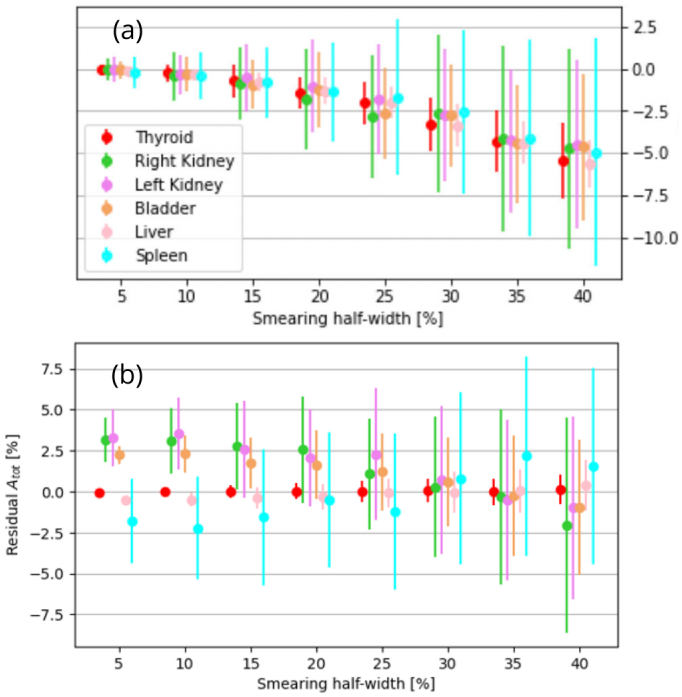


Figure 4: Same as fig. 3 but as a function of the data dispersions considered. In plot (a), there are results obtained with the fit procedure, in plot (b) those obtained with Bayesian unfolding

activity are plotted as a function of the percentage dispersion on data, with starting priors extracted from a uniform distribution with half-width of 50%. Also in this case, the results obtained by the fit approach 4 (a) are compared with the actual results 4 (b).

In the case of the fit procedure, we can see that the A_j^{tot} values are systematically underestimated, with an increasing trend as the half-width of the distribution increases, introducing an offset in the determination of the curves' integrals. The results obtained with Bayesian unfolding exhibit biases with respect to true cumulated activity similar to those obtained in the priors-dependent study, but still less than 2.5%. Notably, the offset found in the cumulated activities for kidneys, spleen, and bladder, are due to the significantly larger activity magnitudes in the thyroid and liver for the studied case. Minor deviations in these two organs' activity estimates cause larger variations in the others, while maintaining the cumulative reduced chi-square of unfolded curves near 1. Overall, the fit procedure yields better results with low smearing rates, whereas Bayesian unfolding outperforms with higher data fluctuations, achieving accuracy within 15%, up to 40% uniform noise in measured signals.

4 Discussion

Application of the Bayesian algorithm in the case of thyroid treatment with ^{131}I demonstrates performances quite comparable to the previous fitting approach. However, the results are considered satisfactory due to the Bayesian algorithm's two main strengths: its capacity for instant-by-instant deconvolution without assuming a predefined functional form for TACs, and its independence from the sampling rate. The only dependency on the latter lies in the assumption that the posterior at time t becomes the prior at time $t + dt$. If the dt interval increases substantially, the reconstruction accuracy may decrease. Despite this, the high sampling rate allowed by WIDMApp's wearable detectors effectively avoid this limitation.

Additionally, the results could be further improved by iterate more time the Bayesian unfolding per each time, using the previous estimation at the same time as prior.

Although in this simple case described by mono-exponential TACs the results do not clearly outperform the fitting approach, the flexibility of the Bayesian algorithm allows for application to more complex scenarios where the functional form cannot be assumed a priori, and traditional methods may fail.

5 Conclusion

This work introduces a novel Bayesian unfolding approach for reconstructing organ-specific time-activity curves (TACs), significantly enhancing personalized dosimetry in Molecular Radiotherapy (MRT). Our findings demonstrate the capability to reconstruct organ TACs with accuracy within 15%, even under conditions of up to 40% uniform noise in measured data and with initial activity uncertainties of 50%. This new methodology surpasses the previous fit technique in that no assumption is made regarding the functional form on the biokinetics of the drug. Subsequent studies will be carried out to test the algorithm's effectiveness in more complex scenarios, which include functional forms consisting of initial uptake and double-exponential decay, and by using the ICRP phantom

References

[1] E. Ao et al. "Improved dosimetry for targeted radionuclide therapy using nonrigid registration on sequential SPECT images." *Med Phys.* 42 (2015), pp. 1060–1070.

[2] S. Morganti et al. "Technical note: A wearable radiation measurement system for collection of patient-specific time-activity data in radiopharmaceutical therapy: system design and Monte Carlo simulation results". *Medical Physics* 48.12 (2021), pp. 8117–8126. DOI: 10.1002/mp.15311.

[3] S. Agostinelli et al. "Geant4 a simulation toolkit." *NIM-A* (2003), 506:4.

[4] J. Allison et al. "Recent developments in Geant4." *NIM-A* (2016), 835:186–225.

[5] W. Snyder et al. "MIRD Pamphlet No. 5 Revised, Estimates of absorbed fractions for monoenergetic photon sources uniformly distributed in various organs of a heterogeneous phantom". *J. Nucl. Med. Suppl.* (1969), pp. 5–52.

[6] I. C. on Radiological Protection. "Age Dependent Doses to Members of the Public from Intakes of Radionuclides: Part 1." *Pergamon Press* (1989).

[7] G. D'Agostini. *Improved iterative Bayesian unfolding*. 2010.

Rapid Auto-Planning for Cervical Brachytherapy: Direct Dwell Time vs. Dose Prediction with Deep Learning Networks

Lance Moore[1], Karoline Kallis[1], Kelly Kisling[1], Nuno Vasconcelos[2] and Sandra Meyers[1]

[1]Department of Radiation Medicine and Applied Sciences, University of California, San Diego, La Jolla, USA
[2]Department of Electrical and Computer Engineering, University of California, San Diego, La Jolla, USA

Abstract Treatment planning for cervical brachytherapy is often performed manually and can take over an hour. Automated planning leveraging deep learning models could significantly reduce planning time, improving the patient experience. The goal of this work was to assess a new approach to automated planning: using a neural network to directly predict dwell times (VGG model). This model was compared to an automated planning method that uses a neural network to predict 3D dose (OPT model), followed by an optimizer to convert dose to dwell times. Models were developed using 1803 prior cervical brachytherapy treatment plans, with a 68%/16%/16% train/test/validation split. The dataset included 7 different types of implants, featuring some combination of tandem, ovoids, ring, and/or needles. The VGG model used a 16-layer 3D VGG regression network, while the OPT model employed a Cascade U-Net. Automated plans were compared to clinical plans using the mean absolute error (MAE) and mean error (ME) in dose and dwell times. The resulting automated plans were very similar to clinical plans for both models, as evidenced by low MAE (OPT=3.4%, VGG=4.7%) and ME in dose (OPT=0.1%, VGG=0.6%), and low MAE (OPT=7.3s, VGG=8.0s) and ME (OPT=0.2s, VGG=0.7s) in dwell times. While the OPT model featured slightly superior performance, the direct dwell time prediction of the VGG model reduced computation time. Both methods of automated planning could provide high clinical value for brachytherapy procedures.

1 Introduction

Brachytherapy is an important component of the standard of care for cervical cancer. Treatments are customized to patient anatomy on imaging, but this treatment planning is often performed manually and can take over an hour [1]. Automated treatment planning could significantly reduce treatment time, decrease the use of hospital resources, standardize treatment, and improve patients' experience.

Automated, knowledge-based planning uses models trained on prior patient data to generate an optimal treatment plan for a new patient's anatomy. Three solutions have been proposed for knowledge-based brachytherapy planning of cervical cancer [2–4]. Two groups have trained deep reinforcement learning models to autonomously adjust parameters such as inverse optimization objectives and weights [3] or dwell times [2]. These methods rely on objective functions that incorporate dose metrics for target and organ contours. However, cervical brachytherapy often delivers prescription dose to additional regions that are not contoured, such as the uterus and parametrium, and therefore this dosing could not be included in such an objective function.

An alternative is automated planning based on 3D dose prediction from a 3D U-Net [4,5]. One group has shown that the 3D dose can be predicted using 3D inputs of standard anatomical contours and applicator geometry for tandem-and-ovoids (TO) [5]. An optimizer can then iteratively determine the dwell times for each dwell position that replicate this 3D predicted dose, generating an automated treatment plan in a few minutes [4]. These U-Nets can learn common dose shapes, avoiding the need for additional target contours beyond the high-risk CTV (HRCTV).

The above method is limited by the accuracy of the CNN, which may predict physically impossible dose distributions. Cortes at al reported reduced accuracy in high dose regions near applicators [5]. Further, it requires a 3D predicted dose intermediate and a separate optimizer, which increases computation time. The purpose of this work was to develop a deep learning model for direct dwell time prediction. Automated plans produced using these two methods were compared to clinical plans to assess performance.

2 Materials and Methods

Deep learning models were developed using 1803 prior high dose-rate brachytherapy treatment plans for cervical cancer for 475 patients (from 2009-2023). The train/test/validation split was 68%/16%/16%. The approximate applicator breakdown in each dataset was 40% TO, 6% TO with 4-10 needles (TON), 14% tandem-and-ring (TR), 16% TR with 3-6 needles (TRN), 3% tandem only (T), 18% T and 3-21 needles (TN) and 2% pure interstitial with 3-18 needles (N). Two automated planning techniques based on deep learning models are described below.

Model inputs

Planning CT images and DICOM-RT files were exported from the treatment planning system and used to generate model inputs. Binary contour masks of HRCTV, vaginal HRCTV (if present), bladder, rectum, and sigmoid were extracted from DICOM structure sets. Applicator masks were produced by extracting dwell position coordinates from the plan DICOM file and setting a 3x3x3 voxel box centered on each dwell position to 1. Glowing masks, which encode the distance of each image voxel to a structure with an inverse square fall-off, were also produced. Voxels inside a structure were set to 1, then for each image voxel, the summed total inverse squared distance from that voxel to every structure voxel was calculated. The intensity of voxels outside the structure were normalized by the maximum value outside, and voxels inside the structure were set to 1. Glowing masks were produced for anatomical contours and the applicator mask.

A dose-rate kernel was produced by exporting the 3D dose matrix for a single dwell position from the treatment planning system and dividing by the dwell time. This kernel was mapped to each dwell position in each applicator to produce 3D single-dwell position dose-rate kernels, which were used as an input for one model. Four applicator dose kernel inputs were produced by summing

dose-rate kernels for all dwell positions in a given applicator type (tandem, ring, ovoids and needles) with uniform dwell time weighting.

Dose prediction and optimizer model

This model (termed "OPT") used a 3D Cascade U-Net [6] with a mean square error (MSE) loss function to predict the 3D dose. The inputs included 5 binary contour masks, a binary applicator mask, and 4 applicator dose kernels. The model consisted of 2 in-series U-Nets, each containing 5 encoding layers and 5 decoding layers (Figure 1A). Each encoding layer contained 2 in-series depth-wise separable convolution blocks. Each block contained a 3D grouped convolution followed by an instance normalization, ReLU activation, a 3D point-wise convolution, a second instance normalization, and a second ReLU activation function. A 3D maxpool operation performed at the end of each encoder block served to reduce the feature map dimensions. In the decoder layers, the convolution blocks were preceded by a transposed convolution to increase the dimensions of the feature maps prior to cross concatenation with the feature maps from the corresponding encoder layer. At the termination of each U-Net, two additional convolution blocks were appended to reduce the number of feature maps and generate the final 3D dose prediction. To employ the cascade mechanism for the network, the penultimate feature maps from the first U-Net were concatenated with the inputs to the second U-Net prior to feeding the images to the model.

Model dose predictions were then converted into deliverable treatment plans using a COBYLA optimizer, which minimized the MSE between the 3D dose calculated for a given set of dwell times and the predicted dose [4].

VGG model

This model (termed "VGG") used a 16-layer 3D VGG [7] regression network to directly predict dwell times. Model inputs included the planning CT image, glowing contour and applicator masks, and single-dwell position dose-rate kernels. Due to memory limitations, a maximum of 200 dwell position kernels could be included, which meant that 4 plans were excluded. If a treatment plan lacked a contour or 1 of the 200 dwell positions, that channel was left blank.

The model followed a standard VGG archetype, with sequential convolutions in the encoder portion of the network and a multi-layer perception (MLP) decoder (Figure 1B). In the encoder, each convolution block consisted of a standard 3D convolution followed by a 3D batch norm and a ReLU activation. A 3D max pool was used after the second, fourth, seventh, tenth, and thirteenth convolution blocks to reduce the feature maps prior to flattening. Before the MLP, a length 200 dwell position token vector was concatenated to the flattened feature vector. This vector contained a one-hot encoding of the number of dwell positions in the plan. This vector injects plan information deep into the network to assist the model in predicting dwell times only for valid dwell positions. The MLP in the decoder consisted of three layers, each followed by a ReLU activation function and a dropout layer with probability 0.1. The decoder produced a vector of size 200, i.e. dwell times for the 200 possible dwell positions. The absolute value of the predictions ensured that no negative dwell times were predicted.

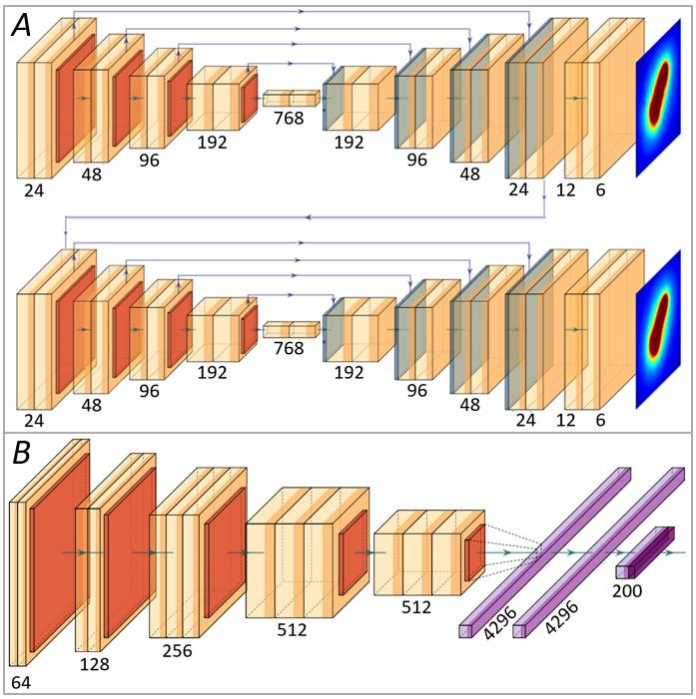


Figure 1. Model diagrams for (A) Cascade U-Net and (B) VGG network. Yellow blocks indicate a convolution block, orange indicates a maxpool, blue blocks indicate a transposed convolution, and purple indicates a fully connected layer. The numbers below the blocks depict the number of filters or channels.

Each predicted dwell time was multiplied by the corresponding dwell position dose-rate kernel and then the 3D predicted dose was computed by summing over all dwell positions. The loss function for this model compared the 3D predicted dose to the 3D clinical plan dose. A 3D squared error matrix (SE) was computed between the predicted ($Dose_{pred}$) and clinical ($Dose_{clin}$) distributions using equation (1) below.

$$SE = (Dose_{clin} - Dose_{pred})^2 \qquad (1)$$

The final MSE loss function was computed from the voxel-wise mean of the SE matrix, weighted by each of the 5 glowing contour masks ($Glow_Mask_i$), per equation (2).

$$MSE = mean(SE + \sum_{i=1}^{5} SE \cdot Glow_Mask_i) \qquad (2)$$

Training methods and hyperparameters

Data augmentation was employed during model training to delay overfitting. Both networks used an AdamW optimizer, a weight decay of 1e-4, and a one-cycle learning rate scheduler to anneal the learning rate over the course of training (peak learning rate 1e-3). Models were implemented using the PyTorch framework and trained for 500 epochs on a single GPU node with four 32GB NVIDIA V100 SMX2 GPUs. The validation set performance was

checked at the end of each epoch and the best performing checkpoint was selected as the final model.

Comparing model performance

Automated plans were compared between the models using the following metrics: the voxel-based mean error (ME, representing bias) between the 3D auto-plan and clinical dose, the voxel-based mean absolute error (MAE) in dose, and the ME and MAE between auto-plan and clinical plan dwell times. The ME in clinical metrics (HRCTV D90 and D2cc of organs) were also compared. A positive ME indicates the clinical dose or dwell times were higher than the auto-plan.

3 Results

Figure 2 shows example dose distributions and dwell times for the two automated plans compared to the clinical plan for three patients. While there were some differences in dwell times, the general trends in the dwell time distributions were similar, and the resulting dose distributions were nearly indistinguishable.

Box plots of MAE in dose over all voxels and dwell times, broken down by model and applicator type, are displayed in Figure 3. Averaged over all plans, the MAE and ME in dose were 23.2 cGy (3.4%) and 0.8 cGy (0.1%) for the OPT plans, and 31.6 cGy (4.7%) and 4 cGy (0.6%) for the VGG plans. MAE and ME in dwell times were 7.3 s (1.4% of the total plan dwell time) and 0.2 s (0.0%) for the OPT plans, and 8 s (1.6%) and 0.7 s (0.1%) for VGG plans.

Figure 4 shows boxplots of ME in clinical metrics for each model, all plans combined. For the OPT model, rectum D_{2cc} showed the lowest ME of -0.2%, while HRCTV D90 had the largest ME of 2.8%. The VGG model showed the best performance for bladder, with a ME of -0.7%, and the largest ME for HRCTV D90 (6.7%).

4 Discussion

In this work, we demonstrate the value of two unique, deep learning-based methods for automated planning of cervical brachytherapy. These models were trained and validated on a dataset of over 1800 treatment plans with 7 different applicator implant types. To our knowledge, this is one of the largest datasets that has been used for dose prediction or automated planning with deep learning in radiotherapy, and the greatest variety of brachytherapy implants that have been included in any single brachytherapy deep learning model thus far. Both methods produced treatment plans that were similar to clinical plans, as evidenced by the near-zero mean errors and low mean absolute errors in dose and dwell times. We believe that these facts qualify this work as a substantial and significant step forward in the field of brachytherapy. Despite the variability in the data due to the numerous applicator types included, performance was quite similar between applicator types, demonstrating that a single model was adequate.

The performance of the OPT model was slightly superior for most metrics. The tradeoff for these superior results is the time required for the optimizer, which took 30 s on average and up to 3.2 minutes (for plans with the most dwell times) on a computer with an Intel(R) Core(TM) i7-8700 3.2GHz processor. In contrast, the VGG model can directly predict dwell times in a matter of seconds, which is very valuable for real-time planning in the clinic. To our knowledge, this is the first deep learning model to directly predict dwell times for automated planning. Another study used a CNN to predict dwell positions and times for plan verification [10], but used the clinical 3D dose as input.

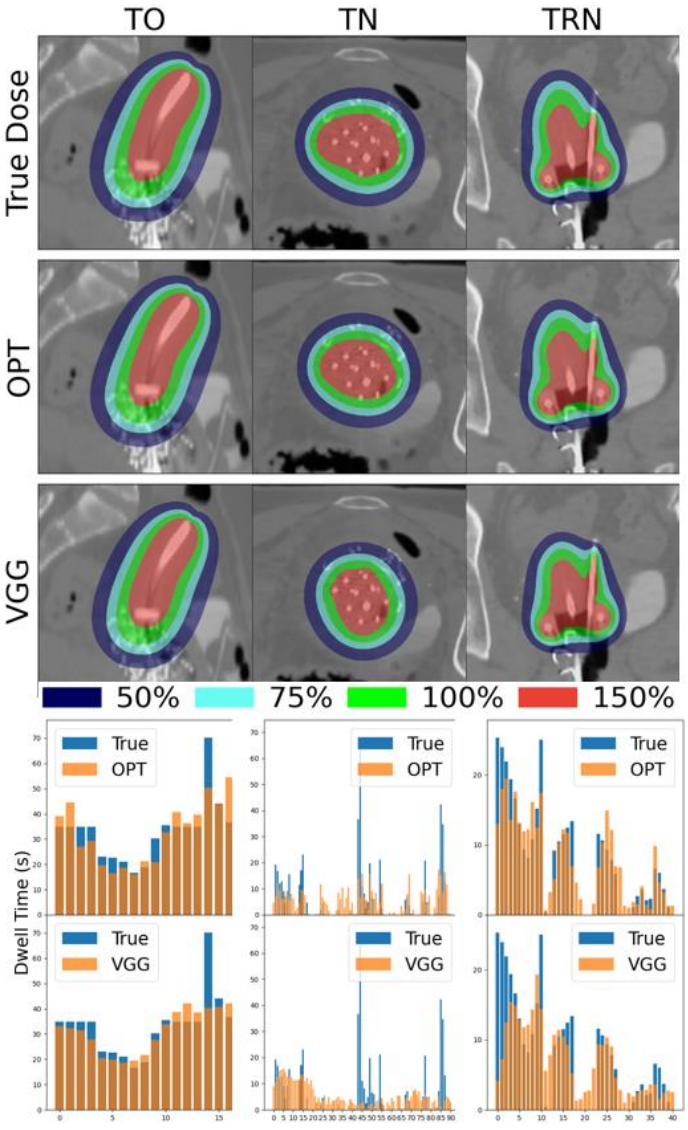


Figure 2. Comparison of dose distributions and dwell times for the two automated plans and clinical plan for three random patients. Isodose color washes (where 100% is the prescription dose) are overlaid over coronal (TRN), axial (TN), and sagittal (TO) CT images.

The VGG model directly incorporates a dose rate kernel exported from the treatment planning system. This kernel is used to compute a physically realistic dose from predicted dwell times for known dwell positions, which is compared to the clinical dose in the loss function. Thus, physical constraints for dose delivery are forced within the VGG model itself. In contrast, the OPT model was not

bound by physics and predicted dose distributions that featured larger errors at high dose. However, it appears that the optimizer was able to correct for most of those errors.

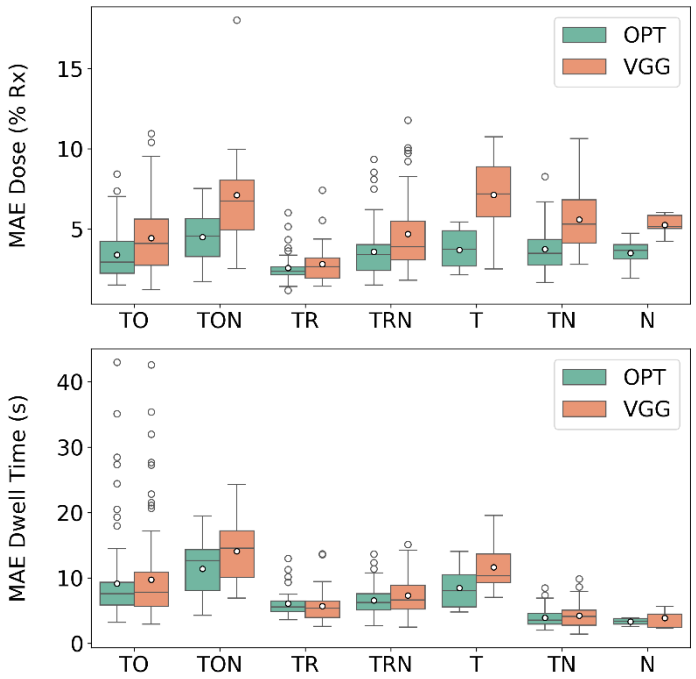


Figure 3. MAE in dose over all voxels (relative to prescription dose) and dwell times, comparing automated plans to clinical plans for each model and applicator. White circles overlaid on boxplots indicate mean values.

One limitation of the VGG method is that it can only support plans with less than 200 dwell positions. We expect that this model could improve further with a few modifications, which will be explored in future work. Currently, the model loads dwell position images into sequential channels, meaning the first dwell position channel is always loaded with the first dwell position in a plan. Shuffling these dwell positions could avoid over-emphasizing the first channels. We employed a dose-based loss function to account for the fact that slight variations in dwell times could result in approximately the same dose. However, we plan to try incorporating dwell time error in the loss function in the future.

In future work, we will continue to refine these models in pursuit of yet more accurate results, and test more sophisticated models. While the OPT model utilized only binary contour and applicator inputs, we found VGG model performance improved when incorporating glowing masks. Future work will explore the use of glowing masks in the OPT model as well. Blinded physician evaluation will be used to evaluate clinical acceptability of automated plans.

5 Conclusion

Two deep learning models were trained to produce automated treatment plans for cervical brachytherapy. The best-performing model predicted the 3D dose as an intermediate step, followed by an optimizer to convert the dose to a deliverable treatment plan. The second model directly predicted dwell times in a matter of seconds, reducing computation time. Both methods produced automated plans that were comparable to clinical plans, and thus could provide great clinical value by significantly reducing planning time.

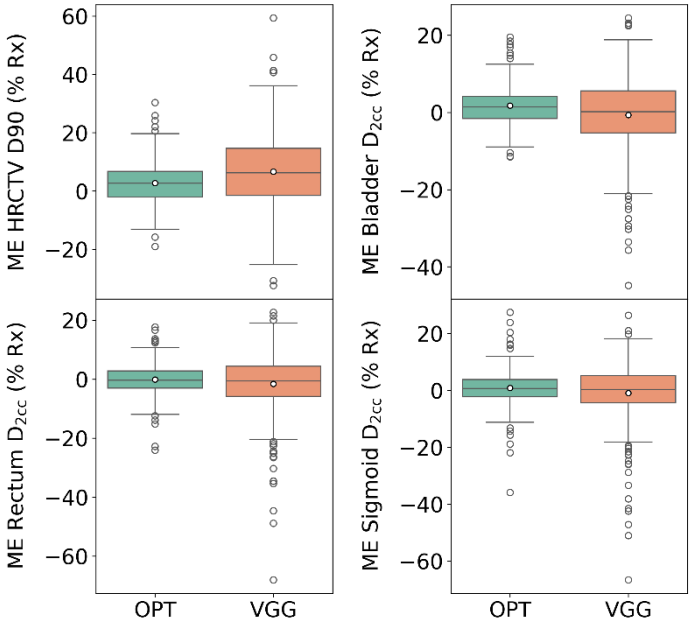


Figure 4. ME in clinical metrics for each model as a percent of the prescription dose, where a positive value indicates the clinical plan metric was higher than the automated plan. White circles overlaid on boxplots indicate mean values.

References

[1] J. Mayadev, L. Qi, S. Lentz, S. Benedict, J. Courquin, S. Dieterich, M. Mathai, R. Stern, R. Valicenti, Implant time and process efficiency for CT-guided high-dose-rate brachytherapy for cervical cancer, Brachytherapy 13 (2014) 233–239. https://doi.org/10.1016/j.brachy.2014.01.004.

[2] G. Pu, S. Jiang, Z. Yang, Y. Hu, Z. Liu, Deep reinforcement learning for treatment planning in high-dose-rate cervical brachytherapy, Physica Medica 94 (2022) 1–7. https://doi.org/10.1016/j.ejmp.2021.12.009.

[3] C. Shen, Y. Gonzalez, P. Klages, N. Qin, H. Jung, L. Chen, D. Nguyen, S.B. Jiang, X. Jia, Intelligent inverse treatment planning via deep reinforcement learning, a proof-of-principle study in high dose-rate brachytherapy for cervical cancer, Phys Med Biol 64 (2019) 115013. https://doi.org/10.1088/1361-6560/ab18bf.

[4] K. Kallis, L.C. Moore, K.G. Cortes, D. Brown, J. Mayadev, K.L. Moore, S.M. Meyers, Automated treatment planning framework for brachytherapy of cervical cancer using 3D dose predictions, Phys Med Biol 68 (2023). https://doi.org/10.1088/1361-6560/acc37c.

[5] K.G. Cortes, K. Kallis, A. Simon, J. Mayadev, S.M. Meyers, K.L. Moore, Knowledge-based three-dimensional dose prediction for tandem-and-ovoid brachytherapy, Brachytherapy 21 (2022) 532–542. https://doi.org/10.1016/j.brachy.2022.03.002.

[6] S. Liu, J. Zhang, T. Li, H. Yan, J. Liu, Technical Note: A cascade 3D U-Net for dose prediction in radiotherapy, Med Phys 48 (2021) 5574–5582. https://doi.org/10.1002/mp.15034.

[7] K. Simonyan, A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition, (2015). https://doi.org/10.48550/arXiv.1409.1556.

[8] J. Fan, L. Xing, Y. Yang, Independent verification of brachytherapy treatment plan by using deep learning inference modeling, Phys Med Biol 66 (2021). https://doi.org/10.1088/1361-6560/ac067f.

Near-to-target organs-at-risk segmentation in cervical high dose-rate brachytherapy via deep learning

Ruiyan Ni[1], Elizabeth Chuk[2,3], Kathy Han[2,3], Jennifer Croke[2,3], Anthony Fyles[2,3], Jelena Lukovic[2,3], Michael Milosevic[2,3], Benjamin Haibe-Kains[1,2,4,$], Alexandra Rink[1,2,3,$]

[1]Department of Medical Biophysics, University of Toronto, Toronto, CA
[2]Princess Margaret Cancer Center, University Health Network, Toronto, CA
[3]Department of Radiation Oncology, University of Toronto, Toronto, CA
[4]Vector Institute for Artificial Intelligence, Toronto, Toronto, CA
[$]Both authors should be considered as corresponding authors

Abstract Deep learning has been used to automate and speed up the manual contouring step for organs at risk (OARs) and targets in radiotherapy. In cervical cancer high dose-rate brachytherapy, OARs far from the radiotherapy target are not considered during treatment planning, which is driven by the OAR dose to 2 cubic centimeters ($D2cm^3$) closest to the target. Accuracy in distal OAR segmentation does not necessarily translate to dosimetric impact. To align with the clinical requirements, in this study, a distance-penalized loss function was developed to guide the network's attention towards the near-to-target OAR regions, ensuring highly accurate proximal OAR segmentations. Models trained with the proposed loss demonstrated improved geometric and dosimetric performance. A clinical acceptability test was conducted, by a physician scoring the auto-contours and recording the time to rate and revise. Additionally, a novel geometric metric, weighted Dice Similarity Coefficient (wDSC), was introduced as a better indicator for $D2cm^3$ accuracy compared to commonly used evaluation metrics (Pearson correlation coefficient, r=-0.55 for wDSC; |r|<0.22 for other metrics).

1 Introduction

Brachytherapy plays a vital role in cervical cancer treatment, providing improved survival rates over other advanced radiotherapy techniques such as intensity-modulated radiotherapy and stereotactic body radiation therapy[1] Delineating the radiotherapy targets and organs at risk (OARs) is the essential basis of obtaining treatment plans and is conventionally performed manually. The emergence of deep learning (DL) technologies demonstrated fast and accurate auto-segmentations on magnetic resonance (MR) images [2], [3].

The OARs treatment planning for brachytherapy is driven by the OAR dose to hottest 2 cubic centimeters ($D2cm^3$), which resides near the target[4]. The contouring accuracy must be high for the OAR regions proximate to the target, while discrepancies in distant regions are not expected to translate to dosimetric impact. Recent studies have shown that optimizing the precision of the regional-focused network is achievable through tuning the loss function in convolutional neural network (CNN) training, such as penalizing the loss function with distance maps[5].

Furthermore, it is essential to utilize appropriate geometric evaluation metrics as indicators of clinical acceptability. Surface Dice Similarity Coefficient (sDSC)[6] and added path length (APL)[7] are vital for assessing clinical acceptability in external beam radiotherapy auto-segmentation, highlighting the time saved and clinical acceptance. Yet, the correlation between these metrics and dosimetry remains understudied, especially in brachytherapy, where the doses received by the OARs are predominantly determined by their proximity to the target regions. This gap underscores the necessity for a new geometric evaluation metric tailored for brachytherapy dosimetry accuracy.

We address these issues by (1) incorporating a distance map to penalize the loss function for enhanced near-to-target OARs segmentation; (2) introducing a novel geometric metric, the weighted DSC (wDSC), which prioritizes accuracy in near-to-target regions, consistent with $D2cm^3$ considerations. Geometric evaluations, dosimetric analysis, and clinical acceptability test were conducted. We found that the CNN model, trained with our distance-penalized loss function, outperformed the baseline model in both geometric and dosimetric performance. Furthermore, wDSC demonstrated the highest correlation with dosimetric accuracy compared to the existing evaluation metrics.

2 Materials and Methods

2.1 Patient data

A dataset of 170 T2-weighted MR images with clinically used contours was built from 56 patients undergoing EBRT and high-dose-rate (HDR) brachytherapy from 2014 to 2019 at the Princess Margaret Cancer Center with Institutional Review Board (IRB) approval. The TNM stage of all the patients ranged from T1b to T4a. All the patients were first treated with EBRT with prescription doses of 45 Gy in 25 fx or 50.4 Gy in 28 fx and received HDR brachytherapy treatment as 7 Gy × 4fx or 8 Gy × 3fx. All patients were treated with interstitial ring and tandem applicators (Elekta) and scanned on a Siemens Verio 3T MR scanner. The median image size is 320×320×38 voxels, and the median image spacing is 0.625×0.625×3.0 mm^3. Clinically used contours, reviewed by experienced radiation oncologists (ROs), included the high-risk clinical tumor volume (HR-CTV), bladder, rectum, sigmoid, and small bowel. The dataset was split into subsets at the patient level: 45 patients (150 cases) as the training set for five-fold cross-validation and 11 patients (20 cases) as the testing set.

2.2 Distance-penalized loss function and training details

Distance-penalized (DP) maps, labelling which voxels were near the target/applicator, were generated to penalize the loss function for indicating the spatial relevance

information. Distance transformation was performed on a binary mask where the clinically used HR-CTV (Figure 1(a)) was labelled as 1, calculating the distance between each voxel to its nearest non-zero voxel (Figure 1(b)). Second, the inverse square law was applied to the distance-transformed map to produce the DP map (Figure 1(c)), which was normalized to a range of [0, 1]. Consequently, voxels nearer to the HR-CTV were allocated higher error penalties, reflecting their increased importance in treatment planning.

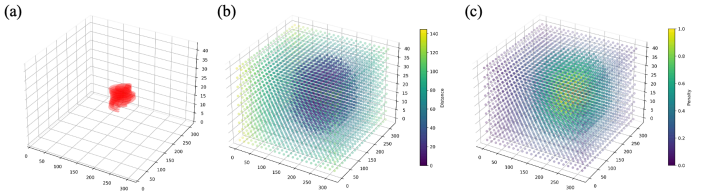


Figure 1. Examples of (a) HR-CTV, (b) Distance transformation based on HR-CTV, and (c) Normalized distance-penalized (DP) map. Points in (b) and (c) are downsampled with slice intervals of (15, 15, 5) along (x, y, z)

The DP map was used to penalize two commonly used loss functions, cross-entropy (CE) loss and the Dice loss part of the DiceCE loss[8]. The penalized loss functions, distance-penalized cross-entropy (DPCE) loss and distance-penalized DiceCE (DPDiceCE) loss, are defined by

$$L_{DPCE} = -\frac{1}{N}\sum_{c=1}^{C} D^c \odot \sum_{i=1}^{N} g_i^c \log p_i^c \qquad [1]$$

$$L_{DPDiceCE} = L_{DPDice} + L_{CE}, \qquad [2]$$
$$where\ L_{DPDice} = 1 - \frac{2\sum_{c=1}^{C} D^c \odot \sum_{i=1}^{N} g_i^c p_i^c}{\sum_{c=1}^{C}\sum_{i=1}^{N} g_i^c + \sum_{c=1}^{C}\sum_{i=1}^{N} p_i^c}$$

where g_i^c and p_i^c are the ground truth label and the network prediction of class c at voxel i. N and C are the number of voxels and classes, respectively. D^c is the distance penalty term of class c from DP map. $\odot$ is the Hadamard product. By incorporating the DP map into the loss function, the pixels that are closer to HR-CTV would be assigned larger weights.

The nnU-Net pipeline[8] was employed with the 3D full-resolution U-Net architecture. Four models were trained using two baseline, non-DP loss functions (CE loss and DiceCE loss) and two DP loss functions (DPCE loss and DPDiceCE loss). The patch size was 256×256×28. The models were trained for 200 epochs with a batch size of 2. The proposed framework was implemented in Python 3.9.7 using PyTorch 1.11.0. The training and testing were conducted using an NVIDIA Tesla V100 GPU with 32GB memory. The training and testing code is available on GitHub: https://github.com/bhklab/BrachyNewLoss

2.3 Geometric evaluation – New metric

We introduced a new evaluation metric, the weighted Dice Similarity Coefficient (wDSC), specifically designed to evaluate near-to-target OAR segmentation performance. Given that D2cm^3 measures the dose received by the 2cc of an OAR nearest to the radiotherapy target, the wDSC emphasizes the accuracy of proximal OAR regions to HR-CTV by incorporating a weighted factor in the original volumetric DSC (vDSC). The weights are derived from the voxel distances to HR-CTV, as represented in the DP map. The wDSC is defined by

$$wDSC = 1 - \frac{\sum_{i=1}^{N} \frac{i}{N} * (1 - vDSC_i)}{N} \qquad [3]$$

where N is the number of cropped volumes; $vDSC_i$ is the vDSC between the i-th cropped volumes of label and prediction; the weight of the i-th cropped volume is $\frac{i}{N}$. As shown in the illustration (Figure 2), the cropping occurs based on the DP map from center to the boundary. Each cropped volume has an associated weight, diminishing as it moves from center to boundary. Therefore, the discrepancies between the prediction and ground truth near the target have a higher impact on the wDSC value compared to those at the boundary. In this study, we used N=100 to crop each case into 100 cropped volumes. Each cropped volume was small enough that this approach can be considered continuous weighting. A Wilcoxon signed-rank test was utilized to analyze the statistical differences in the wDSCs between models trained with CE and DPCE, as well as between models trained with DiceCE and DPDiceCE loss functions, to determine the effectiveness of incorporating a distance penalty term in improving the delineation of proximal-target OARs.

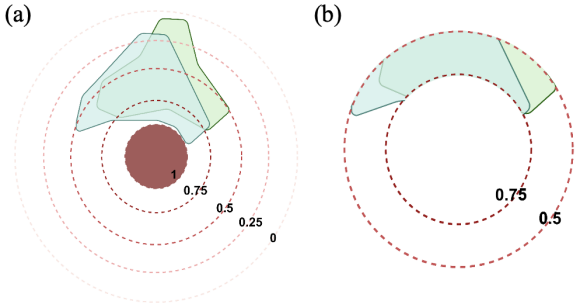


Figure 2. The illustration of wDSC between ground truth (lighter green) and network prediction (darker green). (a) Both are first cropped into N volumes based on the values in the distance map. In this example, N=4; the weights of each cropped volume = (1, 0.75, 0.5, 0.25). (b) A zoom-in version of volumes with weight = 0.75. Assume $vDSC$ = 0.8 between these two cropped volumes; the weighted $vDSC$ loss is 0.75*(1-0.8). The final $wDSC$ is $1 - \frac{total\ weighted\ vDSC\ loss}{N}$, iterating through all cropped volumes. From boundary to the center, the weights are $(\frac{1}{N}, \frac{2}{N}, \ldots, \frac{N-1}{N}, 1)$. In this study, we used N=100.

In addition, commonly used evaluation metrics were used to assess the whole organ segmentation performance, including vDSC, 95% Hausdorff Distance (HD95), surface DSC (sDSC)[6], and added path length (APL)[7].

2.4 Dosimetric evaluation

The dosimetric difference between manual contour and DL-generated contour was calculated to evaluate the dosimetric impact. The clinically used treatment plan was applied to the auto-segmentation to calculate $D2cm^3_{auto}$. The absolute D2cm^3 difference was quantified as $D2cm^3_{diff} = \text{abs}(D2cm^3_{auto} - D2cm^3_{clinical})$, indicating the D2cm^3 discrepancy between DL-based auto-contour and clinical contour. An absolute $D2cm^3_{diff}$ value closer to zero indicated that the auto-contour aligned better with the dosimetric requirements. The correlations between each geometric evaluation metric and D2cm^3 difference were evaluated to explore the best indicator of dosimetric accuracy. The linear regression was performed, employing the Pearson correlation coefficients to quantify the strength of the relationship. Wilcoxon signed-rank test was used to analyze the statistical differences between the clinical D2cm^3 values and the D2cm^3 values obtained from each loss function.

2.5 Clinical acceptability test

A radiation oncology fellow rated the DL contours on a five-point scale[9], where contours with a score of 5 (Use-as-is) or 4 (minor edits) were clinically acceptable and below 4 were not acceptable. The fellow would revise the auto-contour where it failed to meet clinical standards. The time taken for rating and revising was recorded.

3 Results

3.1 Correlation between geometric metric and D2cm^3 difference

The strongest correlation observed was between wDSC and D2cm^3 difference among all geometric evaluation metrics, which exhibited a Pearson correlation coefficient (r) of -0.55, indicating a moderate negative correlation. In contrast, the other four standard geometric evaluation metrics demonstrated weak or negligible correlations with the D2cm^3 difference with r of -0.15 for vDSC, 0.22 for HD95, -0.20 for sDSC, and 0.00 for APL.

3.2 Geometric evaluation

The newly proposed metric, wDSC, indicated that the model trained with the DP loss functions constantly yielded more accurate near-to-target OAR segmentation when compared to their respective non-DP counterparts. DPCE loss outperformed CE loss for the rectum, sigmoid, and small bowel, with mean wDSCs of 0.94 vs. 0.93, 0.89 vs. 0.88, and 0.86 vs. 0.85, respectively. Similarly, DPDiceCE had increased wDSCs, compared to DiceCE in rectum (0.94 vs. 0.93), sigmoid (0.90 vs. 0.89), and small bowel (0.87 vs. 0.85). There was a statistically significant improvement in small bowel with p-values of 0.008 for the CE group and 0.017 for the DiceCE group.

The DP group outperformed the non-DP group across four standard metrics. The most notable improvement was observed between DiceCE and DPDiceCE in small bowel with mean vDSCs of 0.74 and 0.77, mean HD95 of 23.95mm and 19.04mm, mean sDSCs of 0.52 and 0.55, and mean APL of 27.69cm and 27.01cm.

3.3 Dosimetric evaluation

The models trained with DP loss functions constantly demonstrated a reduced mean absolute $D2cm^3_{diff}$ than their non-DP counterparts (Table 1), indicating an enhanced accuracy in dosimetry. Furthermore, while each group had several outliers with large $D2cm^3_{diff}$, DP loss functions had smaller standard deviations, suggesting greater consistency in their performance. The only comparison that yielded a statistically significant difference was between CE and DPCE in the small bowel (p-value=0.012).

Table 1. Absolute $D2cm^3_{diff}$ between non-DP loss function and DP loss function (CE vs DPCE and DiceCE vs DPDiceCE) in four OARs. Smaller value indicates a more accurate D2cm^3. The values are represented as mean (±std). The two groups with statistically significant difference are underlined.

	CE	DPCE	DiceCE	DPDiceCE
Bladder	0.12 (±0.11)	0.10 (±0.10)	0.13 (±0.12)	0.11 (±0.10)
Rectum	0.16 (±0.21)	0.12 (±0.15)	0.12 (±0.14)	0.11 (±0.16)
Sigmoid	0.15 (±0.16)	0.13 (±0.11)	0.11 (±0.09)	0.10 (±0.09)
Small Bowel	<u>0.25 (±0.27)</u>	<u>0.18 (±0.21)</u>	0.22 (±0.26)	0.16 (±0.12)

3.4 Clinical acceptability

The clinical acceptability ratings (Figure 3) showed that DP group had marginally superior performance compared to the non-DP group. The model trained with DPCE loss had more clinically acceptable contours (with a score of 5 or 4) than CE loss in two out of four OARs (rectum and sigmoid). There were equal or higher number of contours rated as clinically acceptable using the DPDiceCE compared to DiceCE loss, except for sigmoid.

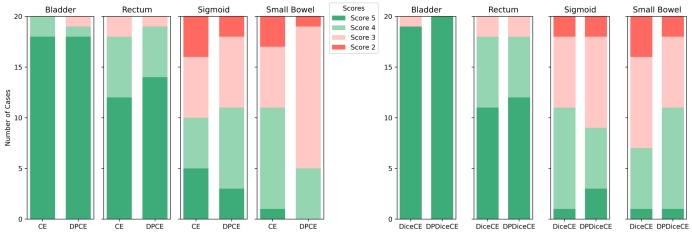


Figure 3. Clinical acceptability ratings between CE and DPCE (left) and between DiceCE and DPDiceCE (right) in

four OARs. Clinically acceptable scores (5 or 4) are in green shaded. Not acceptable scores (3 or 2) are in red shaded.

The average time required for reviewing an entire case was 0.73 minutes (ranging from 0.2 to 3.83 minutes); the average time to revise the contours in one case was 4.64 minutes (ranging from 0 to 19.12 minutes). For each loss function specially, the mean times for review and revision were similar for non-DP and DP loss functions. However, a notable difference was observed in the maximum time required for revision. The DP loss functions demonstrated a reduced maximum revision time (12.22 minutes for DPCE and 10.83 minutes for DPDiceCE), in comparison to the non-DP loss functions (19.12 minutes for CE and 15.27 minutes for DiceCE).

4 Discussion

In this study, wDSC, the new metric we proposed, was found to be the best indicator of D2cm^3 accuracy in image-guided brachytherapy compared to the four standard geometric measures. The models employing the DP loss functions showed improved performance in segmenting OAR regions close to the target, which led to more precise D2cm^3 calculations with reduced variations and a decrease in the time required to correct poorly segmented cases.

This study is the first to focus the CNN model's attention on near-to-target OAR regions, thereby enhancing dosimetric accuracy in brachytherapy. Compared to the two prior studies[2], [3], our model that incorporated the DP loss function achieved comparable or superior whole organ segmentation performance while exhibiting reduced D2cm^3 difference[3]. Furthermore, we included the clinical acceptability test, which revealed that most of the bladder and rectum contours, and approximately half of the sigmoid and small bowel contours, met clinical acceptability criteria. Although some contours failed to meet the clinical standard, their D2cm^3 differences remained within the range of inter-observer variations[10], [11].

Our study introduced and validated the wDSC as a new geometric evaluation metric, demonstrating its superior correlation with D2cm^3 difference in comparison to existing standard metrics. Our findings underscore a potential limitation in relying solely on standard metrics, which may not accurately represent dosimetric acceptance in brachytherapy applications despite being widely used in automatic medical image segmentation studies.

This work has several limitations. First, this was a single institutional study with a relatively small cohort size. More data from patients treated in our institution are being collected and will be used to reduce the model bias. we intend to validate the model's performance across multiple institutions in future studies. Second, wDSC was evaluated exclusively on cervical cases. We aim to expand the applicability of wDSC by validating it on prostate brachytherapy cases. Furthermore, the clinical acceptability test of the model involved only one fellow. More ROs will be involved to assess the inter-observer variations and contribute to a more robust evaluation.

5 Conclusion

We developed a DP loss function to guide the model for accurate near-to-target OAR segmentation in MR-guided cervical brachytherapy. The model with the proposed loss showed improved geometric and dosimetric performance. We introduced a new geometric metric, wDSC, which was demonstrated to be a better indicator of D2cm^3 accuracy in cervical brachytherapy.

References

[1] B. S. Gill *et al.*, "National Cancer Data Base analysis of radiation therapy consolidation modality for cervical cancer: the impact of new technological advancements," *Int. J. Radiat. Oncol. Biol. Phys.*, vol. 90, no. 5, pp. 1083–1090, 2014, doi: 10.1016/j.ijrobp.2014.07.017.

[2] F. Zabihollahy, A. N. Viswanathan, E. J. Schmidt, M. Morcos, and J. Lee, "Fully automated multiorgan segmentation of female pelvic magnetic resonance images with coarse-to-fine convolutional neural network," *Med. Phys.*, vol. 48, no. 11, pp. 7028–7042, Nov. 2021, doi: 10.1002/mp.15268.

[3] S. A. Yoganathan *et al.*, "Automatic segmentation of magnetic resonance images for high-dose-rate cervical cancer brachytherapy using deep learning," *Med. Phys.*, vol. 49, no. 3, pp. 1571–1584, Mar. 2022, doi: 10.1002/mp.15506.

[4] R. Potter *et al.*, "Recommendations from gynaecological (GYN) GEC ESTRO working group (II): concepts and terms in 3D image-based treatment planning in cervix cancer brachytherapy-3D dose volume parameters and aspects of 3D image-based anatomy, radiation physics, radiobiology," *Radiother. Oncol.*, vol. 78, no. 1, pp. 67–77, 2006, doi: 10.1016/j.radonc.2005.11.014.

[5] F. Caliva, C. Iriondo, A. M. Martinez, S. Majumdar, and V. Pedoia, "Distance map loss penalty term for semantic segmentation," *arXiv preprint arXiv:1908.03679*, 2019.

[6] S. Nikolov *et al.*, "Clinically Applicable Segmentation of Head and Neck Anatomy for Radiotherapy: Deep Learning Algorithm Development and Validation Study," *J. Med. Internet Res.*, vol. 23, no. 7, p. e26151, Jul. 2021, doi: 10.2196/26151.

[7] F. Vaassen *et al.*, "Evaluation of measures for assessing time-saving of automatic organ-at-risk segmentation in radiotherapy," *Phys Imaging Radiat Oncol*, vol. 13, pp. 1–6, 2020, doi: 10.1016/j.phro.2019.12.001.

[8] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation," *Nat. Methods*, vol. 18, no. 2, pp. 203–211, Feb. 2021, doi: 10.1038/s41592-020-01008-z.

[9] J. Marsilla *et al.*, "SCARF: Auto-Segmentation Clinical Acceptability & Reproducibility Framework for benchmarking essential radiation therapy targets in head and neck cancer," *bioRxiv*, Jan. 25, 2022. doi: 10.1101/2022.01.15.22269276.

[10] A. L. Damato *et al.*, "Dosimetric consequences of interobserver variability in delineating the organs at risk in gynecologic interstitial brachytherapy," *Int. J. Radiat. Oncol. Biol. Phys.*, vol. 89, no. 3, pp. 674–681, Jul. 2014, doi: 10.1016/j.ijrobp.2014.03.005.

[11] F. K. Duane *et al.*, "Impact of delineation uncertainties on dose to organs at risk in CT-guided intracavitary brachytherapy," *Brachytherapy*, vol. 13, no. 2, pp. 210–218, Mar-Apr 2014, doi: 10.1016/j.brachy.2013.08.010.

Automatic target propagation strategies for MRI-guided cervical brachytherapy

Rita Simões[1], Eva C. Schaake[1], Eva E. Rijkmans[1], Marlies E. Nowee[1] and Tomas M. Janssen[1]

[1] Department of Radiation Oncology, the Netherlands Cancer Institute, Amsterdam, the Netherlands

Abstract The target structures required for cervical cancer brachytherapy treatment planning are delineated by radiation oncologists using both imaging and clinical information. The segmentations are made manually from scratch at the first brachytherapy fraction. For subsequent fractions the segmentations from the first fraction are rigidly propagated. This process is time- and resource-intensive, which makes the workflow inefficient and, since the patient is waiting with the applicator in situ, uncomfortable for the patient. In this work, we compared different strategies for automatic target propagation from the first to the second and third fractions. Registration-based approaches performed the worst, while segmentation-based approaches that take prior patient-specific information into account reached the best performance. In particular, patient-specific fine-tuned deep learning segmentation models are a promising strategy for automating the target propagation workflow.

1 Introduction

The standard-of-care treatment for locally advanced cervical cancer patients consists of external beam radiotherapy with concurrent chemotherapy followed by three to four fractions of image-guided adaptive brachytherapy. At each brachytherapy fraction an MRI is made with applicator in situ. The radiation oncologist manually delineates three target volumes following the EMBRACE II study guidelines [1]: the gross tumor volume (GTV), the high-risk clinical tumor volume (HR-CTV) and the intermediate-risk clinical tumor volume (IR-CTV). The GTV is defined as the macroscopic disease visible on the daily MRI. The HR-CTV is defined as the GTV plus the whole cervix and adjacent residual pathologic tissue, as determined via imaging and clinical examination. The IR-CTV is defined as the pretreatment tumor area plus a margin surrounding the HR-CTV [1]. The IR-CTV is therefore partially defined by information which is not present in the daily brachytherapy MRI images, as in general the tumor regresses during the course of external beam radiotherapy. The target structures delineated manually at the first brachytherapy fraction incorporate these patient-specific clinical decisions. An important question is how to optimally propagate this information from the first to the subsequent fractions.

In our clinic, structures are propagated with rigid registration, after which the radiation oncologist checks and manually edits them. This process, particularly the manual editing, is time and resource intensive and represents a bottleneck in the brachytherapy workflow. As a result, in our clinic a dedicated team of two radiation oncologists, two RTTs and one physicist can typically treat only two patients per day. Importantly, the time spent during manual segmentation is also unpleasant for the patient, as they need to wait immobilized in bed, with the applicator in situ, until the treatment plan is ready to be delivered. Automating (even if only partially) this process is likely to help speed up the entire workflow, improving logistics and patient comfort. Different approaches for automatically propagating patient-specific prior information have been proposed in the literature. The first has been traditionally used in various radiotherapy applications and consists of performing rigid or deformable image registration between the previous and current fraction scans and propagating the segmentations accordingly [2, 3]. Although registration methods inherently take patient-specific information into account, they are often not the best option when large deformations are present [3].

Alternatively, auto-segmentation methods can be employed. However, conventional auto-segmentation methods rely only on a single input image. Recently, strategies to incorporate patient-specific prior information into auto-segmentation models have been proposed, namely in the context of online adaptive radiotherapy. One strategy consists of training an auto-segmentation model with not only the image of the current fraction but also patient-specific information from previous time points as extra input [4]. Another strategy consists of fine-tuning a pre-trained, population-based model with the previous fraction(s) of the same patient, thereby overfitting the model and rendering it patient-specific [5-8].

In this work, we compare various approaches for automated target segmentation during the second and the third cervical brachytherapy fractions: 1) using rigid registration (our current clinical practice), 2) using deformable registration, 3) segmentation model trained on images as single inputs (without patient-specific prior); 4) segmentation model that takes as inputs both the current fraction image and the previous fraction segmentations; 5) patient-specific segmentation models resulting from fine-tuning the model in 3) with the previous fraction image and segmentations.

2 Materials and Methods

DATASETS

Two datasets of axial T2-weighted MRI images and the corresponding target segmentations made by a radiation oncologist and reviewed by another one for clinical practice were retrieved from our clinical database. The first dataset consisted of data of 209 cervical cancer patients (575 brachytherapy treatment fractions; median 3 fractions per patient) treated with brachytherapy between 2011 and 2019. This dataset was used for training and validating the auto-segmentation models. An additional dataset of 29 patients

(79 fractions; median 3 fractions per patient) treated between February 2022 and August 2023 was collected. This dataset was used for independent testing of all different target propagation and segmentation approaches. In this test dataset, the first treatment fraction was segmented by the radiation oncologists manually from scratch, with pre-treatment diagnostic information available. For the second and third fractions, the structures were in clinical practice rigidly propagated from the previous fraction by a radiation technology therapist. The propagated structures were then checked and manually edited by a radiation oncologist. A second radiation oncologist reviewed all structures. Both structure sets (rigidly propagated and manually edited) were retrieved from the database. All segmentations and registrations were performed in Mirada RTx. The institutional review board approved the study (IRBd20276). Informed consent was waived considering the retrospective design.

METHOD 1: RIGID REGISTRATION

The rigidly propagated structures retrieved from our database were defined, in this study, as the outputs of the first target propagation method that we aimed to evaluate. As this method is used in our current clinical practice, it served as a benchmark for comparison.

METHOD 2: DEFORMABLE REGISTRATION

For the second and third fractions, the multi-modality deformable image registration tool available in Mirada RTx was used with its default parameters to register the previous fraction to the current fraction image. The clinical segmentations were propagated according to the resulting deformation field.

METHOD 3: IMAGE-ONLY SEGMENTATION

The first dataset was split into 168 patients (455 fractions) for training and 41 patients (120 fractions) for validation. A 3D nnU-Net model was trained to segment the daily MRI using the default parameters [9]. The inputs to the model were the axial T2 images. The final model was the one which performed best in the validation set over the course of the pre-defined 1000 training epochs.

METHOD 4: IMAGE + PRIOR SEGMENTATION

A subset of the first dataset containing only the second and third fractions was used to train a model that took as inputs both the current fraction image and the previous fraction's segmentations. The resulting dataset consisted of 150 patients (277 fractions) for training and 39 patients (75 fractions) for validation. The toolbox `elastix` [10, 11] was used with parameter file *Param0001* [11] to automatically rigidly register the previous fraction to the current fraction images. The structures were propagated accordingly. A 3D nnU-Net model was then trained for 1000 epochs to segment the daily MRI using both the daily MRI and the registered segmentation of fraction 1 as input channels. The final model was the one that performed best in the validation set during training.

METHOD 5: PATIENT-SPECIFIC FINE-TUNING

For each patient in the second dataset, the previous fraction image and the corresponding clinical segmentations were used to fine-tune the population-based model (method 3), resulting in patient- and fraction-specific overfitted models. Similarly to [5], we allowed for training of all layers. Fine-tuning was stopped after 5 epochs, as it was empirically observed that results changed only negligibly after that. For the prediction of the third fraction, the patient-specific model obtained after training on the first fraction was further fine-tuned for 5 epochs on the second fraction. The nnU-Net framework was used for this purpose.

PERFORMANCE EVALUATION

For the second and the third fractions of each patient in the test set (50 fractions in total), the clinically used manual segmentations were geometrically compared with the segmentations resulting from the various target propagation strategies using the surface Dice (with a tolerance of 3 mm) [12], the mean surface distance (MSD) and the added path length (APL) [13]. These metrics were used as these have been shown to correlate more strongly with manual editing time than e.g. the Dice coefficient [13]. Pairwise comparisons between the performance of the different approaches were tested for statistical significance with the Wilcoxon signed-rank test.

3 Results

The performance of the various target propagation approaches is depicted in Figure 1. Both segmentation methods that employ a patient-specific prior (4 and 5) generally outperformed the registration-based (methods 1 and 2) and image-only segmentation (method 3) techniques. The patient-specific fine-tuned models consistently outperformed deformable registration and the image-only segmentation model. Results were statistically significant ($p \leq 0.05$) for all structures and metrics. Compared to the current clinical practice of rigid registration, the patient-specific models were statistically significantly better for the three structures in at least one of the metrics: surface Dice and MSD for the IR-CTV, surface Dice for the HR-CTV and all metrics for the GTV. This approach also significantly outperformed the image+prior segmentation approach for all metrics and structures, except for the surface Dice and MSD in the GTV. The image+prior segmentation model significantly outperformed the image-only model in all metrics for the GTV and in the surface Dice and MSD for the IR-CTV.

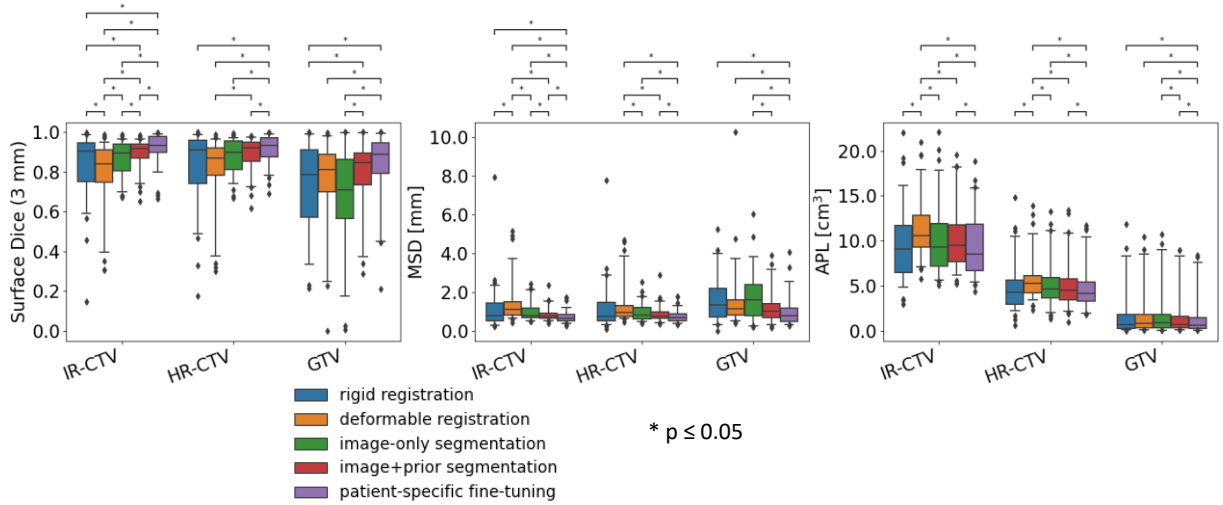


Figure 1 - Performance of the different target propagation approaches for the second and third fractions of the test set patients.

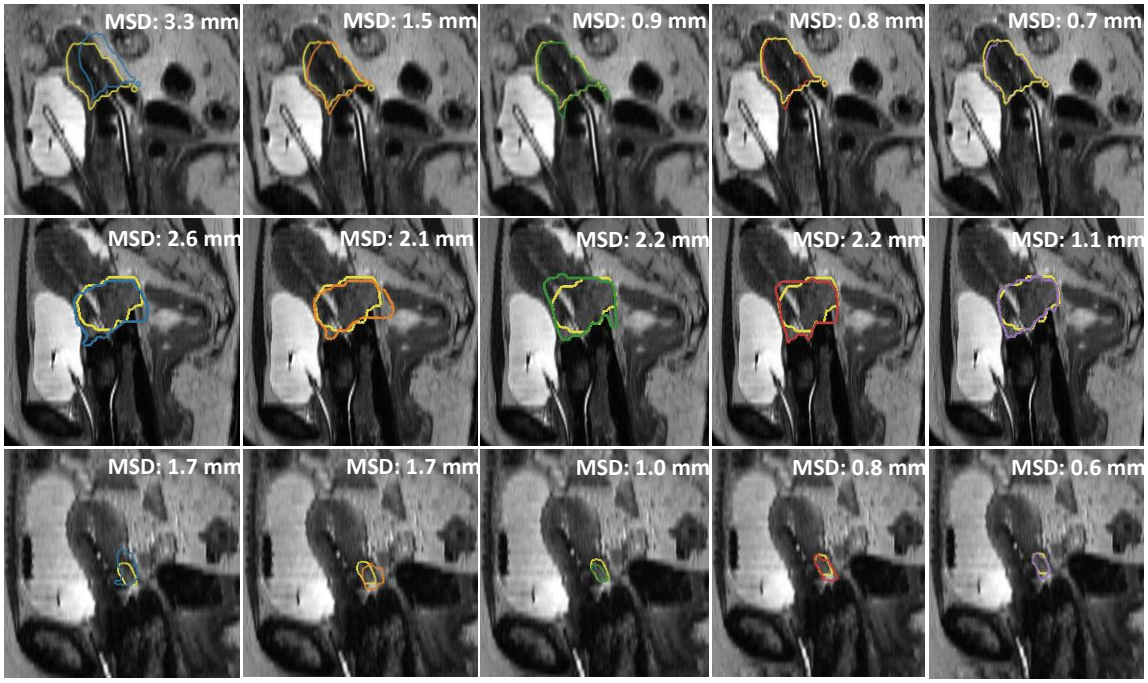


Figure 2 – Examples of IR-CTV (top row), HR-CTV (middle row) and GTV (bottom row) segmentations on the T2 images of three different patients (sagittal view, cropped). The ground truth is depicted in yellow and the outputs of the different target propagation methods are identified by different colors (conform Figure 1): a) rigid registration; b) deformable registration; c) image-only segmentation; d) image+prior segmentation; e) patient-specific fine-tuning.

Figure 2 shows segmentation examples of the IR-CTV, the HR-CTV and the GTV in three different patients. In the case of the IR-CTV, the rigid propagation failed to position the structure. Deformable registration improved upon rigid registration but failed at segmenting high contrast regions such as the uterus boundaries. The image-only segmentation model was able to correctly detect these boundaries but over-segmented the IR-CTV caudally. The patient-specific methods were able to capture the radiation oncologist's intent by ending the delineation caudally at the correct place. In the case of the HR-CTV, the patient-specific fine-tuned models clearly outperformed the other approaches, including the image+prior segmentation model. In the GTV example, the rigidly propagated structure is larger than the ground truth, indicating tumor regression. The deformable registration and the image-only segmentation model respectively over- and under-segmented the GTV considerably. The two patient-specific methods outperformed the other approaches.

4 Discussion

The patient-specific methods outperformed the registration- and image-only segmentation approaches for target propagation in the second and third fractions of cervical cancer brachytherapy. In particular, the patient-specific fine-tuned models yielded the best performance for most metrics and structures.

The registration-based methods performed generally worse than the segmentation-based methods. The IR-CTV is partially defined by information that is not present on the images, so its boundaries are not necessarily strictly anatomical. Additionally, inter-fraction anatomical changes in the pelvic region can be large [14], rendering methods

that are purely based on image registration less suitable for this task. Deformable registration performed actually worse than rigid registration for all structures and metrics. A possible explanation is that the images were registered globally. Defining an area for local registration would have possibly resulted in improved registrations.

The image-only segmentation model performed on average comparably to the rigid registration but was able to improve upon it for the poorest-performing cases. This indicates that an image-only segmentation approach may be a suitable strategy for assisting the radiation oncologists in manually segmenting the first fraction, for which no segmentation starting point is available. The image+prior segmentation model, however, significantly outperformed the image-only segmentation model for most structures and metrics. This demonstrates the benefit of incorporating patient-specific information into the auto-segmentation model.

We hypothesized that the propagation of the IR-CTV would benefit most from the incorporation of patient-specific information, since this structure is defined as the pre-treatment tumor area. Because the tumor often regresses during the course of treatment, this information is generally no longer present in the brachytherapy MRI images. Indeed, for most performance metrics the two patient-specific segmentation methods significantly outperformed the other techniques.

The GTV is defined as the visible tumor on the daily images, however, tumor auto-segmentation is a complex task due to the large heterogeneity in shape, size and texture of these structures. Adding extra patient-specific information might therefore help in simplifying the task. Indeed we found that, for the GTV, the patient-specific fine-tuned segmentation models significantly outperformed the other methods in all metrics. This indicates that tumor auto-segmentation also benefits from patient-specific information, even in the presence of inter-fraction tumor regression.

It is worth noting that the ground truth delineations used in this study were obtained by editing the rigidly propagated structures. The performance of the rigid registration method is therefore likely overestimated, as the radiation oncologists were biased towards these structures. A more realistic estimate would require the manual segmentations to be performed without a starting point. However, this data was not available as this is not done in clinical practice for the second and third fractions.

5 Conclusion

In this study we compared the performance of various automatic approaches for propagating the three target volumes over the course of brachytherapy for cervical cancer patients. The patient-specific approaches outperformed both registration- and image-only segmentation approaches. In particular, fine-tuned segmentation models have the potential to reduce clinical staff workload, thereby improving the treatment efficiency and increasing patient comfort.

References

[1] Pötter R, Tanderup K, Kirisits C, de Leeuw A, Kirchheiner K, Nout R, et al. The EMBRACE II study: The outcome and prospect of two decades of evolution within the GEC-ESTRO GYN working group and the EMBRACE studies. Clinical and Translational Radiation Oncology. 2018;9:48-60. doi:https://doi.org/10.1016/j.ctro.2018.01.001.

[2] Brock KK, Mutic S, McNutt TR, Li H, Kessler ML. Use of image registration and fusion algorithms and techniques in radiotherapy: Report of the AAPM Radiation Therapy Committee Task Group No. 132. Med Phys. 2017;44:e43-e76. doi:10.1002/mp.12256.

[3] Rigaud B, Simon A, Castelli J, Lafond C, Acosta O, Haigron P, et al. Deformable image registration for radiation therapy: principle, methods, applications and evaluation. Acta Oncologica. 2019;58:1225-37. doi:10.1080/0284186X.2019.1620331.

[4] Ferreira Silverio N, van den Wollenberg W, Betgen A, Wiersema L, Marijnen C, Peters F, et al. MO-0798 Evaluation of target autocontouring for MRI-guided online adaptive treatment of rectal cancer. Radiotherapy and Oncology. 2023;182:S668-S9. doi:10.1016/S0167-8140(23)08462-1.

[5] Smolders A, Lomax A, Weber DC, Albertini F. Patient-specific neural networks for contour propagation in online adaptive radiotherapy. Physics in Medicine & Biology. 2023;68:095010. doi:10.1088/1361-6560/accaca.

[6] Fransson S, Tilly D, Strand R. Patient specific deep learning based segmentation for magnetic resonance guided prostate radiotherapy. Phys Imaging Radiat Oncol. 2022;23:38-42. doi:10.1016/j.phro.2022.06.001.

[7] Elmahdy MS, Ahuja T, van der Heide UA, Staring M. Patient-Specific Finetuning of Deep Learning Models for Adaptive Radiotherapy in Prostate CT. 2020. p. arXiv:2002.06927. doi:10.48550/arXiv.2002.06927.

[8] Li Z, Zhang W, Li B, Zhu J, Peng Y, Li C, et al. Patient-specific daily updated deep learning auto-segmentation for MRI-guided adaptive radiotherapy. Radiother Oncol. 2022;177:222-30. doi:10.1016/j.radonc.2022.11.004.

[9] Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods. 2021;18:203-11. doi:10.1038/s41592-020-01008-z.

[10] Shamonin DP, Bron EE, Lelieveldt BP, Smits M, Klein S, Staring M. Fast parallel image registration on CPU and GPU for diagnostic classification of Alzheimer's disease. Front Neuroinform. 2013;7:50. doi:10.3389/fninf.2013.00050.

[11] Klein S, Staring M, Murphy K, Viergever MA, Pluim JP. elastix: a toolbox for intensity-based medical image registration. IEEE Trans Med Imaging. 2010;29:196-205. doi:10.1109/tmi.2009.2035616.

[12] Nikolov S, Blackwell S, Zverovitch A, Mendes R, Livne M, De Fauw J, et al. Deep learning to achieve clinically applicable segmentation of head and neck anatomy for radiotherapy. arXiv preprint arXiv:180904430. 2018.

[13] Vaassen F, Hazelaar C, Vaniqui A, Gooding M, van der Heyden B, Canters R, van Elmpt W. Evaluation of measures for assessing time-saving of automatic organ-at-risk segmentation in radiotherapy. Physics and Imaging in Radiation Oncology. 2020;13:1-6. doi:https://doi.org/10.1016/j.phro.2019.12.001.

[14] Patil G, Br KK, Narayanan G. Inter-Fractional Variations in Volume and Radiation Dose to the Organs at Risk, High-Risk Clinical Target Volume and Implication of Image-Guided Adaptive Planning During Intracavitary Brachytherapy of Carcinoma Cervix. Cureus. 2022;14:e21503. doi:10.7759/cureus.21503.

Geant4 simulation of DNA damage surrounding a DaRT seed

Laura Ballisat[1], Chiara De Sio[1], Lana Beck[1], Susanna Guatelli[2], Dousatsu Sakata[3,2,1], Yuyao Shi[1], Jinyan Duan[1], Jaap Velthuis[1], Anatoly Rosenfeld[2]

[1]*School of Physics, University of Bristol, Bristol, UK*
[2]*Centre for Medical Radiation Physics (CMRP), University of Wollongong, NSW, Australia*
[3]*Division of Health Sciences, Osaka University, Osaka 565-0871, Japan*

Abstract Diffusing alpha-emitters radiation therapy (DaRT) is a brachytherapy technique using ^{224}Ra seeds. This treatment utilises the high linear energy transfer (LET) and short range of α-particles. Treatment planning of DaRT requires a good understanding of the dose from all particles in the ^{224}Ra decay chain. Geant4 and Geant4-DNA are used in this study to simulate a DaRT seed and the DNA damage due to this treatment. Simulations show the majority of the dose is delivered in the first 3 weeks following insertion, therefore there is little additional dose delivered if seeds are not removed following treatment. The dose close to the seed is predominantly due to α-particles, however the contribution of electrons becomes greater at radial distances beyond 4 mm. It is therefore important to include the contribution from all particles in treatment planning. The DNA damage surrounding the seed becomes less complex with distance, due to the change in particle spectrum. The radiobiological effectiveness (RBE) has been calculated as a function of number of double strand breaks (DSBs) and complex DSBs. RBE_{DSB} is approximately 2 close to the seed and tends to unity with distance. RBE_{cDSB} is greater, in the range 5-7, close to the seed but also tends to unity with distance. This should be considered when determining the spacing between seeds to ensure the required therapeutic effect is maintained. The reduced RBE with distance from the source results in less complex, more easily repaired, damage to surrounding normal tissues. Hence DaRT has the potential to be a conformal treatment delivering a high RBE to the treatment volume.

1 Introduction

Diffusing alpha-emitters radiation therapy (DaRT) is an interstitial brachytherapy treatment using seeds containing ^{224}Ra. This treatment takes advantage of the multiple alpha decays in the ^{224}Ra decay chain. α-particles have a high linear energy transfer (LET) and short path length, therefore highly lethal doses can be delivered close to the seed. DaRT also benefits from the additional treatment range due to the diffusion of the daughter nuclei in the decay chain, in particular ^{220}Rn and ^{212}Pb. This increases the effective treatment distance to several millimetres [1] enabling better tumour coverage whilst maintaining good spatial protection of surrounding normal tissue. DaRT is currently undergoing several clinical trials, including a prospective trial treating locally recurrent squamous cell carcinomas [2].

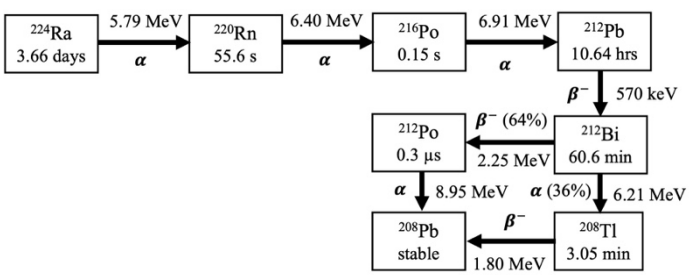


Figure 1 – ^{224}Ra decay chain.

A good understanding of the DNA damage surrounding the seed is needed for the treatment planning of DaRT. All particles in the ^{224}Ra chain, see Fig. 1, contribute to the DNA damage. Due to the diffusion of daughter nuclei, the particle spectrum (the type of particle, the energy of particle and quantity of particles) will vary with distance from the ^{224}Ra seed. The energy and particle type affect the induced DNA damage, therefore an understanding of the particle spectrum is important for the treatment planning of DaRT.

2 Materials and Methods

This study uses Geant4 (version 11.1) [3,4,5] and the low energy extension Geant4-DNA to simulate a DaRT seed and the initial DNA damage due to this treatment. Two individual simulations are used. Firstly, the DaRT seed is simulated, and the particle spectrum saved in a phase space file at specified distances from the seed. Secondly this phase space file is used as the input to the DNA damage simulation.

2.1 DaRT seed simulation

The DaRT seed is simulated as a stainless-steel cylinder with radius 0.15 mm and length 6 mm [6] as shown in Fig. 2. ^{224}Ra ions are placed at a depth of 0.15 nm. The Geant4-DNA physics constructors *G4EmPenelopePhysics* and *G4RadioactiveDecayPhysics* are used to simulate electromagnetic interactions and radioactive decay.

The whole decay chain of ^{224}Ra is simulated. Daughter nuclei from these decays which enter the tumour due to recoil slowly diffuse away from the seed. Only the diffusion of ^{220}Rn, ^{212}Pb and ^{212}Bi is simulated. ^{216}Po and ^{212}Po have relatively short half-lives and therefore it is assumed that diffusion away from their parent nuclei is negligible. ^{208}Tl has a longer half-life, however the diffusion constant is

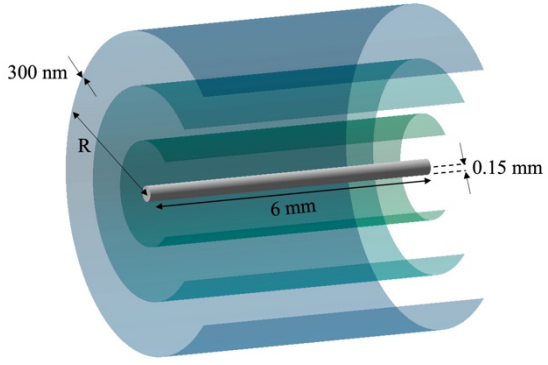


Figure 2 – DaRT seed simulation geometry.

uncertain and has therefore not been included in the simulation. The diffusion of ^{220}Rn, ^{212}Pb and ^{212}Bi is simulated using a random walk with diffusion constants: D_{Rn} = 1.12x10^{-3} mm^{2}s^{-1}, D_{Pb} = 1.22x10^{-5} mm^{2}s^{-1} and D_{Bi}= 6.78x10^{-7} mm^{2}s^{-1} [1].

Some decay products are removed from the tumour environment via biological processes. In particular ^{212}Pb can bind to surrounding molecules such as proteins and be removed via vascular routes. The amount of leakage has been found to vary significantly with tumour size between 12-90% [6]. Therefore, the leakage has been simulated to have a probability of 50%. It is assumed that the leakage of all other nuclei is negligible.

Particles which are not removed due to leakage are all tracked. Any particles crossing rings at specified distances from the seed, see Fig. 2, are saved to the phase space file for the DNA damage simulation.

2.2 DNA damage simulation

A previously developed simulation [7] is used to simulate the direct and indirect damage due to DaRT from which the yield of DNA strands breaks is calculated. This model requires an input particle distribution, see Section 2.1, and a DNA model.

The DNA model consists of a small section of a cell nucleus as shown in Fig. 3. The *fractaldna* python package (version 0.6.0) [8] was used to produce straight and turned segments of solenoid chromatin fibre. These segments are placed along a self-avoiding walk.

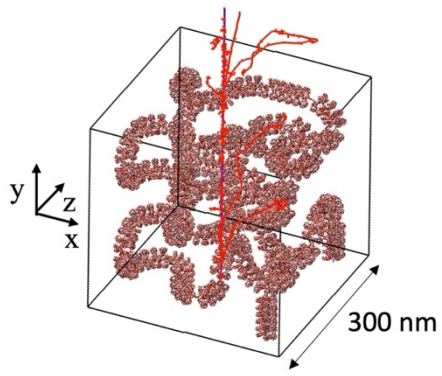


Figure 3 – α-particle crossing the simplified cell nucleus. Deoxyribose molecules are shown as red and grey spheres.

The *G4EmDNAPhysics_option7* physics constructor was used for all volumes, which tracks particles down to the nanoscale in liquid water. Therefore, all volumes in the simulation are considered to be liquid water.

The methodology for simulating direct and indirect damage is described in [7] and is described here briefly.

To calculate the initial distribution of direct stand breaks energy depositions are accumulated within 0.35 nm of each deoxyribose-phosphate molecule. The linear damage model [9] is then applied with a minimum of 5 eV and a maximum of 37.5eV. For indirect strand breaks a 40.5% probability [10,11] of a strand break occurring following a reaction between a hydroxyl radical and the deoxyribose-phosphate backbone is used. The radiolysis of water is simulated up until 5 ns, and radicals more than 9 nm from the deoxyribose-phosphate backbone are removed from the simulation to simulate scavenging by the medium [10,11].

Isolated strand breaks are easy for a cell to repair, whilst many strand breaks in close proximity are potentially more difficult to repair and more likely to lead to cell death [12]. Strand breaks have been classified using the methodology presented in [13]. Double strand breaks (DSBs) are considered to be the most lethal damage. A DSB contains at least one strand break on each of the DNA strands within 10 base pairs. DSBs can be further classified into simple and complex DSB. Only one strand break on each strand within 10 base pairs is considered to be a simple DSB. A DSB containing more strand breaks within 10 base pairs is a complex DSB.

The radiobiological effectiveness (RBE) is a commonly used metric for comparing high LET therapy to conventional low LET therapy. It can be calculated [14]:

$$RBE = \frac{Dose_{endpoint/photon}}{Dose_{endpoint/quality}}$$

where the endpoint is a biological quantity of interest. DSBs are more difficult to repair than simpler strand breaks, and therefore are considered to be the main cause of cell death [12]. Therefore, in this work the endpoint considered is DSB induction. For the low LET photon comparison, the same DNA model was used to calculate the strand break yield from a ^{60}Co source at a radius of 5 mm from the centre of the cell nucleus.

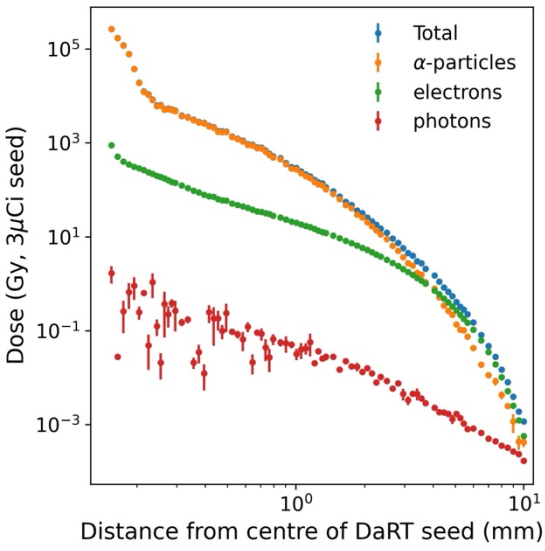


Figure 4 –Dose as a function of radial distance from the centre of a 3 μCi/cm DaRT seed, for all particles, α-particles (including delta electrons), electrons, and photons.

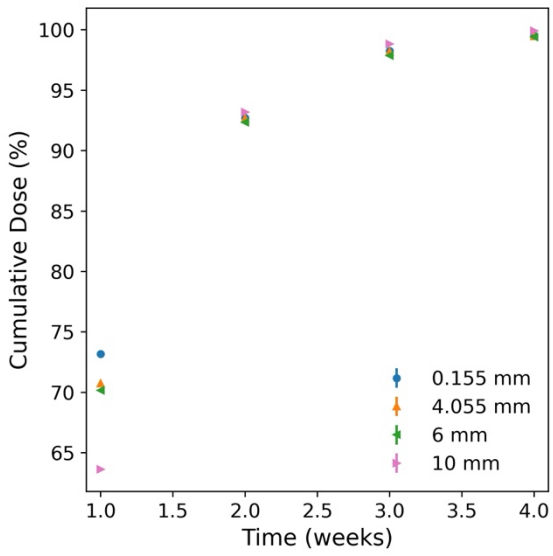


Figure 5 – Cumulative dose as a percentage of the total dose as a function of time since seed insertion.

3 Results and Discussion

Figure 4 shows the dose from the DaRT seed as a function of distance due to α-particles, electrons and photons. There is a sharp fall off in dose with increasing distance. Tissue more than 4.3 mm from the seed receives less than 1 Gy which allows DaRT to be a very conformal treatment. Close to the seed α-particles deposit the majority of the dose. However, the contribution of electrons increases with distance. At distances greater than 4 mm electrons deposit the majority of the dose, whilst photons deliver a much smaller dose at all distances. This change in particle distribution is important for the effectiveness of the treatment, as α-particles produce more complex, harder to repair strand breaks than electrons and photons.

Figure 5 shows the percentage of the total dose delivered at four distances from the seed as a function of time after insertion. After 3 weeks at least 98% of the dose is delivered at all distances from the seed. Therefore, for superficial treatments the seeds can be removed after 3 weeks with little impact on the delivered dose. In contrast, for treatment sites where seed removal is not feasible the extra dose delivered to surrounding tissue is not significant.

As expected, the change in particle spectrum with distance affects the yield of DNA damage. Figure 6 shows the yield of strand breaks, DSBs and complex DSBs normalised to the dose at a given distance from the seed. DNA damage due to α-particles is the major source of damage for all types of strand break close to the seed. However, the contribution due to electrons increases for individual strand breaks and DSB becoming greater than α-particles beyond distances of 3.0 mm and 5.8 mm, respectively. α-particles have a higher LET than electrons and therefore cause more complex damages. The total yield of strand breaks increases with distance from the seed. However, the yield of DSB decreases due to the decrease in dose deposition from α-particles. The complexity of DSB also decreases with distance from the seed. Whilst some DSB are formed due to electrons distal from the seed, there are fewer complex DSB due to the lower LET. Thus, the overall DNA damage complexity decreases with distance from the seed, simpler damage is more easily repaired and therefore the effectiveness of treatment is reduced at greater distances from the seed.

This is reflected in the calculated RBE. Both RBE_{DSB} and RBE_{cDSB} tend to unity at large distances from the seed as shown in Fig. 7. RBE_{DSB} is approximately 2 until 1.0 mm from the seed after which the RBE decreases to approximately unity. RBE_{cDSB} is much higher in the range 5-7 close to the seed, until 1.2 mm from the seed after which

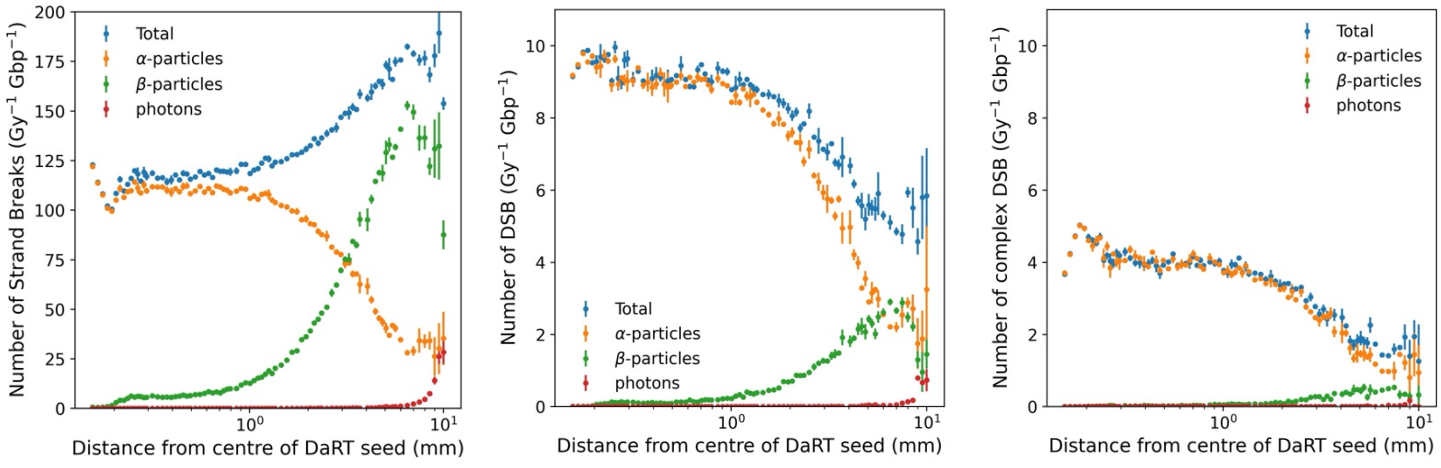


Figure 6 – Yield of (left) total strand breaks, (middle) DSBs and (right) complex DSBs as a function of radial distance from the centre of the DaRT seed for all particles, and due to particles produced due to radioactive decay (subsequent secondaries are included with the decay product).

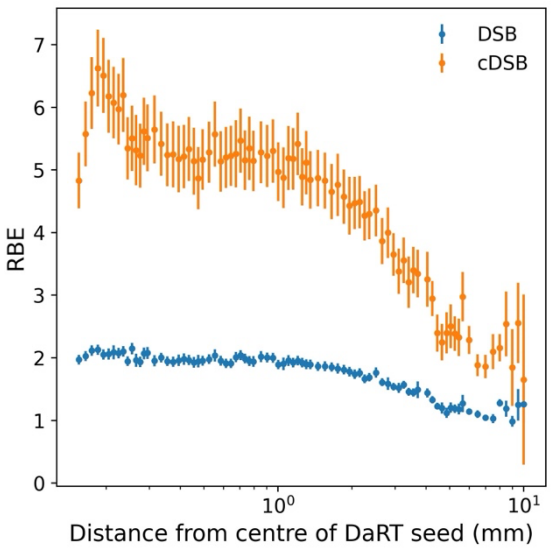


Figure 7 – RBE as a function of radial distance from the centre of the DaRT seed using ^{60}Co as the reference radiation.

it also decreases towards unity. This change in treatment effectiveness should be considered in treatment planning to determine the spacing of seeds such that the therapeutic effect is maintained between seeds. If the distance is too large the RBE between seeds will decrease. Since less complex damages are more easily repaired this could result in a reduction in treatment effectiveness. The decrease in RBE is however beneficial for surrounding normal tissues as less complex damages are easier to repair resulting in a lower probability of normal tissue complications.

5 Conclusion

The high LET and short range of α-particles allow the targeted treatment of cancer. DaRT uses ^{224}Ra seeds, which decays releasing four α-particles per decay. The aim of this study is to better understand the DNA damage surrounding a DaRT seed from all particles in the ^{224}Ra decay chain. To do this a DaRT seed and the DNA damage radially surrounding it have been simulated using Geant4 and Geant4-DNA.

The majority of the dose is delivered in the first 3 weeks following treatment, therefore the extra dose delivered from not removing seeds following treatment is not significant.

It has been shown that the distribution of particles changes as a function of distance from the seed, which affects the DNA damage. Whilst more strand breaks occur at larger distances from the seed, the complexity of damages decreases as a function of distance. The majority of the deposited dose up until a distance of 4.0 mm is due to α-particles, which having a higher LET results in more complex damages. At greater distances however, electrons deliver the greater dose. Such electrons have a lower LET and hence cause less complex damages which are potentially easier for a cell to repair. This causes the RBE to decrease with distance from the source. This is an important consideration for treatment planning to ensure the minimum tumour volume dose coverage considering the differences in RBE. However, this is beneficial for surrounding normal tissue, which receive less dose due to the rapid fall off and the RBE is lower hence DNA damage is less complex and easier to repair. Therefore, DaRT has the potential to be a very conformal treatment whilst still delivering an effective dose to the treatment volume.

References

[1] Arazi L. “Diffusing alpha-emitters radiation therapy: approximate modeling of the macroscopic alpha particle dose of a point source”. Phys. Med. Biol. 65 (2020). DOI: https://doi.org/10.1088/1361-6560/ab5b73.

[2] Popovtzer A et al. “Initial Safety and Tumor Control Results From a "First-in-Human" Multicenter Prospective Trial Evaluating a Novel Alpha-Emitting Radionuclide for the Treatment of Locally Advanced Recurrent Squamous Cell Carcinomas of the Skin and Head and Neck”. IJROBP, 106 (2020), p 571-8. DOI: https://doi.org/10.1016/j.ijrobp.2019.10.048.

[3] Agostinelli S et al. “GEANT4 - A simulation toolkit”. Nucl. Instrum. Methods Phys. Res., 506 (2003), p 250-303. DOI: https://doi.org/10.1016/S0168-9002(03)01368-8.

[4] Allison J et al. “Geant4 Developments and Applications”. IEEE Trans. Nucl. Sci. 53 (2006). DOI: https://doi.org/10.1109/TNS.2006.869826.

[5] Allison J et al. “Recent developments in GEANT4”. Nucl. Instrum. Methods Phys. Res. 835 (2016), p 186-225. DOI: https://doi.org/10.1016/J.NIMA.2016.06.125.

[6] Arazi L et al. “Treatment of solid tumors by interstitial release of recoiling short-lived alpha emitters”. Phys. Med. Biol., 52 (2007), p 5025-42. DOI: https://doi.org/10.1088/0031-9155/52/16/021.

[7] Ballisat L et al. “In-silico calculations of DNA damage induced by α-particles in the ^{224}Ra DaRT decay chain for a better understanding of the radiobiological effectiveness of this treatment”. Physica Medica, 112 (2023). DOI: https://doi.org/10.1016/j.ejmp.2023.102626.

[8] Lampe N. “FractalDNA”. GitHub, (2021). https://github.com/natl/fractaldna.

[9] Friedland W et al. “Simulation of DNA Damage after Proton Irradiation”. Radiat. Res. 159 (2003), p 401-10. DOI: https://doi.org/10.1667/00337587(2003)159[0401:SODDAP]2.0.CO;2.

[10] Shin WG, et al. “A Geant4-DNA Evaluation of Radiation-Induced DNA Damage on a Human Fibroblast”. Cancers, 13 (2021). DOI: https://doi.org/10.3390/cancers13194940.

[11] Sakata D et al. “Prediction of DNA rejoining kinetics and cell survival after proton irradiation for V79 cells using Geant4-DNA”. Physica Medica, 105 (2023). DOI: https://doi.org/10.1016/j.ejmp.2022.11.012.

[12] Georgakilas AG et al. “Induction and repair of clustered DNA lesions: What do we know so far?” Radiat. Res. 180 (2013), p 100-9. DOI: https://doi.org/10.1667/RR3041.1.

[13] Nikjoo H et al. “Computational modelling of low-energy electron-induced DNA damage by early physical and chemical events”. Int. J. Radiat. Biol. 71 (1997) p 467-83. DOI: https://doi.org/10.1080/095530097143798.

[14] Valentin J. “Relative biological effectiveness (RBE), quality factor (Q), and radiation weighting factor (wR): ICRP Publication 92”. Annals of the ICRP, 33 (2003), p 1-121. DOI: https://doi.org/10.1016/S0146-6453(03)00024-1.

A pipeline for generating longitudinal head and neck cancer radiotherapy datasets for machine learning

Kayla O'Sullivan-Steben[1], Odette Rios-Ibacache[1], Luc Galarneau[1,2], and John Kildea[1,2]

[1]Medical Physics Unit, McGill University, Montreal, Canada
[2]Research Institute of the McGill University Health Centre, McGill University, Montreal, Canada

Abstract AI and Big Data have promised to revolutionize radiotherapy. However, there is a lack of readily available large-scale datasets needed to train AI models. Therefore, we developed an image processing pipeline that automatically sorts, registers, and preprocesses DICOM images and DICOM-RT files for use in downstream machine learning and analysis tasks. We focused on the use case of head and neck cancer (HNC), where patients often undergo anatomical changes that necessitate replanning. These anatomical changes can potentially be modeled and predicted with sufficient longitudinal imaging data. Our Python-based pipeline was validated on 370 HNC patients, resulting in a comprehensive HNC dataset containing 5,402 Cone-Beam CTs (CBCTs), 530 planning CTs, and related radiotherapy data. Using this dataset, we can produce various image-related metrics and have begun training deep learning algorithms for HNC replanning predictions. Overall, our pipeline is a useful tool for collecting and utilizing longitudinal real-world radiotherapy datasets for machine learning.

1 Introduction

The use of Artificial Intelligence (AI) in radiotherapy research continues to rise, but its clinical application is often hindered by the lack of large high-quality datasets. Therefore, we developed an image processing pipeline to generate ready-to-use datasets of longitudinal imaging data, with an initial focus on predicting anatomical changes in head and neck cancer (HNC) patients undergoing radiotherapy.

Big Data and AI are rapidly evolving areas of research that have promised to address a variety of challenges in radiotherapy, including treatment planning, patient monitoring, and outcomes prediction [1, 2]. One emerging AI strategy for personalized medicine is the Digital Twin concept, a dynamic virtual representation of an object that is updated continuously with new incoming real-world data to predict the object's future behaviour [3]. Unfortunately, successful development of AI models is often limited by the lack of accessible high-quality real-world radiotherapy data for research use [1, 4]. Digital twinning, in particular, requires both historical data as well as new data for continuous model updating. While databanking initiatives like The Cancer Imaging Archive (TCIA) [5] aim to provide large collections of imaging data for research, they currently lack comprehensive longitudinally-sequential radiotherapy data such as cone-beam CT (CBCT) images for Image-Guided Radiotherapy (IGRT). Pipelines to efficiently produce these datasets would facilitate digital twinning and other AI studies reliant on time-series data.

Furthermore, the process of extracting, cleaning, and curating these large datasets for analysis and machine learning can be laborious and time-consuming. Ethical concerns around patient re-identification often further limit data sharing [1, 4]. Therefore, to comprehensively address all these challenges, we developed an image processing pipeline that automatically sorts, registers, preprocesses, and analyses the DICOM images and DICOM-RT files of patients undergoing radiotherapy so that the processed data may be easily used for future downstream machine learning and analysis tasks.

The first use-case of our pipeline is to prepare a dataset for modelling and tracking anatomical changes of head and neck cancer (HNC) patients during radiotherapy. HNC patients often experience significant anatomical changes throughout treatment due to weight loss and tumour shrinkage, which can affect the planned dose distribution and compromise tumour coverage and organ at risk sparing [6, 7]. In such cases, the patient must have their radiotherapy treatment replanned. However, in many centres, radiotherapy replanning is resource-intensive and the decision to replan is often made with short notice to the planning team, resulting in workflow disruptions and suboptimal treatment planning across multiple patients.

The long-term clinical goal of our use case is to predict the progression of HNC patients' anatomical changes over the course of radiotherapy so that replanning can be anticipated and prepared in advance. Our methodology consists of training a variational autoencoder – a neural network composed of an encoder and decoder – on CBCT imaging data. However, we were unable to find readily available datasets providing complete radiotherapy and imaging data over the course of treatment for a sufficiently large number of HNC patients. Therefore, we designed our pipeline as a tool to transform raw data from the Treatment Planning System (TPS) into machine-learning-ready data for longitudinal image-based studies.

2 Materials and Methods

Our pipeline was built and incrementally tested using the data of HNC patients treated at our centre, which are stored and managed by the ARIA oncology information system (Varian Medical Systems, Inc. Palo Alto, California, USA). The use of these data was approved by our centre's Research

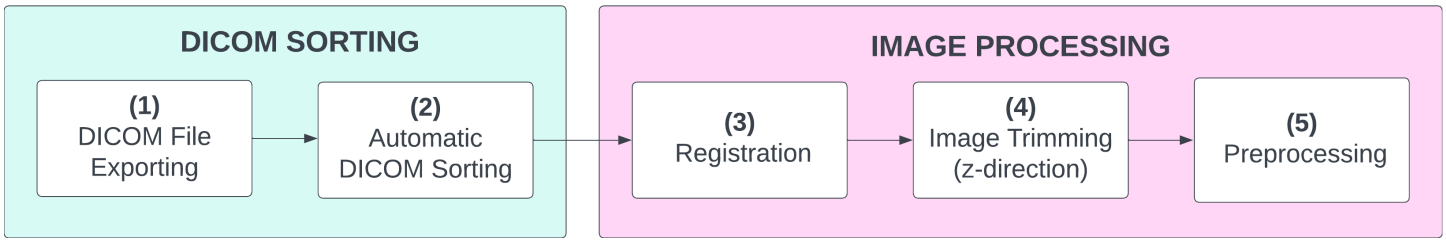


Figure 1: Schematic of our image sorting and processing pipeline.

Ethics Board. To retrieve the list of patients to be included in our study, we first queried the ARIA database for patients who were diagnosed with squamous cell carcinoma HNC and treated at our centre since 2017.

We then designed a pipeline that would start from the list of patient IDs and finish with a fully-preprocessed imaging dataset. The pipeline was coded in Python and consists of five main steps: DICOM exporting, sorting, registration, trimming, and preprocessing (Figure 1). Each step is described below.

1. Exporting: For each patient, all radiotherapy-related DICOM files are (by necessity) manually exported from the TPS. These files include planning CTs, CBCTs, image registration files, and DICOM-RT objects (RT Image, RT Structure Set, RT Plan, RT Dose, and RT Beams Treatment Record). The DICOM files are exported all at once (per patient) into a single unsorted directory and are anonymized by the TPS on export.

2. Sorting: DICOM images of different modalities (CT, CBCT, etc.) and their associated DICOM-RT and registration files are sorted into separate directories based on their metadata. When available, DICOM-RT Structure Set files, whose metadata contain the UIDs of each image slice in the reference scan, are used as an index to sort image slice files. This action accelerates the overall process by eliminating the need to open and read individual image slice DICOM files.

3. Registration: For each patient, the planning CTs and CBCTs from different time points are registered using rigid registration. Our code supports registration using both the DICOM registration files generated by the TPS as well as an optimizer algorithm that employs the SimpleITK library [8].

4. Trimming: CBCTs are cropped along the z-axis so that all images for a given patient cover the same anatomical region. The CBCTs can additionally be left untrimmed or trimmed to a specified reference point or structure in the planning CT Structure Set.

5. Preprocessing: Using OpenCV [9], a binary mask is applied to the patient's body, allowing for geometric calculations (cross-sectional area, perimeter, volume), noise removal (eg: removing the visible headrest), and image centering and cropping in the x-y plane. Additional preprocessing features include applying a window-level to the image, converting Hounsfield Units (HU) to standard pixel values, and resampling and resizing images. Preprocessing ensures that all images are standardized prior to further analysis. Finally, a simplistic outlier detection algorithm is applied using body area metrics and pixel means and variances to identify and (depending on the threshold applied) exclude low-quality and unusable images from the dataset.

To validate our work, we generated a variety of statistics and metrics relating to the imaging data. To evaluate the performance of our pipeline, we measured the execution time for each processing step for a sample patient.

3 Results

In this work, we developed and implemented a DICOM sorting and processing pipeline using Python and open-source libraries, which will be published on GitHub as an open-source tool. The pipeline was successfully tested on 370 squamous cell carcinoma HNC patients who underwent radiotherapy at our institution since 2017. This testing resulted in a comprehensive real-world retrospective radiotherapy and imaging dataset that is primed for use in analysis and machine learning.

In total, this 370-patient dataset contains 5,402 CBCT images, 530 planning CTs, and all associated DICOM-RT and registration files. 142 (38.4%) of the 370 patients had their treatments replanned, making this a reasonable dataset for research in predicting the need to replan HNC radiotherapy treatments.

From the prepared dataset, we can generate statistics on the frequency and timing of image acquisition and treatment delivery using the dates and fraction numbers embedded in the DICOM files. A summary of some of these values are presented in Table 1.

Once the images are registered, they can optionally be trimmed in the z-direction to cover precisely the same anatomical region across all images for a given patient. The amount of anatomical overlap in our dataset is shown in Table 2. An average and median overlap of 87 and 91 out of 93 slices respectively was achieved, indicating a high consistency in the imaged anatomy within an individual patient's scans.

We can also extract useful information from the preprocessed images themselves. For instance, geometric calculations such as body perimeter, cross-sectional area and volume can be calculated using the masking operation in the preprocessing step of the pipeline, allowing for a quick assessment of

	Mean	Std Dev	Median	Min	Max
Number of fractions between CBCT acquisitions	2.3	0.8	2.0	1	9
Days between planning CT and first fraction	14	5	13	5	50
Days between replan CT and first replan fraction	5	3	4	1	22
Fractions completed before replanned treatment	19	7	21	2	30

Table 1: Various statistics relating to timing of imaging acquisition and treatment delivery of the 370-patient prepared dataset.

Overlap	Mean	Std Dev	Median	Min	Max
Slices	87	9	91	33	93
Millimeters	171	16	179	64	183

Table 2: Anatomical overlap in the z-direction between CBCT images of a given patient. Each scan is taken with a set length of 93 slices and a slice thickness of 1.99 mm.

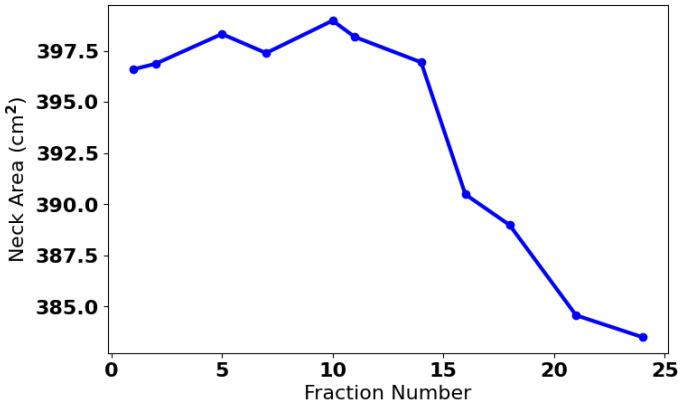


Figure 2: Cross-sectional area of a replanned patient's neck over the course of radiotherapy. The area was calculated on the image slice corresponding to the bottom of the mandible. This patient's neck area began decreasing around the 14th fraction until they were replanned at the 24th fraction.

anatomical changes that may be used for monitoring purposes. Figure 2 shows an example of the change in neck area (at the level of the mandible) of a demonstrative patient over the course of treatment before they were replanned at fraction 24.

Using the neck area metric, along with an outlier threshold set at 3 standard deviations from the mean of the image pixel values, our pipeline successfully identified several outlier images in our dataset, a subset of which are shown in Figure 3.

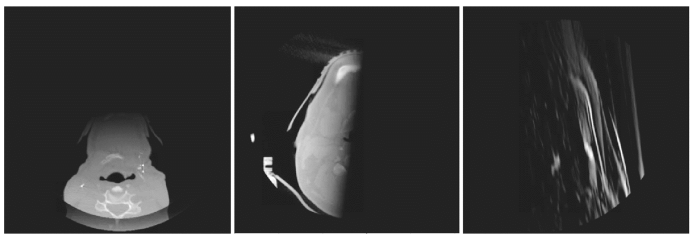


Figure 3: Examples of images successfully identified by our outlier detection algorithm.

Finally, our prepared images, and the metrics computed based on them, can be used as input to machine learning models. We are currently in the process of a training a variational autoencoder model on our prepared dataset.

Pipeline Step	Single Image	Entire Patient
1. Exporting	30 s	180 s
2. Sorting	3 s	26 s
3. Registration	11 s	60 s
4. Trimming	1 s	2 s
5. Preprocessing	1 s	5 s
Total (excl. exporting)	15 s	93 s
Total (incl. exporting)	45 s	273 s

Table 3: Approximate execution times (in seconds) of our pipeline on a sample patient with 1 planning CT and 14 CBCTs. The registration step was performed using the DICOM registration files from the TPS. The code was executed on a machine with an Intel Core Processor (SkyLake, IBRS) and 8 GB RAM.

To ensure the practicality of our pipeline, we measured the time required to process images for a sample patient. The execution times for a patient with 1 planning CT and 14 CBCTs are shown in Table 3. The approximate processing times (including TPS export times) are 45 seconds for a single image and 273 seconds for an entire patient when using the DICOM registration files from the TPS. Approximately 90 additional seconds per image are required to register images without a pre-existing registration file.

4 Discussion

The pipeline described in this work offers a solution for automating and streamlining the workflow of image processing and analysis, reducing the time needed to generate large, high-quality datasets with longitudinal images. Such datasets are essential for training machine learning models and implementing Digital Twins that rely on real-world time-series data. Additionally, since the pipeline begins with the TPS, it is possible for other centres to build similar in-house databases that can be used for research or for quality assurance and improvement purposes.

For research, the ability for different centres to efficiently generate large in-house datasets is especially important given that medical data often cannot be shared with other research groups due to privacy and ethical concerns. These privacy concerns are particularly prominent when dealing with imaging data in the head and neck region where facial recognition

may be possible. With a standardized method for procuring and processing radiotherapy and imaging data, researchers from different centres can more easily implement federated learning and external test validation.

In terms of quality assurance, the resulting prepared data can provide valuable insights and statistics on departmental IGRT and replanning workflows. For instance, statistics on the frequency and timing of image acquisition and treatment delivery could indicate the degree to which a department is following its imaging protocols. Moreover, the integration of outlier detection can reveal how often, and why, low-quality CBCT images are taken. These insights may help identify past issues and adopt preventative measures for the future. Similarly, the replan rate and average replan fraction can be compared with other centres, promoting improved standardization of care.

Our pipeline was designed not only to facilitate the aggregation of retrospective datasets, but also the continued collection of incoming imaging data. With such a pipeline in place, it becomes possible to monitor patients for anatomical changes automatically using image-based metrics such as the area presented in Figure 2.

Once we have completed the training of a varational autoencoder model on our prepared dataset, we aim to use the trained model to map latent space trajectories of HNC patients' longitudinal CBCT images and generate synthetic future CBCT images that will allow the prediction of replanning needs. As an endpoint of our current work, we intend to incorporate both the AI model trained on our dataset and a monitoring system whereby our model is continuously updated with new imaging information for a given patient. Doing so will enable live Digital Twinning for predicting anatomical changes in HNC patients.

One limitation of our work (and TPS) is that raw DICOM files currently need to be manually exported from the TPS. While this task is feasible when building a retrospective dataset as all files for a given patient can be exported at once, it would quickly become cumbersome to do each time a new image is acquired, which is needed for the clinical implementation of a Digital Twin. Thus, future work is needed to implement a fully automatic pipeline.

5 Conclusion

In conclusion, we developed an image processing pipeline that automatically sorts, registers, and preprocesses DICOM images and DICOM-RT files. We demonstrated the feasibility and utility of our pipeline by using it to create a comprehensive dataset of images from 370 HNC patients comprising planning CTs, CBCTs, and radiotherapy data. Our pipeline can facilitate efficient and consistent processing of DICOM files from various sources and modalities, as well as generate useful statistics and metrics about the data. Overall, the datasets produced by the pipeline will allow for the development of AI algorithms trained on longitudinally sequential real-world images.

6 Acknowledgements

This research was supported by the Canada Graduate Scholarships — Doctoral (CGS D) program of the Natural Sciences and Engineering Research Council and by the CREATE Responsible Health and Healthcare Data Science (SDRDS) grant of the Natural Sciences and Engineering Research Council.

References

[1] G. Li, X. Wu, and X. Ma. "Artificial intelligence in radiotherapy". *Seminars in Cancer Biology* 86 (Nov. 2022), pp. 160–171. DOI: 10.1016/j.semcancer.2022.08.005.

[2] S. Siddique and J. C. Chow. "Artificial intelligence in radiotherapy". *Reports of Practical Oncology and Radiotherapy* 25.4 (2020), pp. 656–666. DOI: 10.1016/j.rpor.2020.03.015.

[3] T. Sun, X. He, X. Song, et al. "The Digital Twin in Medicine: A Key to the Future of Healthcare?" *Frontiers in Medicine* 9 (2022).

[4] J. Kazmierska, A. Hope, E. Spezi, et al. "From multisource data to clinical decision aids in radiation oncology: The need for a clinical data science community". English. *Radiotherapy and Oncology* 153 (Dec. 2020), pp. 43–54. DOI: 10.1016/j.radonc.2020.09.054.

[5] K. Clark, B. Vendt, K. Smith, et al. "The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository". en. *Journal of Digital Imaging* 26.6 (Dec. 2013), pp. 1045–1057. DOI: 10.1007/s10278-013-9622-7.

[6] A. Trotti. "Toxicity in head and neck cancer: a review of trends and issues". English. *International Journal of Radiation Oncology, Biology, Physics* 47.1 (Apr. 2000), pp. 1–12. DOI: 10.1016/S0360-3016(99)00558-1.

[7] M. Figen, D. Çolpan Öksüz, E. Duman, et al. "Radiotherapy for Head and Neck Cancer: Evaluation of Triggered Adaptive Replanning in Routine Practice". eng. *Frontiers in Oncology* 10 (2020), p. 579917. DOI: 10.3389/fonc.2020.579917.

[8] B. Lowekamp, D. Chen, L. Ibanez, et al. "The Design of SimpleITK". *Frontiers in Neuroinformatics* 7 (2013).

[9] G. Bradski. "The OpenCV Library". *Dr. Dobb's Journal of Software Tools* (2000).

Demonstration of current best practice for prognostic model validation using the Manchester score

Charlie Cunniffe[1], Gareth Price[1], Matthew Sperrin[2] and Fiona Blackhall[1]

[1]Division of Cancer Sciences, University of Manchester, Manchester, United Kingdom
[2]Division of Informatics Imaging and Data Sciences, University of Manchester, Manchester, United Kingdom

Abstract There is a belief in medical research that predictive and prognostic models could aid with difficult decisions in the clinic. The rise of electronic health records makes it easier to acquire data and results in datasets including patients under-represented in trials and, hence, more difficult to make decisions for. Many models have been developed for lack of thorough validation during development and in external target populations. Current internal and external validation best practices are demonstrated using the Manchester score. This early small-cell lung cancer (SCLC) prognostic model saw clinical use despite not undergoing validation in the 36 years since its development. The model continues to separate patients by predicted survival, and following updates and validation presented here, it can be used to make 6 and 12-month survival predictions for individual patients with SCLC.

1 Introduction

There are acknowledged gaps in clinical knowledge where level 1 trial evidence is unavailable to support decision-making for many patients with SCLC. This leads to a lack of clear guidance on treating and explaining the decision-making process to the patients [1–3]. Predictive models trained on routinely acquired data sets capturing the real-world presentation of treatment and clinical outcomes of previous patients have been proposed as potential solutions to help close this gap as part of clinical decision support systems [4]. However, despite being around for more than 30 years, very few are used clinically. Multiple studies have reported guidelines highlighting issues with development and reporting that further the lack of clinical application [4–9]. Models may be developed using substandard methodology, using factors not regularly measured in the clinic; the question being answered may not be clinically relevant, and the model itself may be too time-consuming or difficult to use [5,10]. Perhaps one of the most important reasons that models do not make it into the clinics is the lack of rigorous external validation [6]. This involves assessing whether a model achieves its intended purpose, often by checking its predictive accuracy. Associated with this are challenging questions around the frequency with which models should be updated or recalibrated over time to maintain relevance as new treatments are developed or populations change [11]. Here, the Manchester score, a prognostic model developed in 1987, is used to demonstrate current validation best practices [12]. The Manchester score aimed to group patients with SCLC by survival to aid with treatment decisions; suggested examples were to provide less intensive treatments to those with poor prognosis to improve their quality of life or help choose patients eligible for clinical trials.

2 Materials and Methods

Cerny et al. began with data from 407 patients with SCLC, taken mainly from previous chemotherapy clinical trials, with the addition of some patients of poorer performance status and more severe disease [12]. A set of 61 pre-treatment variables was chosen by comparing prognostic factors discussed in the literature to the available data. Univariable analysis was used to propose candidate predictors for forward stepwise selection into a multivariable Cox proportional hazards model. The result was a linear predictor shown in Table 1, using six variables (extensive stage, Karnofsky performance status (KPS), serum lactate dehydrogenase, serum alkaline phosphatase, serum bicarbonate and serum sodium levels). An associated nomogram (the 'Manchester Score') based on the Cox model was developed by dichotomising each variable at the upper or lower normal range and assigning points. Good, bad and intermediate prognostic groups were assigned based on the patient's total number of points. The original publication did not provide survival estimates for the Manchester score risk groups nor a baseline hazard to extract them from the linear predictor.

Variable	Manchester score threshold	Linear predictor coefficient
Stage	Extensive disease	+0.63 (if extensive)
KPS	<60	-0.02*KPS
LDH	>450 U/l	+1.17*log10(LDH)
AlkalinePhosphatase	>165 U/l	+0.69*log10(Alk.Phos)
Sodium	<132 mmol/l	-8.43*log10(Sodium)
Bicarbonate	<24 mmol/l	-0.047*Bicarbonate

Table 1. shows the thresholds for using the Manchester score nomogram (add a point for each criterion met) and the equation to calculate the original linear predictor.

This validation used routinely collected structured observational data from over 2000 patients with SCLC seen at a single centre between 2013 and 2022 inclusive. Patients were included in this study if they had a confirmed SCLC diagnosis and had relevant haematology tests. The study included 1783 patients with small cell lung cancer, including those underrepresented in clinical trials. For this study, extensive-stage disease was taken as equivalent to the presence of metastases, and an average of literature conversions from the Eastern Cooperative Oncology Group (ECOG) score to KPS was used [13–16]. Blood tests had to be taken within 31 days prior to the patient's initial consultation.

The validation methodology described here follows the TRIPOD regulations, with reporting guidelines from PROBAST [5,6]. Missing data was handled using multiple imputation to avoid bias from complete case analysis and to maximise the study size [17]. Thirty imputed datasets were created using predictive mean matching, and all results were calculated on each imputed data set individually before being pooled using Rubin's rules [18]. Cerny et al. provided both the nomogram and its underlying linear predictor, which were calculated for each patient. Kaplan-Meier curves were plotted, grouping patients by good, bad, and intermediate prognoses with the same criteria used in the development paper. The score for each patient in the Kaplan-Meier curves was the median across the imputed datasets. The modern survival curves overlayed with the originals to compare survival changes between groups since development. External validation relies on two primary performance metrics: discrimination (the ability to order patients by severity of outcome correctly) and calibration (the ratio of observed to expected outcome probability)[19]. With a 95% confidence interval, discrimination was calculated for both scores using two methods: Harrel's C-index (for time to event prediction) and the area under the curve (AUC) of receiver operating characteristic curve (ROCs) plotted at 6, 12 and 24 months.

Calibration could not be calculated as the development paper failed to report survival estimates for the nomogram groups nor a baseline hazard to accompany the linear predictor. The cumulative baseline hazard for the linear predictor (as presented in the development paper with no rescaling) was calculated for 6, 12, and 24 months to improve the clinical utility and allow for survival probability estimates. These time points were chosen due to the poor median survival time of patients with SCLC, providing probabilities for short and long-term survival. The original paper claimed that the good prognostic group contained all patients surviving beyond 24 months, and the bad prognostic group contained no patients surviving beyond 12 months. Survival estimates at each time point were calculated with optimism adjustment using bootstrap resampling [4,20]. Per imputed dataset, 100 bootstrap datasets were used. Discrimination was not calculated for the internal validation of the updated model as the linear predictor remains the same, yielding the same results.

3 Results

The gender split in the modern cohort is now skewed toward females but is close to 50/50, and the average age of patients has increased by 10 years. An increase in the number of patients with extensive disease is found. There is also a difference in the number of high-performance status patients. The number of patients with KPS>60 in the modern cohort has decreased by ~20%. Overall, the proportion of patients in each Manchester grouping is roughly equivalent. The only notable discrepancy between

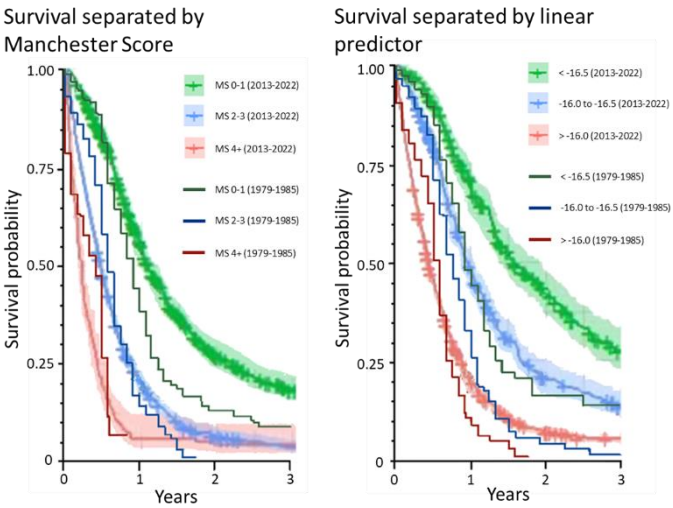


Figure 1, Overlayed Kaplan-Meier survival curves seen in the modern Manchester data set vs those presented in the 1987 development data set, grouped by Manchester Score and linear predictor (as done by Cerny et al.) [12].

cohorts is a 10% decrease in the good prognostic group and a matching 10% increase in the missing data group in the modern cohort compared to the original cohort. The Kaplan-Meier Survival curves display clear separation of the three groups when split by the Manchester score and the linear predictor. All the groups show long-term survivors (not considering any further cancers or relapses). The percentage of patients who survived to one, three and five years was 34%, 7.7% and 3.7%, respectively, within the entire modern cohort. The whole cohort median survival was 8 months. Figure 1 shows the Kaplan-Meier curves up to three years, as seen in the original paper, superimposed with the curves from the current analysis. The separation between the three groups has increased since 1987 for both prognostic scores, and the proportion of long-term survivors has increased. The median survival for the bad and intermediate prognostic groups is relatively similar for both curves; however, for the good prognostic groups, it has increased by approximately two and six months for the Manchester Score and linear predictor, respectively. The Manchester score has an AUC of 0.75, and the linear predictor has an AUC of 0.79 at 6 months. Both decrease by around 0.05 when calculated at 24 months. Harrell's C-index was 0.68 (with 95% CI [0.685, 0.689]) for the Manchester score and 0.70 (with 95% CI [0.69, 0.71]). The baseline hazard and optimism-adjusted calibration values can be seen in Table 2, along with the corresponding 95% confidence intervals. The accuracy of predictions decreases over time as 6 months has the best calibration at 1.001, and 24 months is the worst with 0.795.

4 Discussion

Various criteria exist for evaluating the level of bias in model development (PROBAST) along with the quality of reporting and validation (TRIPOD) [5,6]. The Manchester score development achieved a low risk of bias and poor reporting of model parameters. Here, an external model

validation procedure adheres to development and validation guidelines. This was achieved by using a large unselected observational dataset. Changes in population survival were directly compared by overlaying Kaplan-Meier survival curves. Discrimination was assessed using two methods. However, it was not possible to calibrate the model with solely the information provided in the original publication. Cerny et al. stratified patients by survival but provided no meaning to the groups. A survival estimate was assigned to each prognostic group, and a baseline hazard was calculated to provide individual survival probabilities based on the 1987 linear predictor. Internal validation of the survival estimates was completed using bootstrap resampling as recommended in the literature [4,17].

Time (months)	Cumulative baseline hazard	Baseline hazard 95% confidence interval	Optimism-adjusted calibration	Calibration 95% confidence interval
6	2815117	[2780317, 2837678]	1.001	[0.998, 1.005]
12	7123175	[7033155, 7198284]	0.971	[0.964, 0.978]
24	13580956	[13361570, 13756204]	0.795	[0.784, 0.811]

Table 2. provides baseline hazard estimates, and the calibration values at 6 months, 1 and 2 years with their 95% confidence intervals.

Despite being mentioned in literature discussing treatment methods for SCLC and referenced as an example in further model development papers, the Manchester score has never been validated using performance metrics[21–23]. Three previous analyses have used the Manchester score to stratify patients for Kaplan-Meier survival curves (none have the primary aim of validating the Manchester score) [22–24]. All concluded the score remained significantly associated with survival despite Kaplan-Meier curves presented by El-Hewl et al. showing no differentiation between the intermediate and bad prognostic groups. No papers provide interpretations for the survival of each group, but it can be seen from the graphs that the median survival of each group improves as the studies become more recent. However, the survival seen in the development paper aligns more closely with that of this study rather than the 1997 analysis by Maestu et al. The differences in survival between the 1987 and the modern cohort are likely reduced due to the strict selection criteria used by the trials that provided patients to the development cohort; this results in the development paper presenting disproportionately longer survival times. Survival in the modern cohort agrees with what has been reported in the literature [25–27]. Following the external validation, values for the median survival of each prognostic group or the Manchester score are reported. For the linear predictor model, updating and calibration have provided survival estimates at clinically relevant time points for patients on an individual basis.

During the analysis, stage and performance status conversions had to be utilised as standard measurement practice changed, limiting the accuracy of the validation. However, the current analysis uses fewer selection criteria than in model development, better reflecting the target population. Similarly, multiple imputation allowed all patient data to be utilised, maximising the cohort size. This is dependent on the assumption that data is missing at random. Survival trends are measured for each Manchester prognostic group and the whole unselected cohort. The survival of each prognostic group in the modern cohort has been quantitatively described in terms of overall median survival. This indicates the prognosis in each risk group and the relationship between them. Reporting the coefficients of a predictive model is necessary for future use of the model and validation. However, it is often subject to the Table 2 fallacy; the effects seen in the linear predictor are not causal [28]. A barrier to clinical use in the original model is that baseline hazard is not provided. This means survival estimates cannot be calculated, and the score cannot be externally calibrated in the modern cohort. The extraction of the baseline hazard in the modern cohort allows for survival estimates to be made in the future. Before clinical use, validation in the intended centre and population would be required. However, assessment of the optimism-adjusted calibration using bootstrap resampling is an example of current best practice for internal validation, providing a stable and approximately unbiased estimate of model performance in the represented population. It would be expected that the calibration remains close to one for all time points due to being evaluated in the same data it was developed in; however, the model's predictive ability deteriorates at later time points. This is likely due to a class imbalance at 24 months, as the median survival in the data set is 8 months.

In oncology, there are two distinct meanings for prediction and prognosis. Prediction focuses on assessing the impact of treatment on the outcome, whereas prognosis is aimed at evaluating the outcome in the absence of treatment (null intervention). In practice, assessing outcomes without intervention is difficult since patients typically receive the best available standard of care, usually involving some form of intervention, whether curative or palliative. As a result, prognostic models such as the Manchester Score are often developed without considering treatment factors, even though treatment affects the outcome. To influence treatment decisions with the aim of reducing toxicities for those with poor survival (stated to be the primary suggested use of the Manchester score), a causal analysis looking at the average causal effect of prescribed treatment on toxicities should be conducted. Alternatively, looking at the average causal effect of assessment of Manchester score use vs not use in treatment prescription on toxicity levels in patients with SCLC would provide a clear evaluation of the

impact. Following current validation guidelines, the Manchester score is fit to be used in the clinic to assess prognosis in populations similar to those shown here.

5 Conclusion

Prediction and prognostic models rarely make it into the clinic despite ample model development and promise to aid clinical decision-making and patient communication. One reason for not being used is a lack of validation and poor validation practices in the relevant target populations. To combat these issues, many guidelines have been published on conducting and reporting the development and validation of predictive models. Here, current validation best practice is demonstrated using an early predictive model for survival of patients with SCLC. The Manchester Score is a simple nomogram developed using (now) outdated methods, and information crucial to thorough validation was not reported. External validation was completed, followed by model updating and internal validation, using a large unselected observational dataset from a single centre spanning 10 years. The Manchester score successfully groups patients by overall survival time and has been expanded to make individual predictions at 6 and 12 months post-diagnosis. Factors such as the linear predictor and baseline hazard at clinically relevant time points are reported, allowing the model to be used and validated.

References

[1] K.J. Lang, S. Lidder, Under-representation of the elderly in cancer clinical trials, Br. J. Hosp. Med. 71 (2010) 678–681. https://doi.org/10.12968/hmed.2010.71.12.678.

[2] A. Smart, E. Harrison, The under-representation of minority ethnic groups in UK medical research, Ethn. Health 22 (2017) 65–82. https://doi.org/10.1080/13557858.2016.1182126.

[3] K. Sharrocks, J. Spicer, D.R. Camidge, S. Papa, The impact of socioeconomic status on access to cancer clinical trials, Br. J. Cancer 111 (2014) 1684–1687. https://doi.org/10.1038/bjc.2014.108.

[4] G.S. Collins, P. Dhiman, J. Ma, M.M. Schlussel, L. Archer, B. Van Calster, F.E. Harrell, G.P. Martin, K.G.M. Moons, M. Van Smeden, M. Sperrin, G.S. Bullock, R.D. Riley, Evaluation of clinical prediction models (part 1): from development to external validation, BMJ (2024) e074819. https://doi.org/10.1136/bmj-2023-074819.

[5] R.F. Wolff, K.G.M. Moons, R.D. Riley, P.F. Whiting, M. Westwood, G.S. Collins, J.B. Reitsma, J. Kleijnen, S. Mallett, for the PROBAST Group†, PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies, Ann. Intern. Med. 170 (2019) 51. https://doi.org/10.7326/M18-1376.

[6] K.G.M. Moons, D.G. Altman, J.B. Reitsma, J.P.A. Ioannidis, P. Macaskill, E.W. Steyerberg, A.J. Vickers, D.F. Ransohoff, G.S. Collins, Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and Elaboration, Ann. Intern. Med. 162 (2015) W1–W73. https://doi.org/10.7326/M14-0698.

[7] K.G.M. Moons, J.A.H. de Groot, W. Bouwmeester, Y. Vergouwe, S. Mallett, D.G. Altman, J.B. Reitsma, G.S. Collins, Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modelling Studies: The CHARMS Checklist, PLoS Med. 11 (2014) e1001744. https://doi.org/10.1371/journal.pmed.1001744.

[8] U. of Bristol, QUADAS-2, (n.d.). https://www.bristol.ac.uk/population-health-sciences/projects/quadas/quadas-2/ (accessed January 10, 2023).

[9] J.A. Hayden, D.A. van der Windt, J.L. Cartwright, P. Côté, C. Bombardier, Assessing Bias in Studies of Prognostic Factors, Ann. Intern. Med. 158 (2013) 280. https://doi.org/10.7326/0003-4819-158-4-201302190-00009.

[10] S. Mallett, P. Royston, S. Dutton, R. Waters, D.G. Altman, Reporting methods in studies developing prognostic models in cancer: a review, BMC Med. 8 (2010) 20. https://doi.org/10.1186/1741-7015-8-20.

[11] B. Van Calster, E.W. Steyerberg, L. Wynants, M. Van Smeden, There is no such thing as a validated prediction model, BMC Med. 21 (2023) 70. https://doi.org/10.1186/s12916-023-02779-w.

[12] T. Cerny, H. Anderson, V. Bramwell, N. Thatcher, V. Blair, Pretreatment prognostic factors and scoring system in 407 small-cell lung cancer patients, Int. J. Cancer 39 (1987) 146–149. https://doi.org/10.1002/ijc.2910390204.

[13] ECOG Performance Status Scale, ECOG-ACRIN Cancer Res. Group (n.d.). https://ecog-acrin.org/resources/ecog-performance-status/ (accessed February 6, 2023).

[14] K.T. Prasad, H. Kaur, V. Muthu, A.N. Aggarwal, D. Behera, N. Singh, Interconversion of two commonly used performance tools: An analysis of 5844 paired assessments in 1501 lung cancer patients, World J. Clin. Oncol. 9 (2018) 140–147. https://doi.org/10.5306/wjco.v9.i7.140.

[15] I. De Kock, M. Mirhosseini, F. Lau, V. Thai, M. Downing, H. Quan, M. Lesperance, J. Yang, Conversion of Karnofsky Performance Status (Kps) and Eastern Cooperative Oncology Group Performance Status (Ecog) to Palliative Performance Scale (Pps), and the interchangeability of Pps and Kps in Prognostic Tools, J. Palliat. Care 29 (2013) 163–169. https://doi.org/10.1177/082585971302900305.

[16] G. Buccheri, D. Ferrigno, M. Tamburini, Karnofsky and ECOG performance status scoring in lung cancer: A prospective, longitudinal study of 536 patients from a single institution, Eur. J. Cancer 32 (1996) 1135–1141. https://doi.org/10.1016/0959-8049(95)00664-8.

[17] M. Sperrin, G.P. Martin, R. Sisk, N. Peek, Missing data should be handled differently for prediction than for description or causal explanation, J. Clin. Epidemiol. 125 (2020) 183–187. https://doi.org/10.1016/j.jclinepi.2020.03.028.

[18] A. Marshall, D.G. Altman, R.L. Holder, P. Royston, Combining estimates of interest in prognostic modelling studies after multiple imputation: current practice and guidelines, BMC Med. Res. Methodol. 9 (2009) 57. https://doi.org/10.1186/1471-2288-9-57.

[19] E.W. Steyerberg, Y. Vergouwe, Towards better clinical prediction models: seven steps for development and an ABCD for validation, Eur. Heart J. 35 (2014) 1925–1931. https://doi.org/10.1093/eurheartj/ehu207.

[20] E.W. Steyerberg, F.E. Harrell, G.J.J.M. Borsboom, M.J.C. Eijkemans, Y. Vergouwe, J.D.F. Habbema, Internal validation of predictive models, J. Clin. Epidemiol. 54 (2001) 774–781. https://doi.org/10.1016/S0895-4356(01)00341-9.

[21] T. Blakely, J. Lynch, K. Simons, R. Bentley, S. Rose, Reflection on modern methods: when worlds collide—prediction, machine learning and causal inference, Int. J. Epidemiol. 49 (2021) 2058–2064. https://doi.org/10.1093/ije/dyz132.

[22] R. Hagmann, A. Zippelius, S.I. Rothschild, Validation of Pretreatment Prognostic Factors and Prognostic Staging Systems for Small Cell Lung Cancer in a Real-World Data Set, Cancers 14 (2022) 2625. https://doi.org/10.3390/cancers14112625.

[23] I. Maestu, M. Pastor, J. Gómez-Codina, J. Aparicio, A. Oltra, C. Herranz, J. Montalar, B. Munárriz, G. Reynés, Pretreatment prognostic factors for survival in small-cell lung cancer: A new prognostic index and validation of three known prognostic indices on 341 patients, Ann. Oncol. 8 (1997) 547–553. https://doi.org/10.1023/A:1008212826956.

[24] L.M. El-Helw, T.K. Rogers, M.Q.F. Hatton, Small-Cell Lung Cancer: 8 Years Experience of a Single Multidisciplinary Team, J. Oncol. 2008 (2008) 1–6. https://doi.org/10.1155/2008/150760.

[25] M. Araghi, M. Fidler-Benaoudia, M. Arnold, M. Rutherford, A. Bardot, J. Ferlay, O. Bucher, P. De, G. Engholm, A. Gavin, S. Kozie, A. Little, B. Møller, N. St Jacques, H. Tervonen, P. Walsh, R. Woods, D.L. O'Connell, D. Baldwin, M. Elwood, S. Siesling, F. Bray, I. Soerjomataram, ICBP SURVMARK-2 Local Leads, ICBP SURVMARK-2 Academic Reference Group, International differences in lung cancer survival by sex, histological type and stage at diagnosis: an ICBP SURVMARK-2 Study, Thorax 77 (2022) 378–390. https://doi.org/10.1136/thoraxjnl-2020-216555.

[26] R.J. Stephens, D.H. Johnson, Treatment and Outcomes for Elderly Patients with Small Cell Lung Cancer:, Drugs Aging 17 (2000) 229–247. https://doi.org/10.2165/00002512-200017030-00006.

[27] S. Walters, C. Maringe, M.P. Coleman, M.D. Peake, J. Butler, N. Young, S. Bergström, L. Hanna, E. Jakobsen, K. Kölbeck, S. Sundstrøm, G. Engholm, A. Gavin, M.L. Gjerstorff, J. Hatcher, T.B. Johannesen, K.M. Linklater, C.E. McGahan, J. Steward, E. Tracey, D. Turner, M.A. Richards, B. Rachet, the ICBP Module 1 Working Group, Lung cancer survival and stage at diagnosis in Australia, Canada, Denmark, Norway, Sweden and the UK: a population-based study, 2004–2007, Thorax 68 (2013) 551–564. https://doi.org/10.1136/thoraxjnl-2012-202297.

[28] D. Westreich, S. Greenland, The Table 2 Fallacy: Presenting and Interpreting Confounder and Modifier Coefficients, Am. J. Epidemiol. 177 (2013) 292–298. https://doi.org/10.1093/aje/kws412.

A Non-invansive and Self-managed Microscope for Predicting Radiation-Induced Acute Toxicities in Cancer Patients

Alessandra Catalano[1], Sophie Veronique Materne[1], Fabio Badenchini[2], Eliana La Rocca[3], Eliana Gioscio[1], Francesco Pisani[1], Luca Possenti[1], Barbara Avuzzi[5], Barbara Noris Chiorda[5], Carlotta Giandini[5], Riccardo Roy Colciago[5], Maria Carmen De Sanctis[5], Tommaso Giandini[4], Alessandro Cicchetti[1], and Tiziana Rancati[1]

[1]Data Science Unit, Fondazione IRCCS Istituto Nazionale dei Tumori, Milan, Italy
[2]Unit of GenitoUrinary Oncology, Fondazione IRCCS Istituto Nazionale dei Tumori, Milan, Italy
[3]Department of Radiation Oncology, University of Verona Hospital Trust, Verona, Italy
[4]Division of Medical Physics, Fondazione IRCCS Istituto Nazionale dei Tumori, Milan, Italy
[5]Department of Radiation Oncology, Fondazione IRCCS Istituto Nazionale dei Tumori, Milan, Italy

Abstract Microcirculation is the terminal vascular network of the systemic circulation. It has a pivotal role in the cardiovascular system. Damage to microvessels caused by radiotherapy (RT) contributes to toxicity development in a non-clarified manner. This work aims to investigate quantitatively the role of microcirculation in predicting acute toxicities after breast and prostate cancer RT. We enrolled 247 patients at a single centre for this study RT. We assessed the single-patient baseline microvasculature health status (MVHS) before RT using a microscope coupled to the GlycoCheck™ software. The system records videos (in the sublingual region) showing the live movement of red blood cells in the microvessels, analyses the microvasculature and computes the MVHS value for the overall health of microcirculation. We fitted a logistic model for the association between MVHS and acute toxicities, and we found a quantitative relationship between microvasculature health status and radio-sensitivity for patients both breast and prostate cancer patients.
The information obtained from the sublingual microscope could help personalise predictive models for toxicity and tailor them to the functional status of each patient.

1 Introduction

Microcirculation is an intricate network of arterioles, capillaries and venules that supply and drain blood from every tissue and organ in the body. The particular architecture of these microvessels varies depending on the specific structure of the vital organ they serve. An essential aspect of microcirculation is its pivotal role as the functional core of the cardiovascular system, facilitating the exchange of oxygen, carbon dioxide, nutrients, hormones, water and drugs, and as a causal factor in developing some pathologies.
Radiotherapy (RT) can contribute to microcirculation dysfunction. Indeed, the damage to the microcirculation causes endothelial cell dysfunction, increased vessel permeability and alterations in vasoconstriction and vasodilation, impacting oxygen and nutrient diffusion within the tumour and, consequently, the treatment efficacy. Moreover, RT also damages the microvessels of healthy organs, interfering with the damage recovery and toxicity manifestation process. Several trials have indicated an increased risk of normal tissue complications in patients with pre-existing medical conditions (hypertension, diabetes, use of cholesterol-lowering drugs, use of drugs for cardiac morbidity, obesity) or specific habits (smoking, alcohol abuse, low physical activity) that negatively impact the stability of the vascular system [1–5].
This work aimed to investigate quantitatively the role of microcirculation in predicting acute toxicities after breast (BCa) and prostate (PCa) cancer RT. Current RT techniques and plan optimisation for these two districts allow dose distributions highly conformed to the clinical targets, with small proportions of normal tissue receiving relevant doses. For these reasons, dose parameters are weakly associated with side effects, and the search for patient-specific features highlighting exceptional radiosensivity in still an active field of research. We included a quantitative measure of the single-patient microvascular health status (as measured by GlycoCheck™) in a predictive model for acute toxicities.

2 Materials and Methods

2.1 The study cohort

This prospective and observational study enrolled 247 patients between February 2021 and November 2023. The study cohort consisted of 143 BCa and 104 PCa patients. Women were treated with chemo RT in 25% of cases. RT was applied with a dose boost in 50% of treatments (33% as a sequential boost and 17% as a concomitant boost). Dose per fraction to PTV was 2.6-2.67 Gy in 85% of patients, 5.5 Gy in 10% of treatments and conventional 2 Gy for the remaining 5% of cases. Regarding prostate patients, 25% of them underwent adjuvant RT with 70 Gy in 35 fr. The remaining part was treated with a dose/fr in the 2-2.65 Gy range and a prescribed dose between 65 and 78 Gy.
Before RT, we trained patients to perform the baseline sublingual-microvasculature measurement (details in section 2.2). Clinicians collected the patient's condition at the baseline and at the RT end following CTCAE v4.0. We focused on the worst toxicity symptoms in each cohort, namely grade ≥ 2 erythema for BCa and simultaneous gastrointestinal (GI) and genitourinary (GU) grade ≥ 1 toxicity for PCa.

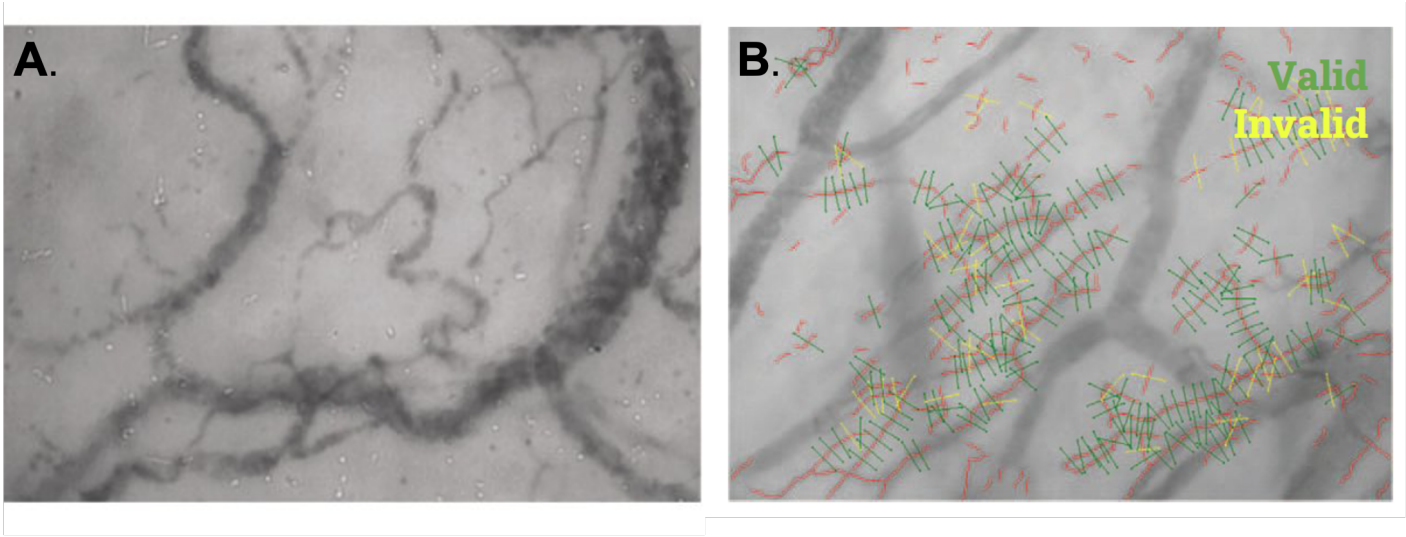


Figure 1: Glycocheck image acquisition. **A.** Frame of the videos recorded by the Glycocheck SDF Camera. **B.** After the image acquisition, the software automatically undergoes several quality check to select the vascular segments with sufficient quality for further analysis. Invalid segments (yellow) are distinguished from the valid vascular segments (green) [6].

2.2 The Instrument: Sublingual Video-Microscopy

The image acquisition of the sublingual microvasculature is performed through the GlycoCheck™ system. It comprises a handheld, non-invasive sidestream dark-field (SDF) camera coupled with a computer with the GlycoCheck software™ (GlycoCheck, Maastricht, the Netherlands). The SDF camera uses green light-emitting stroboscopic diodes (LED) to detect the haemoglobin of passing Red Blood Cells (RBC). The system acquires images in the sublingual region and extrapolates parameters representative of systemic microvascular health. The Glycocheck Software facilitates automatic image acquisition, identifying micro-vessels (with a thickness below 30 μm) during the process and analysing vascular segments along the length of the identified vessels. A single bench of measures consists of 40 frames encompassing 300 vascular segments. Next, the patient relocates the camera to a different position to record an additional acquisition. In this way, the spatial heterogeneity of the sublingual microcirculation is ensured. After acquiring images, the software analyses the vascular segments that satisfy predefined image quality criteria.
The software computes functional and morphological parameters, such as vessel diameter distribution, RBC velocity (VRBC), capillary density, capillary blood volume (CBV) and perfused boundary region (PBR) (see the Additional Material for equations).

2.3 Microvasculature Health Score

The crucial systemic measure obtained by the Glycocheck Software is the Microvascular Health Score (MVHS), which gives information about the overall health of the microvascular system. It is a synthetic and global score ranging from 0 to 15.
The MVHS computation uses all the functional and morphological parameters obtained during the recordings, such as the CBV, the capillary blood volume recruitment (CR) and the PBR [7]. More in detail, the software computes the ratio between the dynamic capillary blood volume CBV_{dyn} and PBR_{4-25} flow corrected (PBR_{4-25}-FL), so

$$MVHS_{dyn} = \frac{CBV_{dyn}}{PBR_{4-25} - FL}. \quad (1)$$

This parameter provides an overall quantitative evaluation of microvascular health. The higher the MVHS, the better the microvascular condition.

2.4 MVHS Imputation and Statistical analysis

The analysis focused on the association of MVHS with acute toxicity. MVHS summarises all the microcirculation parameters, and is automatically normalised for all patients.
The software's computation of MVHS needs a prolonged video acquisition, and it is not calculated for less compliant subjects. Consequently, we developed an imputation model to extrapolate the MVHS for patients with missing information. We based the imputation on all the available parameters. We trained the model on the entire population and tested the results on patients with the GlycoCheck™ computed MVHS. The training process was iterated until the average error between the observed and the imputed values was below 0.05.

3 Results

Forty-one/143 (28.7%) BCa patients developed grade ≥ 2 skin erythema. Sixty-three/104 (60.6%) PCa patients experienced grade ≥ 1 in both GI and GU domains.
The GlycoCheck exam was self-performed by patients. MVHS, requiring a longer video registration, was calculated

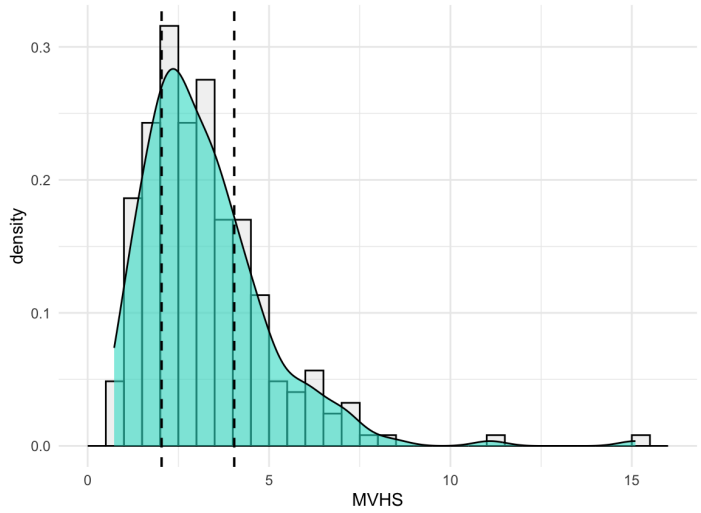


Figure 2: Density plot for the MVHS. The dashed lines represent the 1st and the 3rd quartile of MVHS distribution.

for 60% of patients, while it was obtained by multiple imputation models for the remaining part. After 250 iterations, the model reached good stability with an average error between the predicted and the observed value of MVHS equals -0.04, making the imputed values reliable.
The median MVHS was 2.87, and the 1st and 3rd quartiles were 2.04 and 4.04, respectively (see Figure 2 for the whole distribution). Stratifying for cancer type, the median MVHS was 3.06 in the BCa and 2.75 in PCa.
The logistic model coefficients were $\beta_0 = 0.58$ (constant), $\beta_1 = -0.29$ (slope for MVHS, $OR = 0.81$ for one point increase in MVHS; $p - value = 0.002$). As expected, a higher MVHS (i.e., a healthy microcirculation) protects from side effects. MVHS box plots for BCa and PCa patients stratified for acute toxicity are depicted in Figure 3.

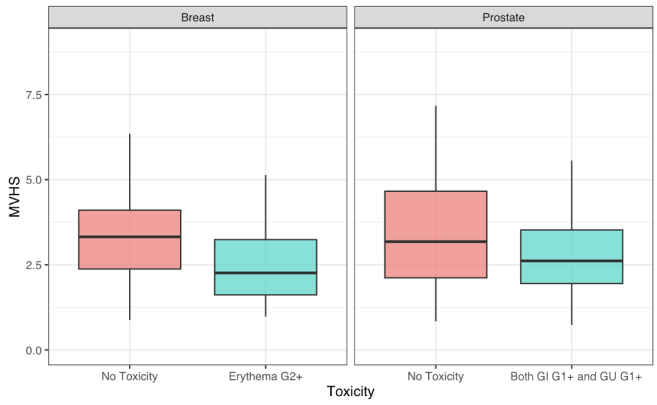


Figure 3: Distribution of MVHS between the two cohorts of patients and separated for toxicity (both $p - value < 0.0001$)

4 Discussion

We tested and quantitatively assessed the MVHS through a non-invasive tool and established the relationship between the microvasculature condition and acute radiation-induced toxicity in a cohort of BCa and Pca patients. The model confirmed an association between altered microvasculature and the damage repair mechanisms following RT, independently from the treatment region and patient sex. Indeed, MVHS distributions for breast and prostate populations were similar (p-value from t-test = 0.36), showing a negligible impact due to sex. In contrast, MVHS distributions for patients experiencing toxicity were significantly different in both the cancer cohorts (p-value <0.0001 in both cases).
From a quantitative point of view, a reduction of 1 value in the MVHS scale corresponded to an (averaged) increased risk of 23.5% of developing radiation-induced symptoms. We can define three risk classes by stratifying patients through the clinical definition of MVHS (see points of Figure 4). The low-risk group (patients having MVHS higher than 4) in which the probability of developing toxicity is 28%; the second class, representing the average patient (MVHS between 2 and 4), is at moderate risk with a toxicity rate equal to 41%. The last class (MVHS less than 2) with compromised microcirculation had a toxicity rate of 59%.
A new analysis is ongoing, including a cohort of 75 head-neck cancer patients and investigating dysphagia, xerostomia and mucositis. Considering the heterogeneity of RT treatments (and dose distributions in organ at risk) in these patients, we will include the MVHS in dose-response models for this cohort.

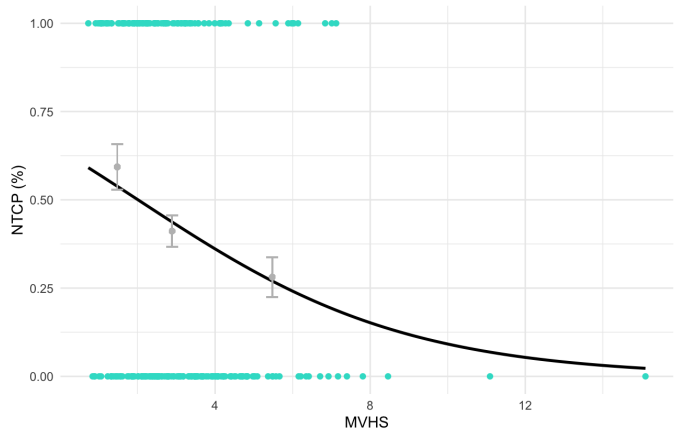


Figure 4: Normal Tissue Complication Probability as a function of MVHS. The three points show the three risk classes (low, moderate and high risk).

5 Conclusion

The study highlights the possibility of implementing a non-invasive instrument in clinical practice to gather information about microcirculation status and define the toxicity risk in different tumour districts using a patient-specific MVHS. The MVHS model predicts acute toxicity, proving a quantitative relationship between microvasculature status and radio-sensitivity in an organ-agnostic way.

The systemic functional information derived by the sublingual microscope could boost the personalisation of predictive models and tailor them to the single-patient functional status.

6 Acknowledgments

The study was supported by AIRC Investigator Grant no. 21479.

6.1 Additional material: formulas

The GlycoCheck software computes two categories of parameters: those specific to each vessel diameter class and those calculated across all diameter classes ranging from 4 to 25 μm, yielding a singular value for each subject's measurement.
Within the first category are three parameters. First, the capillary density, the RBC velocity (VRBC), and the perfused boundary region (PBR). In the second category, the GlycoCheck Software computes the microvascular blood flow, the total valid perfused microvascular density, and the parameters essential for deriving the microvascular health scores. They are the relative capillary blood volume absolute, relative, static and dynamic. In formulas:

$$CBV_{abs} = \ capillary\ density * \pi r^2$$

$$CBV_{rel} = \frac{VRBC(D \geq 10\mu m)}{VRBC(D \leq 7\mu m)}$$

$$CBV_{stat} = CBV_{abs} * CBV_{rel}$$

$$CR = 1 - slope(VRBC(D \leq 7\mu m), VRBC(D \geq 10\mu m))$$

$$CBV_{dyn} = CBV_{stat} * (1 + CR).$$

The static PBR (PBR_{4-25}) is the average PBR calculated across all diameter classes, while the dynamic PBR (PBR_{4-25} flow corrected) is the PBR_{4-25} with $VRBC(D \geq 10\mu m)$ set to 0. This minimises the possible flow-dependent variability in the PBR estimation, as the RBC penetration into the luminal glycocalyx surface can be velocity-dependent.

References

[1] M. D. Santis, F. Bonfantini, F. D. Salvo, et al. "Factors influencing acute and late toxicity in the era of adjuvant hypofractionated breast radiotherapy." *The Breast* 29 (2016), pp. 90–95. DOI: 10.1016/j.breast.2016.07.013.

[2] F. Palorini, T. Rancati, C. Cozzarini, et al. "Multi-variable models of large International Prostate Symptom Score worsening at the end of therapy in prostate cancer radiotherapy." *Radiother Oncol.* 118 (Jan. 2016), pp. 92–8. DOI: 10.1016/j.radonc.2015.11.036.

[3] D. G, V. den Bergh L, A.-M. A, et al. "The benefits of including clinical factors in rectal normal tissue complication probability modeling after radiotherapy for prostate cancer." *Int J Radiat Oncol Biol Phys.* 82 (Mar. 2012), pp. 1233–42. DOI: 10.1016/j.ijrobp.2011.03.056.

[4] S. Tucker, L. Dong, W. Bosch, et al. "Late rectal toxicity on RTOG 94-06: analysis using a mixture Lyman model." *Int J Radiat Oncol Biol Phys.* 78 (Nov. 2010), pp. 1253–60. DOI: 10.1016/j.ijrobp.2010.01.069.

[5] C. Anna, R. Tiziana, C. Alessandro, et al. "EP-1594: Development of multivariable models for acute toxicities in nasopharyngeal cancer radiotherapy". *Radiotherapy and Oncology* 123 (May 2017), S858–S859. DOI: 10.1016/S0167-8140(17)32029-7.

[6] D. Lee, B. D. M. van den Berg, M. Boels, et al. "Deeper Penetration of Erythrocytes into the Endothelial Glycocalyx Is Associated with Impaired Microvascular Perfusion." *Phys Med Biol* (2014). DOI: https://doi.org/10.1371/journal.pone.0096477.

[7] Rovas, A., Sackarnd, et al. "Identification of novel sublingual parameters to analyze and diagnose microvascular dysfunction in sepsis: the NOSTRADAMUS study." *Crit Care 25* 112 (2021). DOI: https://doi.org/10.1186/s13054-021-03520-w.

Data preparation pipeline for artificial intelligence in Radiation Oncology: Lessons from a large Prospective Imaging trial

Hannah Mary Thomas T[1], Amal Joseph Varghese[1], Manu Mathew[1], Devakumar D[2], Aparna Irodi[3], Joel T[1], Timothy Peace[1], Henry Finlay Godson[1], Rajesh Isiah[1], Simon Pavamani[1], Balu Krishna Sasidharan[1],

Departments of [1]Radiation Oncology, [2]Nuclear Medicine, [3]Radiology, Christian Medical College, Vellore, Tamil Nadu, India

Abstract

The integration of Artificial Intelligence (AI) into clinical practice demands robust data preparation pipelines, especially in Radiation Oncology where challenges in data quality and access persist. This work presents a bottom-up approach for AI implementation in Indian Radiation Oncology, focusing on a prospective observational trial for predicting locoregional recurrence in head and neck cancer. The data preparation pipeline encompasses patient consenting, clinical data collection, image acquisition, annotation, curation, de-identification, and storage. The workflow, established in 2020, has successfully imaged over 1100 patients. The Data preparation involves metadata analysis, automated quality control, and integration with open-source platforms like Orthanc DICOM server and XNAT. While providing practical insights for data preparation for AI research, the work emphasizes the importance of tailored frameworks for specific healthcare contexts, outlining current practices and future considerations in the evolving landscape of AI in Radiation Oncology.

1 Introduction

The development of AI solutions that are reproducible and transferrable to clinical practice require access to large scale data for model training and optimization. Despite Radiation Oncology departments generating large volumes of imaging data and clinical routinely, quality and access to this data for AI research poses significant challenges in practice.

Implementing Artificial Intelligence (AI) in a clinical environment is a complex task. It requires resources, time, technical capacity, and multidisciplinary perspectives. Many AI frameworks outline a set of broad principles and values, but do not provide practical guidance on how these principles can be implemented and trade-offs negotiated. Such enabling factors were yet to be articulated within the context of Radiation Oncology in India, signaling the need for a bottom-up study, and localized models. In this work we report our data preparation pipeline for AI projects that involve imaging and clinical data. It was aimed to integrate seamlessly within the existing clinical workflow.

This data preparation pipeline was created as part of a prospective observational imaging trial that is on-going with the primary endpoint being Radiomics models for predicting locoregional recurrence in locally advanced head and neck cancer.

2 Materials and Methods

We have subdivided the data preparation pipeline for AI projects into multiple tasks that are typically performed in sequence. (See Figure 1).

Patient consenting: This study was approved by the Institutional Review Board. The patients treated in Radiation Oncology for locally advanced head and neck cancer were explained the observational nature of the trial and consent was obtained for their clinical and imaging data to be used for research purposes.

Clinical Data collection and curation: Clinical data are stored in disparate sources. The demographics and clinical investigations are collected and stored in our electronic medical records and managed in our hospital information system. Radiation Oncology related data such as radiation treatment notes, treatment related toxicity, treatment outcomes etc. are collected and stored in the Aria Oncology Information system. Study specific clinical data are curated, pseudonymized and stored in a Redcap database[1].

Image acquisition: Planning CT images were acquired on a Siemens SOMATOM AS. The contrast CT acquisition protocols were standardized for Head and Neck cancer for both machines. This was done after extensive phantom experiments and optimized for field of view, tube current, tube voltage, slice thickness and Image reconstruction kernels ideal for treatment planning and imaging research.

Image annotation: The CT images were annotated using Eclipse™ Treatment Planning System (Varian Medical Systems, Palo Alto, CA). Templates were created for consistent naming convention using the Eclipse Structure Templates. The gross tumor volume (GTV) was manually delineated by junior radiation oncologist and approved by senior oncologists. All contours selected for the radiomics study were reviewed by radiation oncologists with more than 15 years of experience. Repeat blind segmentation was performed for 20% of the cases to account for inter-observer variability.

Image data curation: The CT and tumour contours were moved from the Oncology Information System to a lightweight, enterprise-grade Orthanc DICOM server. The DICOM server was deployed in a Docker container in a dedicated Research Workstation.
The DICOM metadata of the stored medical images were collected. This was automated by creating a bespoke python module to analyze the DICOM metadata, specifically to verify the images for Study Description, Modality Specific image parameters (e.g. CT) and its associated series (DICOM segmentation), perform data quality control, ensure image integrity, query the images, and store the file paths in a JSON format. We also use script to check for any inconsistently labelled RT structure sets and rename them before further analysis.

Image de-identification: The curated images were pseudo anonymized by using MIRC Clinical Trial Processor (CTP), an open-source java-based application specifically designed to remove protected health information from DICOM objects[2].
All protected health information was replaced with pseudonyms or unique identifiers and the master table was well-secured and separately stored with access provided only to the researchers involved.

Image storage and Backup: The images and associated DICOM objects from TPS were backed up via S3 bucket on a private cloud setup on the Institutional research cluster and automated using customized python scripts.

The pseudonymized images were sent directly from CTP to XNAT, which is an open-source imaging informatics platform[3]. XNAT was set up in a Docker container and it allowed data to be stored for separate studies. It also provided programmatic script access to perform AI related workflows using the data from XNAT and route the analyzed data back to the database.

3 Results

Figure 1 shows the data preparation pipeline. The image acquisition, image annotation and basic storage to research environment was set up in 2020. It was used to prospectively image about 1100 patients till date treated for head and neck cancer in the department of Radiation Oncology. Other components in the workflow have been included and audited asynchronously.

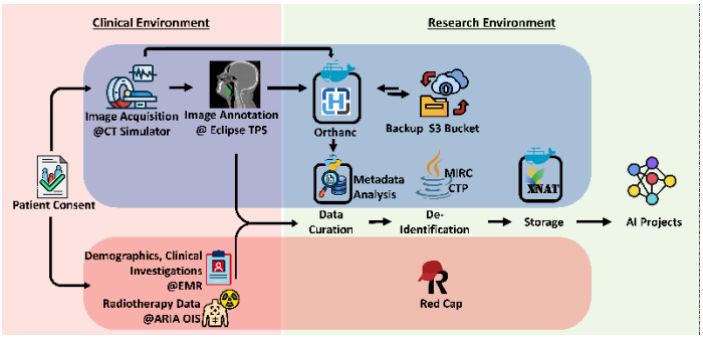


Figure 1: Data preparation pipeline for AI research in Radiation Oncology department

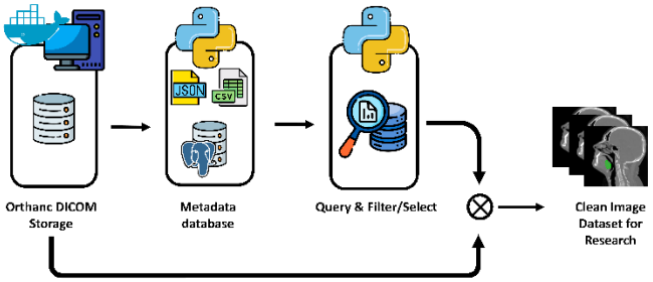


Figure 2: Image data curation workflow

Figure 2 describes the Image Data curation workflow. The Metadata was collected for images received on the Orthanc Storage. The metadata was analyzed to include only patients related to the study in question. The DICOM metadata tags included for analysis used Personal Health identifier information (Hosp ID, Patient Name etc.) Study (Study Instance UID, Date of study etc.) Series related information (Series description, Modality etc.) SOP Instance UID unique for each File and the corresponding File Path. After the completion of metadata analysis, the cleaned data was moved forward for de-identification and storage.

4 Discussion

The healthcare sector itself is not a monolith and often necessitates frameworks tailored to specific use cases or applications. In this work we share the data preparation pipeline that allows us to integrate clinical and research environments to get the data ready for AI projects.

The choice to use open-source platforms in the research environment was to ensure those processes are translatable to other institutions/groups when needed. 3D Slicer was initially set up as our DICOM listener for image data curation. However, although 3D Slicer offers many capabilities, it was not a lightweight solution and could not

handle receiving large volumes of data. We migrated to Orthanc DICOM server for receiving the data from the TPS for the as it seamlessly receives and stores data whilst large volume of data transfer happens from our multiple imaging data sources (e.g. CT simulator, TPS, PACS) for different studies.

The range of DICOM metadata tags available for querying and/or retrieval from any DICOM based PACS is limited and designed for everyday operational use rather than for audit or research purposes. As with other reported studies[4–6] we found query retrieve to be simpler on the DICOM metadata when it is stored independently from the image. We created a bespoke python-based module that could be selectively used to ensure data integrity (e.g.. ensuring availability of RTSTRUCT related to an imaging series). This allows us to check for missing data or audit the imaging protocol. This module can also be utilized as an automated procedure for all patients within the study of choice prior to visual inspection.

As reported by other research groups[7,8], XNAT was chosen for our research data since the access is built on a REST architecture which offers flexibility while interacting with the data. For example, it allows us to integrate the radiomics feature extraction on the data stored and conveniently stores the radiomics features back into the database for each individual image set. It also provides the flexibility to introduce federated learning-based AI models to learn on the data.

There are some limitations and challenges that still need to be resolved. For e.g. the data entry for clinical data is still manual entry which is both time-consuming and error prone. Next steps include creation of API that could capture some of the data fields from the EMR directly to the Redcap database.

5 Conclusion

Our work describes a meticulous data preparation pipeline tailored for AI integration in Radiation Oncology.
This work emphasizes the importance of context-specific frameworks in addressing clinical challenges. The presented pipeline contributes to the discourse on AI implementation in Radiation Oncology, providing insights for future advancements in the field.

References

1. REDCap. https://www.project-redcap.org/.
2. MIRC CTP - MircWiki. https://mircwiki.rsna.org/index.php?title=MIRC_CTP.
3. XNAT - Home. https://www.xnat.org/.
4. Santos, M. & Rocha, N. P. Outcomes from Indexing Initiatives of Medical Imaging DICOM Metadata Repositories. A Secondary Analysis. *Procedia Comput. Sci.* **138**, 203–208 (2018).
5. Mackenzie, A., Lewis, E. & Loveland, J. Successes and challenges in extracting information from DICOM image databases for audit and research. *Br. J. Radiol.* **96**, 20230104 (2023).
6. Kathiravelu, P. *et al.* A DICOM Framework for Machine Learning and Processing Pipelines Against Real-time Radiology Images. *J. Digit. Imaging* **34**, 1005–1013 (2021).
7. Gutman, D. A. *et al.* Web based tools for visualizing imaging data and development of XNATView, a zero footprint image viewer. *Front. Neuroinformatics* **8**, 53 (2014).
8. XNAT Soup, a data finding utility used to visualize col | Open-i. https://openi.nlm.nih.gov/detailedresult?img=PMC4034701_fninf-08-00053-g0007&query=&req=4.

Transfer Learning for the Detection of Lung Lesions in ^{18}F FDG PET/CT Images using GraphNet

Maryam Fallahpoor[1], Biswajeet Pradhan[1,2]

[1]Center for Advanced Modelling and Geospatial Information Systems, Faculty of Engineering and IT

[2]Earth Observation Centre, Institute of Climate Change, Universiti Kebangsaan Malaysia, Bangi 43600, Malaysia

Abstract

Early detection of lung cancer relies on advanced imaging techniques such as PET/CT scan, with ^{18}F FDG being one of the most commonly used radiotracers in PET/CT imaging. Deep learning has been demonstrated to yield promising results in automated lung lesion detection. Challenges, including data scarcity, computational costs, and robust deep learning models to detect small and complex patterns, hinder fully training deep learning models for accurate lung lesions detection. To address these challenges, this study employs sixty ^{18}F FDG PET/CT images for small and complex lung lesion detection using transfer learning with GraphNet model, alongside accurate lung segmentation and co-registration performed using Python-based codes for pre-processing. We achieved an accuracy of 82% and a sensitivity of 78.79%, which shows the potential of transfer learning with GraphNet for identifying small lesions in PET/CT scans. This approach may also help to detect radiation-induced complex lesions, offering valuable insights for treatment monitoring.

Keywords: Lung cancer; Deep learning; PET/CT; Transfer learning; GraphNet

1 Introduction

Lung cancer stands as one of the most lethal malignancies worldwide, highlighting the critical importance of early tumor detection in its early stage. The integration of ^{18}F fluorodeoxyglucose with the valuable imaging modality positron emission tomography/computed tomography (^{18}F FDG PET/CT) is one of the most successful imaging in lung malignancy detection. The radioactive nature of the ^{18}F FDG radiotracer necessitates consideration of radiation protection protocols to minimize potential risks to both patients and healthcare providers and also the development of automated procedures for diagnosis [1]. Due to the prevalence of pulmonary lesions, there is an increasing demand for approaches that guarantee both high sensitivity and efficiency in identification of FDG-avid lung lesion foci. It has been proven that artificial intelligence, specifically its application of deep learning, has yielded outstanding results in detecting lung-related malignancies using PET/CT images [2]. However, some significant challenges are present. First, limited PET/CT data availability often is a challenge for developing robust models. High computational costs, time-consuming training processes, and poor generalization are among the other challenges associated with fully training deep learning models. Additionally, there is a need to utilize a model capable of detecting small structures and patterns to identify the small size of lesions, which have the potential to develop into malignancies. In this case, transfer learning has proven to be a promising technique for using pre-trained models to address data scarcity, particularly in medical imaging applications. While valuable studies have explored transfer learning with PET/CT images [3-7], a notable drawback is the lack of employing a deep learning model specifically tailored to detect small and complex structures within the lung. In this study, we aim to detect lung lesions from ^{18}F FDG PET/CT images using transfer learning and GraphNet [8] for the detection of small lesions and complex patterns in the lung, as it addresses this critical gap in current research and holds promise for improved detection accuracy in PET/CT scans.

2 Materials and Methods

We used 60 ^{18}F FDG PET/CT images, including 30 lung cancer patients and those patients who had malignant lung lesions and 30 normal lung images. Patient preparation and image acquisition adhere to PET/CT imaging guidelines to ensure the attainment of high-quality images [9]. In this retrospective study, patients underwent 18F FDG PET/CT scans as their clinical examinations. The dataset was pre-processed by cropping and resizing all images to a standardized dimension of 120 x 120 x 80. Subsequently, PET and CT images were co-registered, and for more accurate results, lung areas were accurately segmented from the both PET and CT images using Python-based code (Figure 1). The transfer learning process using GraphNet involved using a pre-trained deep learning model trained on a comprehensive dataset of similar ^{18}F FDG PET/CT images. In this study, the last layers of the deep learning model used in full training were frozen to facilitate transfer learning, allowing the model to adapt its knowledge to the specific characteristics of the lung cancer dataset.

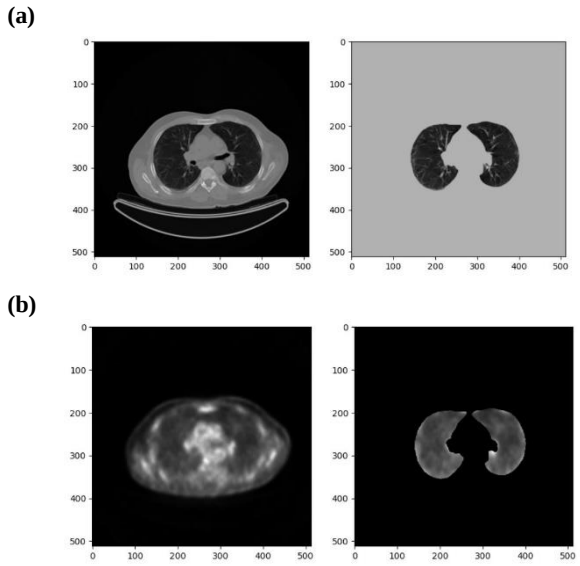


Figure 1. (a) Original CT scan and segmented lung, (b) Original PET scan and segmented lung.

3 Results

Statistical results showed the capability of transfer learning for lung lesion detection. The performance of the proposed approach revealed an impressive accuracy rate of 82% in correctly detecting lung lesions. We reached a sensitivity of 78.79%, showing the proportion of true positive identifications relative to all actual positive cases. The specificity of the present approach was 85.19% for lung lesion detection that indicates the proportion of true negative identifications relative to all actual negative cases.

4 Discussion

The results of the study demonstrate promising outcomes of using transfer learning with GraphNet for the detection of lung lesions, specifically small lesions, in ^{18}F FDG PET/CT images. By leveraging pre-existing knowledge from a large dataset of similar images, the model can adapt its understanding to the specific characteristics of the lung cancer dataset, leading to enhanced diagnostic performance. This capability could extend to detect radiation-induced lesions or changes in tissues affected by radiation therapy. For example, the model could potentially aid in identifying the impact of radiation on lung injuries or changes in tumor metabolism post-radiation therapy, enhancing the ability to monitor treatment response and assess treatment-related complications.

5 Conclusion

This study sheds light on future investigations, suggesting that transfer learning using GraphNet can be a valuable tool for optimizing small and complex tumor detection in cases where data scarcity poses a significant challenge.

References

1. Alenezi, A. and K. Soliman, *Trends in radiation protection of positron emission tomography/computed tomography imaging.* Annals of the ICRP, 2015. **44**(1_suppl): p. 259-279.
2. Schwyzer, M., et al., *Automated detection of lung cancer at ultralow dose PET/CT by deep neural networks–initial results.* Lung Cancer, 2018. **126**: p. 170-173.
3. Capobianco, N., et al., *Transfer learning of AI-based uptake classification from 18F-FDG PET/CT to 68Ga-PSMA-11 PET/CT for whole-body tumor burden assessment*. 2020, Soc Nuclear Med.
4. Kumar, A., et al., *Co-learning feature fusion maps from PET-CT images of lung cancer.* IEEE Transactions on Medical Imaging, 2019. **39**(1): p. 204-217.
5. Apostolopoulos, I.D., N.D. Papathanasiou, and G.S. Panayiotakis, *Classification of lung nodule malignancy in computed tomography imaging utilising generative adversarial networks and semi-supervised transfer learning.* Biocybernetics and Biomedical Engineering, 2021. **41**(4): p. 1243-1257.
6. Wallis, D., et al., *An [18F] FDG-PET/CT deep learning method for fully automated detection of pathological mediastinal lymph nodes in lung cancer patients.* European Journal of Nuclear Medicine and Molecular Imaging, 2022: p. 1-8.
7. Han, Y., et al., *Histologic subtype classification of non-small cell lung cancer using PET/CT images.* European journal of nuclear medicine and molecular imaging, 2021. **48**: p. 350-360.
8. Pu, M., et al. *GraphNet: Learning image pseudo annotations for weakly-supervised semantic segmentation*. in *Proceedings of the 26th ACM international conference on Multimedia*. 2018.
9. Boellaard, R., et al., *FDG PET/CT: EANM procedure guidelines for tumour imaging: version 2.0.* European journal of nuclear medicine and molecular imaging, 2015. **42**: p. 328-354.

Optimising physics-informed CRNNs for longitudinal volumetric medical image data

Denis Page[1], Eliana Vasquez Osorio[1], Robert Chuter[1,2], Andrew Green[1,3], and Alan McWilliam[1]

[1]Advanced Radiotherapy, University of Manchester, Manchester, UK
[2]Radiotherapy Related Research, The Christie NHS Foundation Trust, Manchester M20 4BX, UK
[3]European Bioinformatics Institute, Wellcome Genome Campus, Cambridge, CB10 1SD

Abstract Anatomical changes during radiotherapy can make highly conformal treatment plans, calculated on pre-treatment anatomy, sub-optimal or unsafe to deliver. Daily images of patients are evaluated against the treatment plan to evaluate delivery. Where delivery is deemed unsafe, re-planning will occur which, in a reactive workflow, can draw finite resources from elsewhere or may cause delays in patient treatment. In this work, we present the first use of a physics-informed convolutional recurrent neural network for longitudinal volumetric medical image data. This work aims to optimise the hyperparameters in our model for medical image data. The best performing model produced anatomical predictions at the next five treatment fractions with a mean squared error of 0.077 on images normalised between 0 and 1.

1 Introduction

Radiotherapy is delivered over a number of daily fractions. Throughout treatment, many patients will undergo changes to their anatomy, which can be systematic, for example weight loss, or random, for example rectal gas changes. Such changes can leave treatment plans optimised on a prior anatomical state sub-optimal to deliver. To ensure plans remain safe to deliver, they are evaluated on daily images, conventionally cone-beam CT (CBCT), taken prior to treatment delivery. If the plan is no longer considered suitable due to large anatomical changes, re-planning will occur. In a reactive workflow, this can draw limited resources from elsewhere [1]. Ideally, resources should be pre-allocated for a planned re-optimisation of the treatment plan. However, to enable this, we need workflows that can identify which patients and when anatomical changes are likely to manifest. This will allow efficient ahead-of-time resource allocation to reduce the clinical support and workflow disruption currently required for adaptive radiotherapy (ART).

To address this, we have drawn on machine learning methods originally deployed for analogous tasks in precipitation forecasting [2][3] and video frame prediction [4]. We adapt these frameworks and implement the first example of a physics-informed convolutional recurrent neural network (CRNN) to be used on longitudinal volumetric medical image data. Our example uses the daily CBCT images for head-and-neck cancer patients, predicting anatomy five treatment sessions into the future.

Adopting approaches shown to be successful outside of radiotherapy can help drive research into developing clinically beneficial machine learning applications, but every domain comes with its unique challenges. To implement image-based machine learning methods for clinical use, they must be adapted and optimised for medical image data. For novel approaches, the optimal design choices may not be well documented, and grid or random searches when tuning many hyperparameters may be prohibitively time intensive. This is particularly true of generative tasks using large models to handle complex data sets which may take days to train. In this study, we present our optimisation workflow to select appropriate hyperparameters for generative machine learning approaches to produce forecasts of volumetric longitudinal medical image data by aggressively pruning low performing models.

2 Materials and Methods

For this study we used an in-house data set, with ethics approval from ukCAT, a research database established at The Christie NHS Foundation Trust. It consisted of 427 series of 10 CBCT daily scans, totalling 42700 3D CBCTs, with a voxel spacing of 1 x 1 x 1 mm. CBCTs were acquired with standard in-house settings of 100 kV and 16 mA using ELEKTA XVI version 5.52. The maximum temporal gap between consecutive scans is 3 days, consistent with a break in treatment over a weekend. Each series of CBCTs was normalised to the associated planning CT. First by rigid registering of each CBCT to the planning CT using a match clip box around the vertebrae. Then secondly we used the brainstem contours to define a region used to normalise the image signal on the CBCT to the CT. The brain stem was chosen due to its centrality and consistent signal. During pre-processing for the CRNN, the image signal was normalised between 0 and 1.

2.1 Model

The CRNN architecture used in this work combines convolutional neural networks (CNNs), which effectively model spatial relationships in image data, and recurrent neural networks (RNNs), which are designed to handle temporal relationship in longitudinal data. The architecture used in this study is based on the ConvLSTM introduced by Shi et al [2], a CRNN able to more effectively model long range tempo-

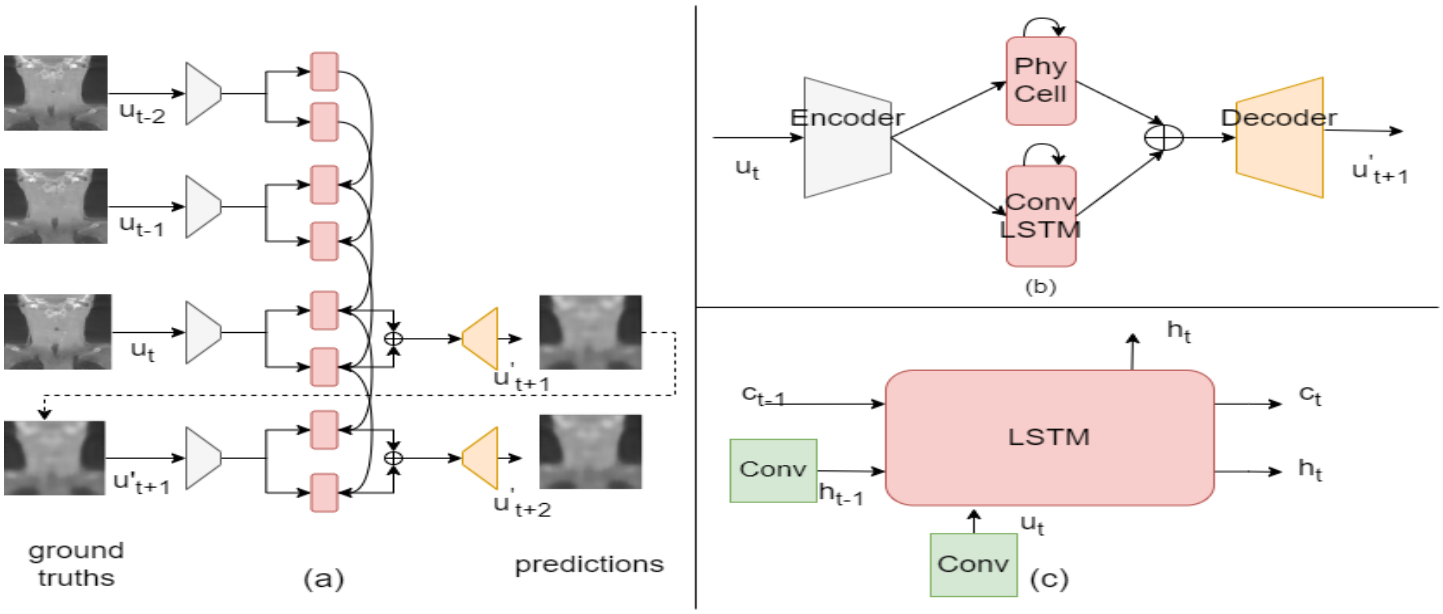


Figure 1: A schematic of the architecture employed in this work. (a) depicts the sequence to sequence architecture, unfolded in time. u_t represents the and input a time t and $u^{'}$ terms represent predictions that are also used as inputs for the proceeding time step. Arrows represent the flow of information both through the network and through time. The dashed line indicates that the input is only passed to the ConvLSTM. (b) shows the recurrent PhyCell which models physical dynamics in input data and the ConvLSTM cell, more closely examined in (c), which uses the cell state, c, to control the flow of information longitudinally through the network, also outputting the hidden state, h both as a prediction and as an input to the next layer.

ral dependencies, and the PhyDNet introduced by Le Guen and Thome [5], which employs a parallel-branch approach to encode PDEs into the model which constrain outputs to a physically plausible region. The ConvLSTM constitutes one of the branches of the PhyDNet, which is turn is part of a sequence-to-sequence model, taking 5 consecutive daily CBCTs as an input, and generating the next 5 CBCTs in the series. The generated images are compared to ground truths to measure prediction accuracy. A schematic of the architecture is displayed in Figure 1.

The final model was implemented using PyTorch Lightning 2.0.1 [6] and has 691M parameters. All training was performed on a 24GB NVIDIA GeForce RTX 3090 Turbo and an AMD Ryzen 7 3800X 8-Core Processor. With 42700 volumes, 21GB of RAM was required.

2.2 Optimisation workflow

Exhaustively tuning all hyperparameters by selecting a range of values for each and producing models for every combination would represent the most comprehensive examination possible. This would, however, be prohibitively time consuming. When making a grid search of the hyperparameter space, a large number of hyperparameters can quickly increase the time and computational resources required. The number of runs can be expressed by

$$\#(n_1 \cdot n_2 \cdot \ldots \cdot n_{n_{trials}}) \quad (1)$$

where n_1 is the number values a hyperparameter can take for a single trial and n_{trials} is the number of trials. Instead, hyperparameter tuning was split into six "phases" with the greatest conditional relationships, and were conducted consecutively. Within each phase, multiple trials are conducted and within each trial one or multiple hyperparameters are optimised. The best performing value is then used for all subsequent tuning. Simultaneous or isolated optimisation is chosen depending on the interdependence between hyperparameters. Where possible, we have tried to avoid simultaneous training to keep optimisation time to a minimum. Initial configuration for the hyperparameters are taken from PhyDNet and are updated with the tuned value at the end of each trial. The trials are detailed in Table 1. We select a range of values which feature in literature describing approaches to analogous tasks, both within and beyond radiotherapy.

Each trial involved training multiple models with each hyperparameter configuration, and was conducted using 100 sequences (roughly one quarter of the data set) for 80 training epochs. The 100 sequences were split 70/20/10 for training, validation and test sets, and the mean squared error (MSE) of each model was evaluated on the test set and is used as the trial result. Trials were conducted using the Tune [7] library, allowing for all model training to be scheduled in advance and run consecutively.

At the end of each trial, the optimised hyperparameter value(s) are noted and the five best performing model configurations were selected for full training using all 427 sequences, again split 70/20/10, for 300 epochs. The performance of the more extensively trained model was then compared to the corresponding optimisation trial model to investigate whether choices made to improve the accuracy of trial models corresponds to improved accuracy in fully trained models.

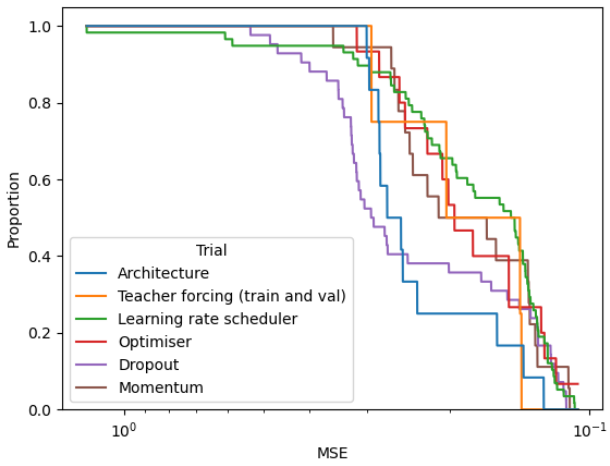


Figure 2: The proportion of six groups of hyperparameter-investigating experiments for MSE values between 1 and 0.1 (horizontal axis decreases to the right). A lower MSE (further to the right) indicates better performance, with flatter curves indicative of a greater mix of suitable and poorly performing hyperparameter options.

Hyperparameters with well documented optimal values were not tuned further.

3 Results

Using our methodology, the tuning of the hyperparameters presented here was conducted in a training time of 340 hours. Applying Equation 1 to the six trials presented here would require 32400 hours.
An example of the input, ground truths and predictions of the network is displayed in Figure 3.

3.1 Hyperparameter tuning

The MSE of all predictions made by models used in eight tuning experiments evaluated on a test set is displayed in Figure 2. Each model produced during this stage required around 2.5 hours to train. The best performing model achieved an MSE of 0.103. Other loss functions were investigated, but using MSE produced the highest accuracy and most consistent training. The best performing values for the eight most influential hyperparameters are displayed in Table 1. Where differences in accuracy could not be determined, all possible values are provided. The proportion of models reaching MSE thresholds, grouped by phase, is displayed in Figure 2.

3.2 Full training

Each model produced during this stage required around 40 hours to train. Of the five model configurations selected for full training, the lowest MSE was 0.0766, and the highest was 0.1637. Three showed an improvement on the corresponding model produced during optimisation trials, one showed

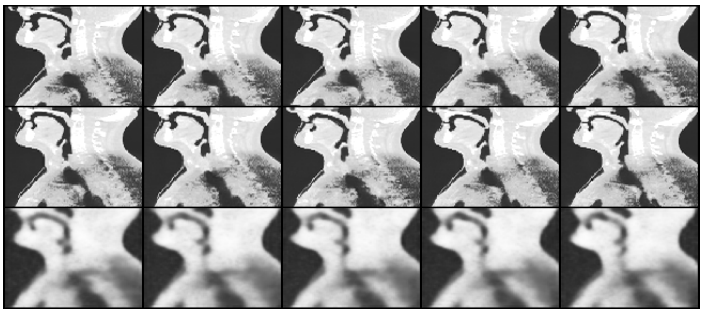


Figure 3: An example of the input to the network (top row), the ground truths (middle row) and predictions (bottom row) of the network. Each prediction is compared to the corresponding ground truth to calculate the loss, which is MSE during training.

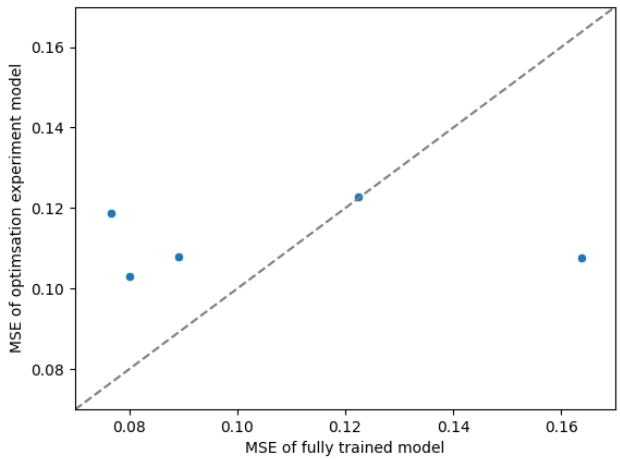
Table 1: The value(s) of eight influential hyperparamter in the top five performing models during optimisation.

Hyperparameter	Best experimental value(s)
Number of layers/nodes per layer (PhyCell, ConvLSTM)	[512,512,512,512,512,512,256, 256,128,128,64,64],[729,729]
Dropout (Encoder-Decoder, PhyCell, ConvLSTM)	0.1, 0.1, 0.25
Teacher forcing (train, validation)	-1% per epoch, -1% per epoch
Optimiser	RMSProp
Momentum (for RMSProp)	0.4
Starting learning rate	1×10^{-4}
Learning rate scheduler	ReduceOnPlateau CosineAnnealingLR+decay
Learning rate reduction factor	0.5

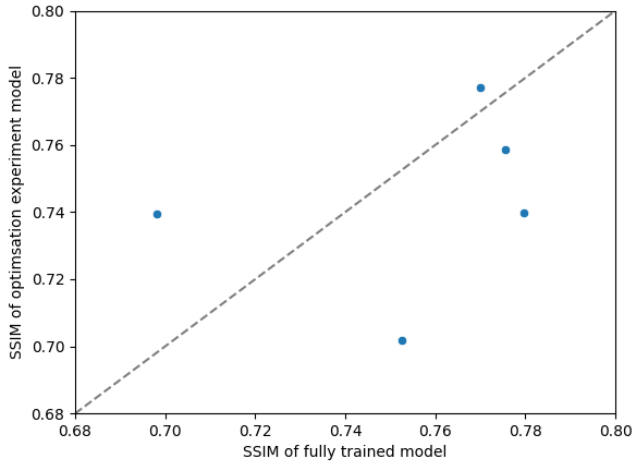
almost identical performance and one showed an reduction in prediction accuracy. A comparison of model performance between short and full training for the top five model configurations is displayed in Figure 4 using MSE and the structural similarity index measure (SSIM).

4 Discussion

In this study, we presented a physics-informed CRNN approach to predicting gross anatomical changes in HN patients being treated with radiotherapy. This represents the first time a physics-informed CRNN had been optimised for such a task. We outlined our methodology for optimising a machine learning model that has been adopted in other fields specifically for longitudinal radiotherapy data, and presented 8 key hyperparamater choices. The significantly shorter training time required to tune the model, compared to an exhaustive exploration, represents a reduction in the computational resources required to optimise the model, and increases the number of approaches that can be investigated in a given time frame.
Four of the five fully trained model configurations performed as well, or improved upon their corresponding trial models, providing evidence that the optimisation performed on subsets of the data for less training time is applicable to models trained for longer on full data sets, and only one of the models



(a)



(b)

Figure 4: A comparison of the MSE (a) and SSIM (b) for the fully trained models (horizontal axis) against the MSE for the corresponding hyperparameter configuration in the optimisation experiment (vertical axis). Points in the top left portion for MSE and in the bottom right for SSIM represent instances where the fully trained model outperforms the model produced during tuning.

showed a reduction in prediction accuracy. A possible explanation for the reduced prediction accuracy is over-fitting, despite hyperparameters associated with dropout and regularisation being optimised. This could indicate that the proposed workflow may not be valid for all hyperparameters, particularly those aiming to reduce over-fitting. Further work to better select hyperparameters used to reduce over-fitting may be required to guarantee robustness to over-fitting, where the effect may become more acute with a greater number of training epochs.

It is also important to consider that, as investigated hyperparameters are frozen after each trial, it is possible that the path selected may have left optimal parameter choices out of the testing pool. Whilst our approach attempts to mitigate this by combining related hyperparameters and deliberately ordering trials, these decisions are dependent on the workflow designer and may, themselves, not be optimal. The trade-off between speed and intensity of optimisation should be considered by the architect of the model.

The optimisation of the PhyDNet to medical image data does improve prediction accuracy, however, the uncertainty in anatomical predictions is not yet suitable for use in clinical predictions. Future work will examine any non-analogous aspects of intended use cases for PhyDNet and anatomical change due to radiotherapy. One such difference may be that the longitudinal data produced during treatment is not uniformly temporally spaced as it is for the frames of a video, increasing uncertainty in the predictions. We are currently working on adapting the network architecture to handle time-dependent inputs.

5 Conclusion

We have implemented an aggressive pruning workflow to optimise the novel use of a physics-informed CRNN on volumetric longitudinal medical image data using a series of small, ordered model training experiments. The optimisation improved the model performance compared to the baseline model published for use on similar tasks and predictions of patients' future anatomy up to one week were produced, taking around two weeks of training to complete. However, hyperparameters that dictate model performance more heavily at later training epochs may not always be able to be tuned using this workflow. We demonstrate a significant reduction in the time and computational resources required compared to a grid search, providing hyperparameter values for training large models.

References

[1] B. J. Slotman et al. "Overview of national guidelines for infrastructure and staffing of radiotherapy. ESTRO-QUARTS: work package 1". *Radiotherapy & Oncology* 75.3 (2005), pp. 349–354.

[2] X. Shi, Z. Chen, H. Wang, et al. "Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting". *CoRR* abs/1506.04214 (2015).

[3] Z. Zhang and Y. Dong. "Temperature Forecasting via Convolutional Recurrent Neural Networks Based on Time-Series Data". *Complexity* 2020 (2020), pp. 1–8.

[4] N. Srivastavan, E. Mansimov, and R. Salakhudinov. "Unsupervised learning of representations using LSTMs". *International Conference on Machine Learning (ICML)* 2.5 (2015), pp. 843–852.

[5] V. Le Guen and N. Thome. "Disentangling Physical Dynamics From Unknown Factors for Unsupervised Video Prediction". *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020, pp. 11471–11481. DOI: 10.1109/CVPR42600.2020.01149.

[6] W. F. et al. "PyTorch Lightning". *GitHub. Note: https://github.com/PyTorchLightning/pytorch-lightning* 3 (2019).

[7] R. Liaw, E. Liang, R. Nishihara, et al. "Tune: A Research Platform for Distributed Model Selection and Training". *arXiv preprint arXiv:1807.05118* (2018).

i-SART: A Prototype AI-based Tool to Enhance Patient Safety in Radiotherapy

Anastasia Sarchosoglou[1], Natalia Silvis-Cividjian[2], Yijing Zhou[2], Periklis Papavasileiou[1], Athanasios Bakas[1], and Evangelos Pappas[1]

[1]Department of Biomedical Sciences, Radiology & Radiotherapy Sector, University of West Attica, Athens, Greece
[2]Department of Computer Science, Vrije Universiteit Amsterdam, The Netherlands

Abstract In the complex field of radiotherapy (RT), ensuring patient safety is crucial. Risk management is a valuable tool for aiding RT departments in achieving this goal by systematically addressing hazards in their systems. Disseminating information about potential failures has been recognized as paramount to improving patient safety globally. This paper presents i-SART, a novel web application that aims to facilitate proactive risk management and share knowledge about potential failure modes (FMs) in a timely and efficient way. Based on the Failure Modes and Effects Analysis (FMEA) methodology, i-SART features a database of 419 FMs. The innovation lies in the integration of AI techniques that engage RT professionals in an interactive learning process by generating discussions about potential failures and safety measures. The web application is a work in progress and will be released soon for user evaluation. We anticipate i-SART to effectively promote proactive learning in FMs, enhance patient safety, and pave the way for advanced AI-based risk management in RT.

1 Introduction

In the ever-evolving landscape of radiotherapy (RT), where precision is paramount, ensuring patient safety is crucial. The meticulous approach required in RT demands a comprehensive strategy to mitigate potential risks and enhance the overall safety in the patient pathway.

In RT, stringent safety standards and robust measures govern the use of radiation. Although accidents are rare, errors may occur at various stages, occasionally compromising clinical outcomes. The International Commission on Radiological Protection (ICRP) underscores the potential risks, especially those associated with new technologies [1].

To minimize the risk of failures and prevent unintended exposure, the implementation of risk management in RT has been regulated in many countries [2, 3], utilizing both reactive and proactive methodologies. Reactive risk management employs Incident Learning Systems (ILS) to collect and analyze adverse events, errors, or near misses. Notable international and national ILS, such as SAFRON (SAFety in Radiation Oncology) [4], ROSEIS (Radiation Oncology Safety Education and Information System) [5], and RO-ILS (Radiation Oncology-Incident Learning System) [6], provide valuable insights about incidents to the RT community. Proactive risk management involves the implementation of systematic methods that assist organizations in managing their weaknesses by identifying and addressing potential failures before they occur [7].

Approaches used in RT include Failure Mode and Effects Analysis (FMEA) and its variations, such as Failure Mode and Effects Criticality Analysis (FMECA) and Healthcare Failure Mode and Effects Analysis (HFMEA), as well as other methods like Fault Tree Analysis (FTA) [7], Hazard and Operability Analysis (HAZOP) [8] and the Systems Theoretic Accident Model and Process (STAMP) [9, 10].

Although many departments have successfully implemented proactive risk management, others face challenges due to limited knowledge of assessment methodologies and a lack of resources. Moreover, proactive risk management involves forward-thinking, focusing on identifying and preventing potential hazards before they manifest, which is particularly challenging especially when implementing new technologies or techniques. The World Health Organization's (WHO) Global Action Plan for Patient Safety (2021-2030) [11] emphasizes the proactive identification and mitigation of risks in healthcare, aiming to enhance patient safety. It calls for the dissemination of such knowledge, yet the vast majority of knowledge about potential risks is not disseminated and remains compartmentalized within the departments. While national and international ILS provide learning about incidents post-occurrence, there is no similar platform that promotes proactive knowledge before potential failures turn into errors.

Our research aims to bridge this gap by developing a web application, the i-SART, an Intelligent Safety Assistant in RT that facilitates proactive risk management and enhances patient safety awareness. Central to this effort is a database that gathers a large number of potential failure modes (FMs) defined as the ways or modes a process or system can fail. Innovative to our approach is the integration of AI techniques that engage RT professionals in an interactive learning process about FMs and safety measures. The advantage of i-SART is that knowledge about potential failures can be disseminated in a timely and efficient way, worldwide. We envision i-SART as a free-to-access tool for all RT users fostering collaboration and future improvements. This is a collaborative effort between the Faculty of Biomedical Sciences of the University of West Attica and the Faculty of Computer Sciences of Vrije Universiteit Amsterdam and is a work in progress.

2 Materials and Methods

Design of i-SART

The design of i-SART was based on FMEA methodology, selected for its ability to yield both qualitative and quantitative results. In an FMEA process study, a dedicated team identifies FMs, along with their causes and effects. The team assesses the FMs in terms of their likelihood of occurrence (O), lack of detectability (D), and potential severity (S). This evaluation aids in prioritizing FMs for implementing risk mitigation strategies.

I-SART used the seminal report of the American Association of Physics in Medicine (AAPM), the AAPM TG100

[12], as a benchmark. This report was developed as a proactive risk management FMEA-based methodology specific to RT.

Initially, we created a process map that depicts chronologically the sub-processes and their steps within the main RT process. The i-SART process map includes 8 sup-process and 61 steps.

1. Patient Assessment / Initial Consultation (13 steps).
2. Scheduling procedures (4 steps).
3. Imaging for RT planning (14 steps).
4. Treatment planning (TP) (10 steps).
5. Pre-treatment QA procedures (3 steps).
6. Treatment session (10 steps) (Fig. 1).
7. On-treatment quality management (4 steps).
8. Post-treatment procedures & Follow-up (3 steps).

Additionally, the specific procedure of the advanced technology/technique MR-guided adaptive RT (MRgART) was separately mapped in 12 steps.

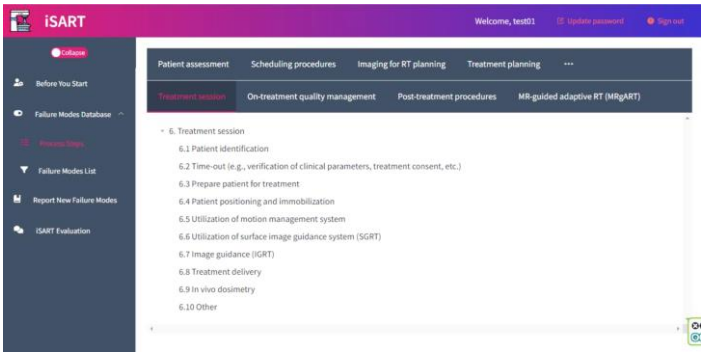


Figure 1: Sample of i-SART interface showing the steps in the treatment session sub-process.

At each step, the collected FMs (as described in the results section), their severity, causes and effects, and mitigation strategies were stored in the database. To ensure a harmonized severity (S) across all selected FMEA studies, we created a modified scale in the form of traffic lights for ease of visualization (Fig. 2).

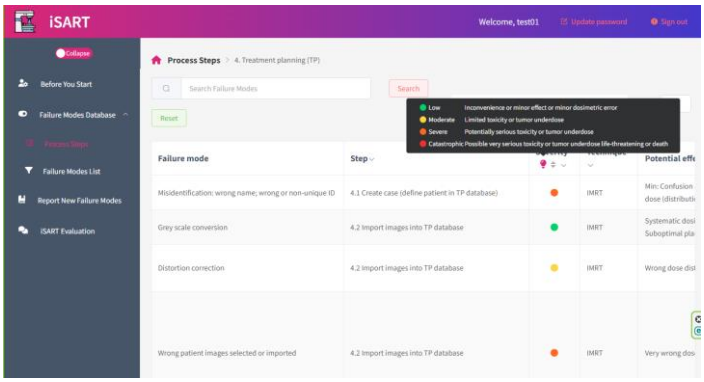


Figure 2: Sample failure modes within the treatment planning sub-process, indicating their severity using a traffic light format.

Data collection

Our research included a sample of FMEA studies that met our eligibility criteria: diversity in external photon beam RT techniques, country representation, studies conducted in English, and the provision of sufficient data to align with our design's objective of offering comprehensive information on FMs.

Technical Description i-SART

Technically, i-SART is a cloud-hosted web application with two user roles: administrator and user, each having distinct interfaces and data accesses. The administrator, who is an RT expert, manages the system and the application database.

The back-end of the web application was developed using Python 3.9 and the Django REST framework. Secure information transfer between client and server was ensured through JSON Web Token authentication. A relational database MySQL, was chosen for persistent storage of user and FM information.

The user interface of i-SART was created using Vue.js, a JavaScript based framework. FM search, filtering, and sorting functions are available. There is also an option for the user to insert in the database a new FM validated by the i-SART administrator. The administrator can also monitor statistics on FMs in various charts. Additionally, all users can evaluate the tool and provide comments and suggestions for improvement. The general workflow of i-SART is shown in Fig. 3.

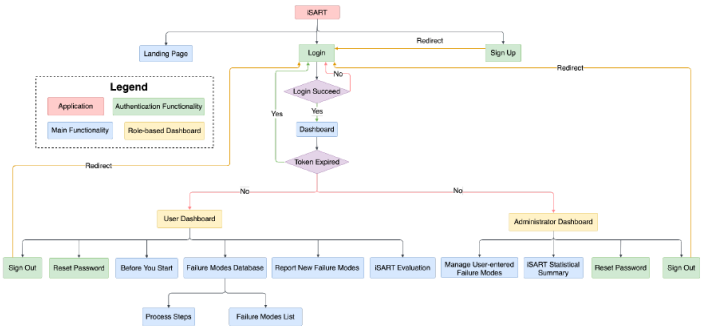


Figure 3: The general workflow of i-SART.

Experimental work on synthetic FMs

To augment the database for potential future use in machine learning, we experimented with the possibility of generating synthetic (artificial) FMs using Generative AI (GenAI) techniques such as Hidden Markov Models, ChatGPT-3, and Generative Adversarial Network (GAN) [13]. Specifically, for GAN, we employed the seqGAN model [14] with a total of 2,305 records extracted from our FMs database and incident data from SAFRON [4] (Fig. 4).

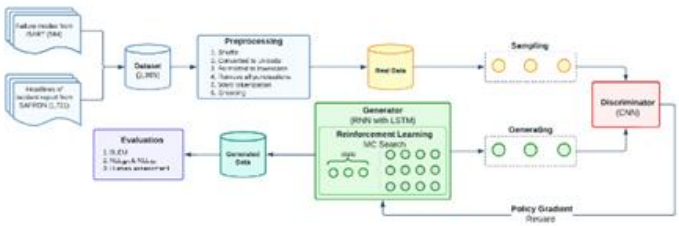


Figure 4: The architecture of the GAN-based synthetic FMs generation experiment.

Current work

We are currently working on integrating an AI tool, specifically a chatbot, into i-SART. This addition aims to elevate user interaction by generating engaging discussions about FMs. Users will have the opportunity to seek insights, clarification, and practical solutions based on white papers, seminal reports, and relevant literature, all within the i-SART. This feature not only promotes proactive learning but also serves as a valuable resource for users seeking timely and context-specific information.

3 Results

Initially, a total of 728 FMs were gathered from 10 FMEA articles and reports. The seminal AAPM TG100 report [12], which provides potential FMs in Intensity Modulated Radiotherapy (IMRT), served as the foundational source, Complementing this, seven articles were included that explored FMs in: RT process [15], 3D Conformal RT [16], lung Volumetric Modulated Arc Therapy (VMAT) [17], Surface Guided deep inspiration breath -hold (DIBH) breast RT [18], Stereotactic Radiosurgery/Radiotherapy [19] and MR guided adaptive RT [20, 21]. Additionally, data from two unpublished reports from the UK investigating breast VMAT and lung Stereotactic Body Radiation Therapy (SBRT) were incorporated.

To identify duplicate FMs across the articles and reports, we utilized a semi-automated approach, initially employing AI-based Natural Language Processing (NLP) algorithms. Results were subsequently manually reviewed by researchers to confirm duplications. We also merged common-ground FMs that included supplementary information while excluding ambiguous or insufficiently detailed FMs from our database. The total number of FMs in i-SART is currently 419 as illustrated in Fig. 5.

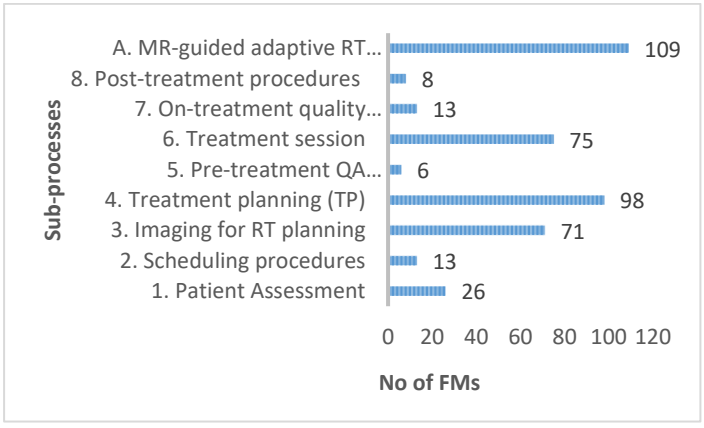


Figure 5: Number of FMs per sub-processes from 1 to 8. MR-guided Adaptive RT is indicated by the letter A.

Our experiments with GenAI to augment the FM database resulted in very few interesting and sensible new FMs, most probably due to the small size of the training dataset, which will hopefully increase, as i-SART will gain more users. The results were not incorporated into the database. However, these explorations lay the groundwork for potential future work.

We will soon release the application for user evaluation to assess its effectiveness and identify areas for improvement. We anticipate that users will contribute new FMs to enhance the database, fostering continuous growth and knowledge exchange.

4 Discussion

Proactive risk assessment within RT has garnered recognition as an essential tool for preventing incidents and errors. While national and international ILS offer valuable insights into lessons learned from past events, there is a noticeable gap in proactive learning on potential risks. Our research aimed to bridge this gap through the development of i-SART, an intelligent Safety Assistant in RT.

Built upon the FMEA and AAPM TG-100, i-SART takes a systematic approach to create a large database of potential FMs in RT process. The integration of an AI chatbot will further enhance user engagement by generating discussions about FMs and safety measures. Additionally, the tool provides a platform for users to contribute new FMs, promoting continuous knowledge exchange. It may also become useful for RT vendors by offering insights into potential risks based on real-world practice. Furthermore, i-SART has the potential to complement free–access ILS, contributing to an overarching approach to risk management for the wider RT community.

Acknowledging its limitations, i-SART currently does not encompass all existing techniques and technologies in RT, focusing on the most widely used ones. Due to the limited number of studies used, generalizations cannot be made, and the tool cannot be considered as a 'one-size-fits-all.' It is indicative and informative, prompting reflection on the user's practice while aiding in the identification and mitigation of potential risks promoting proactive learning.

Looking ahead, we anticipate i-SART to drive further innovations in integrating AI techniques and providing a substantial dataset for machine learning applications, FMs predictions and synthetic FMs generation while also encompassing other proactive risk management methodologies.

5 Conclusion

In this paper, a prototype of i-SART is presented, which is a web application designed to promote proactive learning in

patient safety in RT. By integrating AI capabilities and employing the FMEA methodology, i-SART provides a dynamic platform for proactive risk management. We anticipate i-SART to enhance patient safety awareness and facilitate valuable knowledge exchange, particularly crucial as new techniques and technologies continually emerge. Envisioning the continuous evolution of i-SART, our hope is i-SART to become an essential safety assistant in the hands of every RT professional.

References

[1] Ortiz Lopez P, Cosset J, Dunscombe P, Holmberg O, Rosenwald J-C, Ashton PL, et al. Preventing Accidental Exposures from New External Beam Radiation Therapy Technologies. ICRP Publication 112. Ann ICRP. 2009;39(4):1–86.

[2] The Council of the European Union. Council Directive 2013/59/Euratom of 5 December 2013 laying down basic safety standards for protection against the dangers arising from exposure to ionising radiation, and repealing Directives 89/618/Euratom, 90/641/Euratom, 96/29/Euratom, 97/43/Euratom a. Off J Eur Union. 2014;57(L13):1–73.

[3] International Atomic Energy Agency. Radiation Protection and Safety in Medical Uses of Ionizing Radiation. IAEA Safety Standards Series No. SSG-46. Vienna; 2018. Available from: http://www-ns.iaea.org/standards/

[4] Safety in Radiation Oncology (SAFRON) | IAEA. Available from: https://www.iaea.org/resources/rpop/resources/databases-and-learning-systems/safron

[5] Radiation Oncology Safety Education Information System. Available from: https://roseis.estro.org/

[6] RO-ILS- American Society for Radiation Oncology (ASTRO) - American Society for Radiation Oncology (ASTRO). Available from: https://www.astro.org/Patient-Care-and-Research/Patient-Safety/RO-ILS

[7] European Commission. Radiation Protection N° 181. General guidelines on risk management in external beam radiotherapy. Luxembourg; 2015. Available from: http://europa.eu

[8] Pawlicki T, Dunscombe PB, Mundt AJ, Scalliet P. Quality and Safety in Radiotherapy. Florida: CRC Press; 2011. 632 p.

[9] Pawlicki T, Samost A, Brown DW, Manger RP, Kim GY, Leveson NG. Application of systems and control theory-based hazard analysis to radiation oncology. Med Phys. 2016;43(3):1514–30. Available from: http://dx.doi.org/10.1118/1.4942384http://scitation.aip.org/content/aapm/journal/medphys/43/3?ver=pdfcov

[10] Silvis-Cividjian N, Verbakel W, Admiraal M. Using a systems-theoretic approach to analyze safety in radiation therapy-first steps and lessons learned. Saf Sci. 2020 Feb 1;122. DOI: 10.1016/j.ssci.2019.104519

[11] World Health Organization. Towards eliminating avoidable harm in health care. Draft Global Patient Safety Action Plan 2021-2030. World Health Organization. Geneva: Available from: https://cdn.who.int/media/docs/default-source/patient-safety/gpsap/final-draft-global-patient-safety-action-plan-2021-2030.pdf?sfvrsn=fc8252c5_5

[12] Huq MS, Fraass BA, Dunscombe PB, Gibbons JP, Ibbott GS, Mundt AJ, et al. The report of Task Group 100 of the AAPM: Application of risk analysis methods to radiation therapy quality management. Med Phys. 2016;43(7):4209–62. Available from: http://doi.wiley.com/10.1118/1.4947547

[13] Harshvardhan G, Gourisaria MK, Pandey M, Rautaray SS. A comprehensive survey and analysis of generative models in machine learning. Comput Sci Rev. 2020;38:100285. Available from: https://doi.org/10.1016/j.cosrev.2020.100285

[14] Yu L, Zhang W, Wang J, Yu Y. SeqGAN: Sequence generative adversarial nets with policy gradient. 31st AAAI Conf Artif Intell AAAI 2017. 2017:2852–8.

[15] Pobbiati C, Monturano M, Vavassori A, Gerardi M, Fiore MS, Rondi E, et al. Clinical risk analysis of the patient's path in an Advanced Radiotherapy Center (A.R.C.) through F.M.E.A. method. J Biomed Pract [Internet]. 2019;3(1):38–69. Available from: https://www.ojs.unito.it/index.php/jbp/article/view/3376

[16] Muhammad SG, Sharawy AI, Abdalla MH, Hakim A, Meckawy RG, Hakim ANK. Failure Mode and Effect Analysis for Three Dimensional Radiotherapy At Ain Shams University Hospital. Int J Adv Res. 2021;9(12):971–83.

[17] Gilmore MDF, Rowbottom CG. Evaluation of failure modes and effect analysis for routine risk assessment of lung radiotherapy at a UK center. J Appl Clin Med Phys. 2021;22(5):36–47.

[18] Bright M, Foster RD, Hampton CJ, Ruiz J, Moeller B. Failure modes and effects analysis for surface-guided DIBH breast radiotherapy. J Appl Clin Med Phys. 2022;23(4). Available from: https://pubmed.ncbi.nlm.nih.gov/35112445/

[19] Sarchosoglou AA, Papavasileiou P, Bakas A, Stasinou D, Pappas E. Failure modes in stereotactic radiosurgery. A narrative review. Radiography.2022;28(4):999–1009. Available from: https://doi.org/10.1016/j.radi.2022.07.007

[20] Klüter S, Schrenk O, Renkamp CK, Gliessmann S, Kress M, Debus J, et al. A practical implementation of risk management for the clinical introduction of online adaptive Magnetic Resonance-guided radiotherapy: SUPPLEMENTARY MATERIAL. Phys Imaging Radiat Oncol. 2021;17:53–7.

[21] Liang J, Scripes PG, Tyagi N, Subashi E, Wunner T, Cote N, et al. Risk analysis of the Unity 1.5 T MR-Linac adapt-to-position workflow. J Appl Clin Med Phys. 2023;24(3):1–15.

Deep learning-based tumor shape variation prediction of lung cancer patient's pre-radiation treatment response

Sangwoon Jeong[1], Kyungmi Yang[2,3], Youngyih Han[2,3], Hee Chul Park[2,3], Sang Hee Ahn[2*]

[1] Department of Health Sciences and Technology, SAIHST, Sungkyunkwan University, Seoul, Korea
[2] Department of Radiation Oncology, Samsung Medical Center, Seoul, Korea
[3] School of Medicine, Sungkyunkwan University, Seoul, Korea

Abstract Tumors decrease in size and change in shape with the progress of radiation therapy. Changes in tumor size are checked during treatment through Cone-beam Computed Tomography (CBCT). We predict CBCT based on deep learning using treatment plan information. A total of 376 data sets were used, and one data set was planning CT (pCT), dose map, RT structure, and CBCT. The study used a modified pix2pix model. Generator inputs are pCT, dose map, and RT structure, Generator output is predicted CBCT, and actual CBCT and predicted CBCT are distinguished through discriminator. In the PSNR, SSIM, and FID of the test sets, the averages were 10.03, 0.25, and 359.90, and the best performances were 14.11, 0.36, and 290.80. This study predicted tumor changes over the course of treatment using treatment plan information. Prediction of tumor changes can be used as a reference for when to change treatment plans.

1 Introduction

Radiation therapy (RT), a critical component in cancer treatment, varies in duration, ranging from as brief as 4 days to as extended as 30 days. Radiation therapy can cause changes in tumor morphology. The response of tumors to radiation therapy is influenced by a multitude of factors, including the tumor's initial size, the patient's unique physiological characteristics, and the type of concurrent chemotherapy administered. In [1], gross tumor volume (GTV) decreased by 24.7% in the first repeat computed tomography (CT) of lung cancer patients, and GTV decreased by 44.3% in the second repeat CT. Persisting with the initial treatment plan without accommodating these dynamic changes in tumor characteristics could potentially diminish therapeutic efficacy and inadvertently harm surrounding normal tissues. To address this, adaptive radiation therapy (ART) has been introduced, which considers these evolving tumor dynamics.

ART introduces an innovative approach by incorporating a planning CT (pCT) scan midway through the treatment process, enabling the re-establishment of the treatment plan. This revised plan is tailored to optimize treatment efficacy, taking into account changes in tumor volume and the patient's body during the course of therapy.

In a recent retrospective study, Hoppen, Lea, et al. [2] investigated the efficacy of applying daily online ART in patients with stage 3 non-small cell lung cancer. Their findings demonstrated significant improvements in local tumor control and a concurrent reduction in toxicity.

A notable advancement in this field is the Ethos™ (Varian Medical Systems, Palo Alto, CA), which facilitates treatment adaptation by generating new plans using in-device CT imaging and automated segmentation. However, this process is time-consuming, and the daily requirement for a new CT scan imposes a significant burden on the patient. In our approach, we leverage deep learning techniques to determine the most opportune moments for implementing ART.

In conventional ART methods, daily monitoring of tumor volume and anatomical changes is conducted using cone beam CT (CBCT) acquired during the treatment process. A new treatment plan is formulated when these changes are deemed significant. Our methodology utilizes RT structure, dose maps, pCT, and patient data derived from the treatment plan to generate CBCTs that correspond to treatment date alterations (Fig. 1). These generated CBCTs serve as viable substitutes for actual CBCTs in monitoring tumor size alterations and act as indicators for determining the optimal timing to establish a new treatment plan. Essentially, this implies that changes in tumor size can be preemptively identified using deep learning prior to administering radiation treatment.

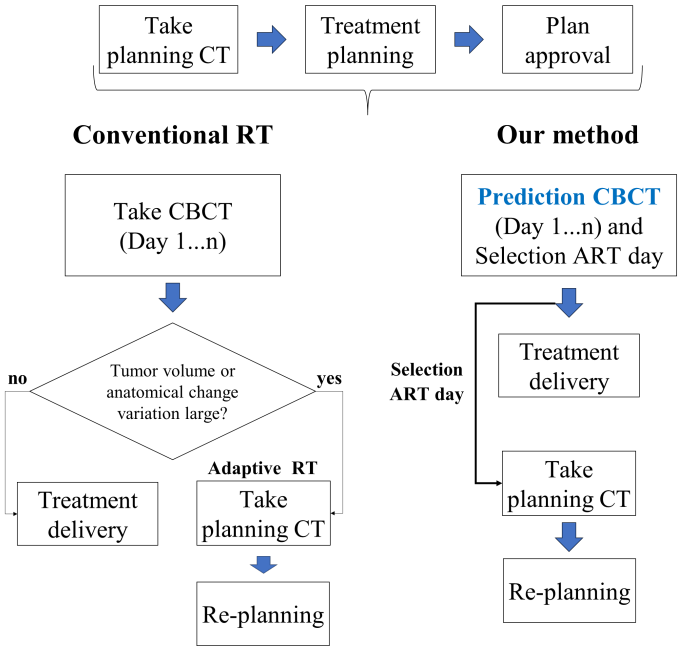


Figure 1. Comparison of the workflow of our method with a conventional radiotherapy workflow.

2 Materials and Methods

2.1 Dataset

The study protocol, including patient recruitment and data collection methods, was approved by the Institutional Review Board of the Samsung Medical Center (IRB number 2023-04-049-001), utilizing a retrospective data collection approach. Our dataset comprised data from 16 lung cancer patients treated at Samsung Medical Center. Among these patients, 9 underwent ART, resulting in a total of 25 treatment plans (16 initial plans, 9 adaptive plans) and 376 CBCT images. The treatment plan data encompasses pCT, RT structure, and dose map images. In the RT structure, only the Planning Target Volume (PTV) was utilized, and the dose map reflects the cumulative dose as of each treatment date. The treatment dates varied across patients, ranging from 5 to 20 days. Since a CBCT is acquired at each treatment session, the distribution of CBCTs corresponds to the treatment dates. For our analysis, three types of treatment plan images were used as input data, with CBCTs serving as the output data (Fig. 2).

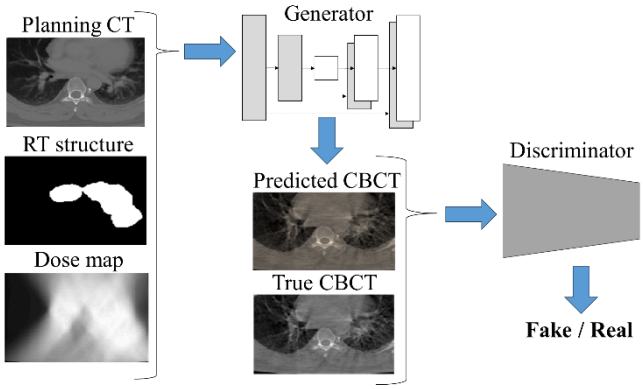


Figure 2. Deep learning-based CBCT prediction model.

2.2 Preprocessing

In the preprocessing stage, treatment plan data and CBCT were aligned at the same spatial location through image registration and then resampled to a uniform resolution of 256 x 256. For the z-axis, we selected 64 slices centered around the PTV. Subsequently, image cropping was performed to adjust the image size to 128 x 128 x 64, with a pixel size of 2.5 mm. The dimensions of the input data for our deep learning model were set to 3 x 128 x 128 x 64, while the output data size was maintained at 128 x 128 x 64. For testing purposes, the first day CBCT images of the 13 patients who underwent ART were used, and the remaining data served as the training dataset.

2.3 Network

For CBCT prediction, our model employs 'pix2pixCon', a variant of the 'pix2pix' framework, which itself is based on Generative Adversarial Networks (GANs). Pix2pix is widely recognized as an image-to-image translation model operating under the principles of conditional GANs [3]. The pix2pixCon model consists of two key components: a generator, responsible for creating CBCT images, and a discriminator, tasked with differentiating between generated and actual CBCT images. Notably, the generator in pix2pixCon adapts the modified U-net architecture, while the discriminator employs a multi-scale approach, utilizing both original and half-scale discriminators. The loss functions employed in this model include L1 loss, the Structural Similarity Index (SSIM), and cGAN loss. For optimization, the Adam optimizer was used, with a learning rate designed to decrease gradually as the training progresses.

$$Loss\ function = cGAN + L1 + SSIM$$

2.4 Evaluation

The results are evaluated through SSIM, Peak signal-to-noise ratio (PSNR), and Fréchet Inception Distance (FID) of actual CBCT and predicted CBCT. The ideal value, PSNR is the maximum value minus the minimum value, SSIM is 1, and FID is 0.

3 Results

Figure 3 visualizes the three central slices from the test dataset. While the predicted CBCT images appear somewhat blurrier compared to the actual CBCTs, they accurately replicate the CT's shape, including the tumor's morphology. The region of interest (ROI) difference effectively mirrors the tumor's Hounsfield Unit (HU) values seen in the actual CBCTs.

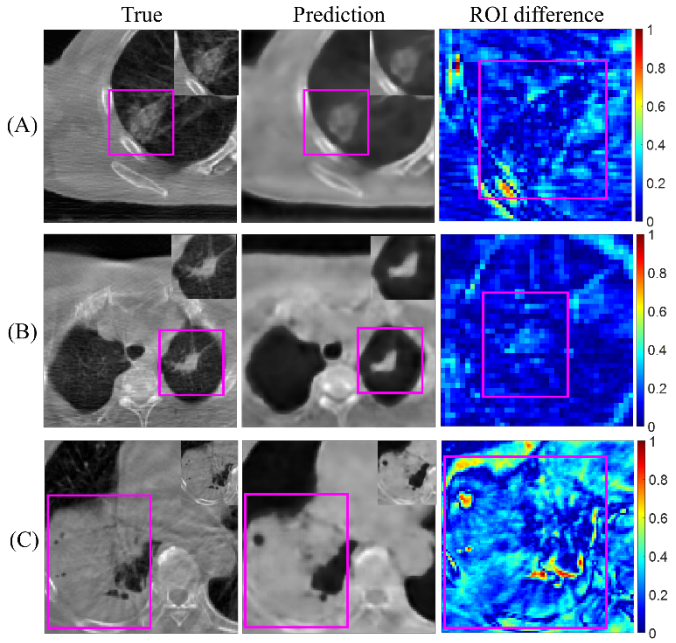


Figure 3. Comparison of predicted CBCT and true CBCT.

Table 1 presents the quantitative evaluation results for the entire test dataset. Among the cases, Case 9 exhibited the highest PSNR at 14.12. Case 1 scored the best in terms of SSIM and FID. However, it is noteworthy that none of the cases achieved values close to the ideal benchmark.

Table 1. PSNR, SSIM, and FID analysis results for the all-test case.

Case	PSNR	SSIM	FID
1	12.89	0.36	290.8
2	7.07	0.17	369.13
3	10.86	0.26	347.28
4	12.41	0.29	414.59
5	4.13	0.14	328.86
6	9.62	0.23	367.97
7	8.38	0.2	367.7
8	10.8	0.29	422.86
9	14.12	0.32	329.98
Mean	10.03	0.25	359.91
STD	2.95	0.07	39.38
Ideal	48	1	0

4 Discussion

In our study, we utilized treatment plan data and the pix2pixCon model to prediction CBCT images over the course of treatment. As illustrated in Figure 3, the predicted CBCT shape in Cases (A) and (B) closely resembles the actual CBCT, with a notably smaller ROI discrepancy in the tumor area compared to surrounding regions. However, in Case (C), characterized by a larger tumor size, a significant difference in ROI is observed. Our findings suggest that when the tumor shape is distinctly visible, the predicted image bears a close resemblance to the actual image. The efficacy of CBCT prediction using the pix2pixCon model hinges on selecting the appropriate tumor type. Despite similarities in overall tumor shape even in instances of large tumors, our results indicate a need for further model refinement and comprehensive data analysis.

Table 1 presents a comparative analysis of the performance metrics across the entire test dataset. It is evident from the data that in all cases, the PSNR and SSIM values did not surpass 50% of the ideal benchmark. Additionally, the FID scores are reported in the three-digit range. The evaluation includes not only the tumor but also the surrounding area. As can be seen from the ROI difference, there is a large difference in the area surrounding the tumor. If only the tumor area is compared, it appears that better results will be obtained.

Given that the efficacy of using predicted CBCT in actual treatment hinges on its congruence with the actual CBCT images, the SSIM, as an evaluative metric, holds significant importance. To enhance performance, it is imperative to consider modifications to the model's structure and to incorporate additional patient data.

This study is subject to certain limitations. As depicted in Figure 3, the diversity in tumor shapes poses a challenge; accurate prediction becomes difficult without training on a wide variety of tumor morphologies. Predominantly small-tumor datasets hinder our model's ability to accurately predict size reductions in larger tumors. Moreover, in cases where the tumor shape is indistinct, there is a risk of it being indistinguishable from the lung parenchyma. To overcome these challenges and enhance prediction accuracy, we aim to employ additional datasets and ensure a more balanced data representation in future studies.

Nevertheless, applying our research to ART could enable the prediction of tumor responses to radiation doses before treatment. This would allow for more precise dosing prescriptions and facilitate the implementation of patient-specific, customized treatments. Previous research has explored the use of time-series models like long short-term memory [4] and recurrent neural network [5] for predictive purposes. However, our study represents a pioneering effort in employing a generative model, specifically a GAN, for the prediction of radiation dose responses. This novel approach marks a significant advancement in the field by integrating dose information through a GAN model. Personalized Ultra fractionated Stereotactic Adaptive Radiotherapy (PULSAR) has recently shown excellent effectiveness in treating tumors along with immunotherapy [6]. The prediction of tumor size change in our study can be used as an indicator of tumor change for using PULSAR in actual treatment.

In this study, 9 CBCT images from the first day of ART treatment were employed as test data. These initial day CBCTs are hypothesized to closely resemble the treatment pCT, primarily because the treatment is conducted based on the previous day's plan information. Going forward, our plan involves using the predicted CBCT for formulating subsequent treatment plans and conducting ART treatments along with Dose-Volume Histogram (DVH) analysis. Additionally, we aim to elevate the quality of CBCT images, to the level of pCT by incorporating a deep learning-based image quality enhancement model. This approach will enable us to conduct a comprehensive analysis of the treatment plan and DVH for CBCTs across all treatment days, thereby establishing quantitative indicators to determine the most suitable dates for implementing ART.

5 Conclusion

This study predicted tumor changes over the course of treatment using treatment plan information. Prediction of tumor changes can be used as a reference for when to change treatment plans. Further study will improve the accuracy by increasing the number of data, develop better data processing methods, and refining the prediction model.

Financial Support

This research was supported by the National Research Foundation of Korea grant funded by the Korean government (MSIT) 2022R1F1A1070271.

References

[1] Fox J, Ford E, Redmond K, Zhou J, Wong J, Song DY. Quantification of tumor volume changes during radiotherapy for non–small-cell lung cancer. International Journal of Radiation Oncology* Biology* Physics. 2009 Jun 1;74(2):341-8.

[2] Hoppen, L., Sarria, G. R., Kwok, C. S., Boda-Heggemann, J., Buergy, D., Ehmann, M., ... & Fleckenstein, J. (2023). Dosimetric benefits of adaptive radiation therapy for patients with stage III non-small cell lung cancer. Radiation Oncology, 18(1), 1-11.

[3] Isola P, Zhu JY, Zhou T, Efros AA. Image-to-image translation with conditional adversarial networks. InProceedings of the IEEE conference on computer vision and pattern recognition 2017 (pp. 1125-1134).

[4] Lee, D., Hu, Y. C., Kuo, L., Alam, S., Yorke, E., Li, A., ... & Zhang, P. (2022). Deep learning driven predictive treatment planning for adaptive radiotherapy of lung cancer. Radiotherapy and Oncology, 169, 57-63.

[5] Li, R., Roy, A., Bice, N., Kirby, N., Fakhreddine, M., & Papanikolaou, N. (2021). Managing tumor changes during radiotherapy using a deep learning model. *Medical physics*, *48*(9), 5152-5164.

[6] Moore C, Hsu CC, Chen WM, Chen BP, Han C, Story M, Aguilera T, Pop LM, Hannan R, Fu YX, Saha D. Personalized ultrafractionated stereotactic adaptive radiotherapy (PULSAR) in preclinical models enhances single-agent immune checkpoint blockade. International Journal of Radiation Oncology* Biology* Physics. 2021 Aug 1;110(5):1306-16.

Deep learning for in vivo EPID dosimetry classification: Relating gamma analysis and DVH metrics for lung cancer patients

Femke Vaassen[1], Sebastiaan Nijsten[1], Richard Canters[1], Frank Verhaegen[1], Cecile J.A. Wolfs[1]
[1] Department of Radiation Oncology (Maastro), GROW School for Oncology and Reproduction, Maastricht University Medical Centre+, Maastricht, The Netherlands

Abstract
A deep learning classification model was trained on 2D γ-maps to relate 2D γ-analysis to clinically relevant dose volume histogram (DVH) metrics for in vivo treatment verification using electronic portal imaging device (EPID) dosimetry. The aim of the model was to classify γ-maps into two categories: 'error' and 'no error'. It is hypothesized that it should be possible to detect 5% 3D dose differences with this method. Measured and predicted portal dose images for lung cancer patients were compared per beam using γ-analysis. Ground truth classification was based on DVH metrics derived from 3D dose recalculations on daily cone-beam CT. The γ-maps together with their corresponding label were used as input for convolutional neural networks (CNNs). Results showed that training performance was good, but indicated overfitting, seen by lower performance on the validation set. A ground truth classification based on exceeding clinical constraints for OARs resulted in the best performing model. This study showed that applying deep learning for relating 2D γ-results to DVH metrics has the potential for implementation of decision protocols in clinical practice, but the CNN architecture implemented here is unable to result in useful models.

1 Introduction

For lung cancer patients, anatomical changes during radiotherapy treatment can result in significant differences between planned and delivered dose [1]. In around 23% of lung cancer patients, anatomical changes such as pleural effusion, atelectasis, or a tumor shift occur at some point during the course of radiotherapy [2]. This high occurrence rate of anatomical changes (which could result in dose differences in target coverage of more than 5%), together with the complex nature of today's radiotherapy techniques, makes it essential to perform in-treatment, i.e. in vivo, verification of delivered dose to ensure treatment quality for the patient [3,4]. The electronic portal imaging device (EPID) has previously been shown to accurately and efficiently perform in vivo dose verification in 2D, 3D or even 4D [5-7].

Measured EPID doses can be compared to predicted/planned dose distributions using a gamma (γ) analysis. The result of this comparison is usually summarized in one or a few numbers, providing fast decision making based on fixed thresholds using gamma pass rate. However, this leads to loss of 2D and/or 3D spatial, dosimetric and time-resolved information. Several studies have shown previously that threshold-based methods have difficulty relating gamma pass rates to clinically used metrics such as dose volume histogram (DVH) values [8-11].

Artificial intelligence (AI) methods, more specifically machine and deep learning techniques, have been studied extensively, also for use in radiotherapy quality assurance (QA) [12-14]. These approaches use large amounts of (multidimensional) input data, and learn features from this data. Application of AI techniques to treatment verification data is promising, as the full gamma analysis result could be used instead of the condensed form. Focus has not only been on detecting treatment-related errors based on gamma analysis and corresponding gamma pass rate, but also on identifying the causes of these errors using deep learning (DL) methods [15-17].

In this study, a deep learning classification model is trained on 2D γ-maps, investigating the possibility to relate 2D γ-analysis to clinically relevant (changes in) DVH metrics for in vivo treatment verification using EPID dosimetry. This would eliminate the need for a 3D dose recalculation based on a either the planning CT, which is lengthy and time-consuming on itself, or on a cone-beam CT (CBCT), which might not always be present. Next to this, EPID dosimetry gives information on the dose delivery itself. The aim of the model was to classify the 2D γ-maps into two categories: 'error' and 'no error'. It is hypothesized that it should be possible to detect 5% 3D dose differences with this method.

2 Materials and Methods

2.1 Input data

EPID measurements from 3086 fractions of 363 lung cancer patients (380 unique treatment plans), treated with VMAT, were converted to portal dose images (PDIs) using in-house developed software [7]. Predicted transit PDIs were created based on the treatment plan and planning CT. Measured and predicted PDIs were compared per arc in a time-integrated fashion using γ-analysis with 5%/3mm criteria, resulting in 5844 2D γ-maps per set of gamma criteria. The γ-maps were clipped to -3 and +3, cropped using a 10% low dose threshold, resized to 128x128 pixels, and normalized using min/max normalization.

The full dataset was split into training (70%), validation/hyperparameter optimization (10%) and hold-out test sets (20%), with all images of one patient belonging to the same set. Next to this, 5-fold cross validation was also performed in the training and validation sets combined to evaluate model robustness.

2.2 Ground truth classification

The ground truth (GT) classification ('error' vs. 'no error') was based on DVH metrics derived from 3D Monte Carlo dose recalculations (XVMC, [18]) on daily CBCT images, including contour propagation from planning CT to CBCT using Mirada Workflow Box (Mirada Medical Ltd., Oxford, UK). Different GT classification methods (Table 1) were

implemented for both the organs-at-risk (OARs) and the clinical target volume (CTV), based on exceeding the clinical planning constraints as used during treatment planning. When this planning constraint was already exceeded in the treatment plan, the value of the DVH metric at start of treatment was used as action threshold. In this way, it is assumed that the dose should not deviate more than the planning constraint achieved. A fraction was classified in the 'error' class when one or more of the DVH metrics exceeded the specified action threshold, and in the 'no error' class when all DVH metrics were within the constraint. In Table 2, the number of fractions belonging to the 'error' and 'no error' classes for all of the GT classifications is shown.

Table 1 - Overview of the DVH metrics and their limits used in this work. Dmax values correspond to a volume of 0.03cc.

Structure	Metric	Limit
Clinical constraints OARs		
Lungs-GTV	D_{mean}	< 20 Gy
	V_{5Gy}	< 60 %
Heart	D_{mean}	< 10 Gy
	D_{max}	< 66 / 76 Gy[a]
Mediastinal envelope	D_{max}	< 66 / 76 Gy[a]
Esophagus	D_{max}	< 66 / 76 Gy[a]
Spinal cord	D_{max}	< 51.9 / 54 Gy[a]
Brachial plexus	D_{max}	< 66 / 76 Gy[a]
Absolute dose differences OARs		
Per OAR as specified above	ΔD_{mean}	> 2 Gy
	ΔD_{max}	> 5 Gy
Relative dose differences CTV		
CTV	$\Delta D_{95\%}$	> 2 %
CTV	ΔD_{mean}	> 5%
Combination CTV or OARs		
CTV	$\Delta D_{95\%}$	> 2 %
CTV	$\Delta D_{95\%}$	> 2 Gy
Per OAR as specified above	ΔD_{mean}	> 2 Gy
	ΔD_{max}	> 5 Gy

[a]*Different fractionation schemes: 24x2.75 Gy / 33x2 Gy*

2.3 Deep learning model

The γ-maps together with their corresponding GT label were used as input for convolutional neural networks (CNNs). The CNN consisted of multiple blocks of two convolutional layers and a max pooling layer. Rectified linear unit (ReLU) activation was used for all these layers. The exact CNN architecture and corresponding hyperparameters were optimized using Bayesian optimization [19]. Binary Cross Entropy (BCE) is the commonly used loss function in binary classification tasks. However, in this study, class imbalance was present and Sigmoid Focal Cross Entropy (SFCE) loss may be more suitable for binary classification in imbalanced datasets [20]. The performance of these two loss functions was compared. The Adam optimizer was used to optimize the model loss. The final fully-connected layer consisted of one single node with sigmoid activation, giving the output as a number between 0 and 1 as the probability that the input image belongs to the 'error' class. Early stopping was applied based on validation loss to prevent overfitting on the training set. Results of model training were evaluated using the area under the curve (AUC) of the precision-recall curve (AUC_PR).

Table 2 - The number of fractions per ground truth classification method belonging to the 'error' and 'no error' classes.

GT classification	# fractions in 'error' class (%)	# fractions in 'no error' class (%)
Clinical constraints OARs	1930 (62.5%)	1156 (37.5%)
Absolute dose difference OARs (ΔD_{mean} > 2Gy, ΔD_{max} > 5Gy)	1166 (37.8%)	1920 (62.2%)
Relative dose difference CTV $\Delta D_{95\%}$ > 2 %	1475 (47.8%)	1611 (52.2%)
Relative dose difference CTV ΔD_{mean} > 5 %	235 (7.6%)	2851 (92.4%)
Combination: CTV $\Delta D_{95\%}$ > 5% or CTV $\Delta D_{95\%}$ > 2Gy or OARs (ΔD_{mean} > 2Gy and ΔD_{max} > 5Gy)	1245 (40.3%)	1841 (59.7%)

3 Results

Results showed that training and validation performance of the CNN was good for the GT classification based on clinical constraints using a 5%/3mm γ-criteria, but also tended towards overfitting (Figure 1). SFCE as loss function showed better performance on the validation set compared to BCE as loss function. 5-fold cross validation results showed considerable overfitting on the training set.

Confusion matrices comparing GT DVH-classification with CNN-classification using SFCE as loss function and 5%/3mm γ-criteria are shown in Figure 2 for the training and validation sets. For the classification threshold of 0.5, it can be seen that the CNN classification results in a high number of false positives (i.e. a fraction is classified in the 'error' class, when the DVH classification is 'no error'), both for the training and the validation set. This number of false positives is lower when the classification threshold is optimized. However, still a significant number of fractions is wrongly classified.

For the other GT classifications shown in Figure 1 results showed moderate to bad performance, especially on the validation set. This shows that there is considerable overfitting of the models on the training set, leading to low AUC_PR values on the validation set.

4 Discussion

The results of this study showed that it could be possible to relate γ-maps and DVH metrics using a CNN, although there are still some limitations that need to be considered.

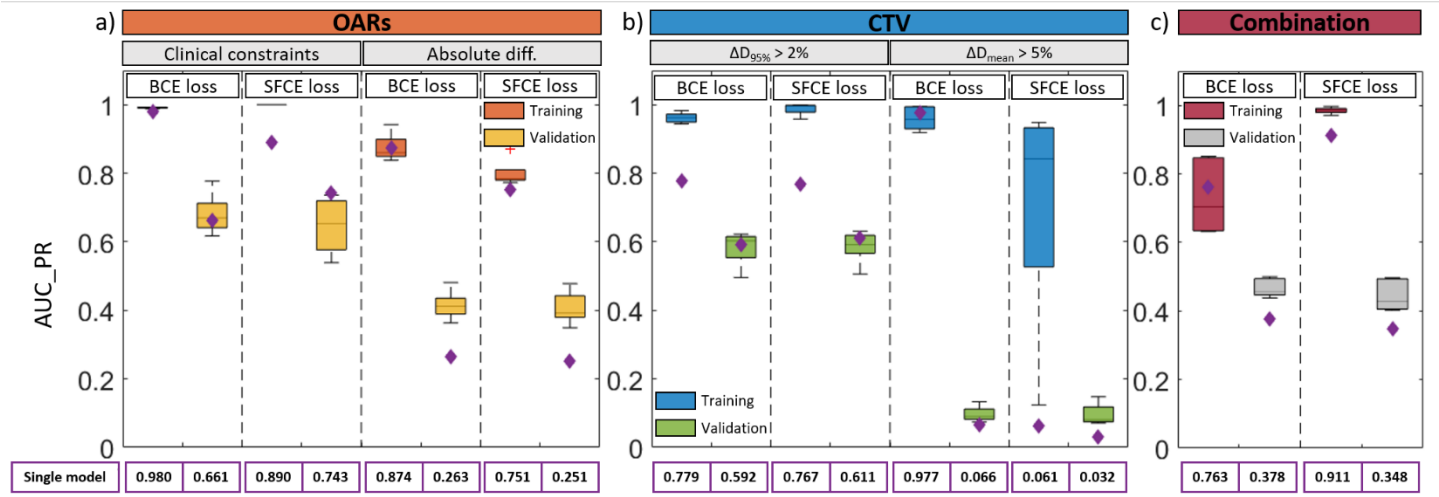


Figure 1 - Results of model performance on the training and validation set, shown here for the two different loss functions and only for the 5%/3mm γ-criteria. a) GT classification for the OARs, b) GT classification for the CTV, c) GT classification for the combination of OARs and CTV (as specified in Table 2). Model performance is presented in terms of AUC_PR for the single model as purple diamonds and for the 5-fold cross validation as boxplots.

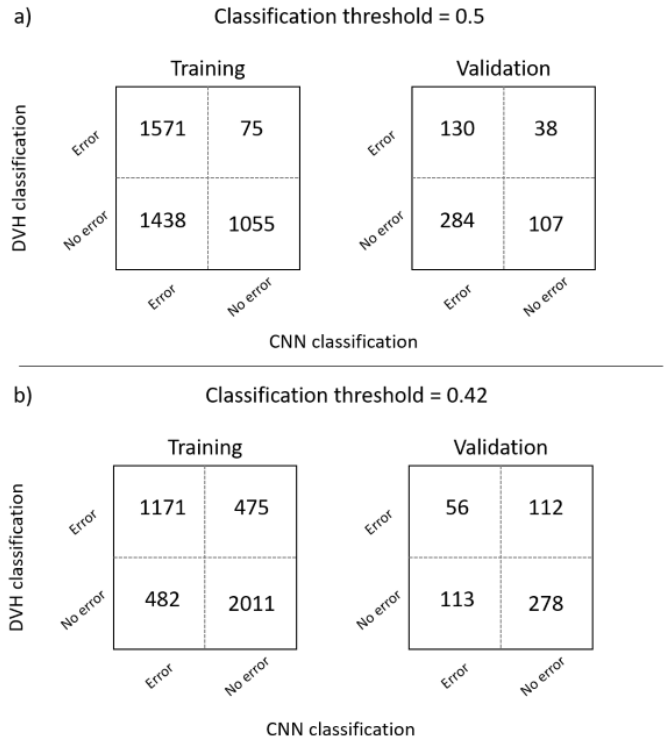


Figure 2 - Confusion matrices comparing the CNN-classification with the GT DVH-classification for a) the commonly used classification threshold 0.5 and b) an optimized classification threshold (0.42) based on equal precision and recall.

The performance of the models shown here was reasonably good on the training set, but lower on the validation set. This illustrates that overfitting on the training data occurred, even though early stopping was applied during the training process when the validation loss did not decrease any more for 20 epochs. Performance on an independent hold-out test set was not shown, but it is hypothesized that model performance would drop even further.

To prevent overfitting during model training, dropout regularization was already applied. The dropout fraction for the models shown in this study ranged between 0.1 and 0.5. This range shows also dropout fractions of relatively low numbers, indicating that the dropout as implemented currently did not improve the model training. Other methods to prevent overfitting could be considered, such as using a different, more complex network architecture.

In clinical practice, DVH metrics are commonly used to assess the dosimetric impact of anatomical changes and make clinical decisions. However, this process of decision making often remains subjective, depending on the preference of the radiation oncologist for a specific patient. In this study, several methods for GT classification into the 'error' class were implemented and tested. Using the clinical constraints appeared to be the method with the best-defined action thresholds, but a disadvantage of this approach is that large dose deviations may go undetected when the dose to a specific organ is still under the clinical constraint. Different approaches for GT classification with different thresholds were implemented in this study, e.g. relative dose differences for the clinical target volume or a combination between dose differences to the OARs and the CTV. However, these approaches did not lead to useful models compared to the hypothesis.

Next to using DVH metrics for GT classification, different approaches could be considered, such as biological metrics. Tumor control probability (TCP) and normal tissue complication probability (NTCP) might be more suitable, and it has been shown that dosimetric predictors of toxicity can be identified and can lead to clinically useful dosimetric thresholds [21].

Another limitation of this study is that the GT classification on DVH metrics is performed per fraction, whereas the PDIs are acquired per beam. Classification of these PDIs is thus also performed per beam, where each beam in one fraction received the same error/no error label. However, it could be possible that a dosimetric error would only be visible in one beam. Next to that, for VMAT arcs, the PDIs are a 2D summation over all angles used in the arc, blurring out the high dose region and penumbra. Consequently, dose differences can average out over the arc, resulting in no dose

difference in the PDI when there is a dose difference in a certain DVH metric. 3D measurements or even time-resolved EPID dosimetry might provide additional information that could improve the error classification. In a recent study, it was also shown that more simple dose comparison methods such as absolute or relative dose differences instead of gamma analysis could be a more optimal input for a deep-learning network to detect errors [22].

Multilabel classification was implemented using multiple output labels to try to further improve the CNN classification. Six of the OARs with their corresponding clinical constraints were used, and the final dense layer consisted of six nodes with sigmoid activation. However, evaluation results (not shown) were worse compared to the models reported here. There are also other deep learning approaches that might be more suitable for relatively small datasets, such as transfer learning. With this approach, a CNN that is trained on a different but large dataset, is used and only the last layers are retrained on the smaller dataset of interest. This approach has proven to be beneficial and useful for applying deep learning to small datasets [23]. This approach has been implemented in this study using a CNN trained to identify error types for lung cancer patients (evaluation results not shown), however this did not lead to higher AUC_PR scores [17]. More research is needed to study other models that could be used for transfer learning.

5 Conclusion

This study showed that applying deep learning for relating 2D γ-results to DVH metrics has potential for implementation of decision protocols in clinical practice. However, the CNN architecture implemented here is unable to result in useful models. More research is needed to prevent overfitting by looking into more complex network architectures or using different dose comparisons as input data.

References

[1] Rozendaal RA, Mijnheer BJ, Van Herk M, Mans A. In vivo portal dosimetry for head-and-neck VMAT and lung IMRT: linking γ-analysis with differences in dose-volume histograms of the PTV. Radiother Oncol 2014;112:396–401. doi: 10.1016/J.RADONC.2014.03.021.

[2] Møller DS, Khalil AA, Knap MM, Hoffmann L. Adaptive radiotherapy of lung cancer patients with pleural effusion or atelectasis. Radiother Oncol. 2014;110(3):517-522. doi: 10.1016/j.radonc.2013.10.013.

[3] Persoon LC, Egelmeer AG, Ollers MC, et al. First clinical results of adaptive radiotherapy based on 3D portal dosimetry for lung cancer patients with atelectasis treated with volumetric-modulated arc therapy (VMAT). Acta Oncol. 2013;52(7):1484-1489. doi: 10.3109/0284186X.2013.813642.

[4] Olaciregui-Ruiz I, Beddar S, Greer P, et al. In vivo dosimetry in external beam photon radiotherapy: Requirements and future directions for research, development, and clinical practice. Phys Imaging Radiat Oncol 2020;15:108–16. doi: 10.1016/J.PHRO.2020.08.003.

[5] Mans A, Wendling M, McDermott LN, et al. Catching errors with in vivo EPID dosimetry. Med Phys 2010;37:2638–44. doi:10.1118/1.3397807.

[6] Podesta M, Nijsten SMJJG, Persoon LCGG, et al. Time dependent pre-treatment EPID dosimetry for standard and FFF VMAT. Phys Med Biol 2014;59:4749–68. doi: 10.1088/0031-9155/59/16/4749.

[7] Nijsten SMJJG, Van Elmpt WJC, Jacobs M, et al. A global calibration model for a-Si EPIDs used for transit dosimetry. Med Phys 2007;34:3872–84. doi: 10.1118/1.2776244.

[8] Nelms BE, Zhen H, Tomé WA. Per-beam, planar IMRT QA passing rates do not predict clinically relevant patient dose errors. Med Phys. 2011;38(2):1037-1044. doi: 10.1118/1.3544657.

[9] Zhen H, Nelms BE, Tome WA. Moving from gamma passing rates to patient DVH-based QA metrics in pretreatment dose QA. Med Phys. 2011;38(10):5477-5489. doi: 10.1118/1.3633904.

[10] Stasi M, Bresciani S, Miranti A, et al. Pretreatment patient-specific IMRT quality assurance: a correlation study between gamma index and patient clinical dose volume histogram. Med Phys. 2012;39(12):7626-7634. doi: 10.1118/1.4767763.

[11] Persoon LC, Podesta M, Hoffmann L, et al. Is integrated transit planar portal dosimetry able to detect geometric changes in lung cancer patients treated with volumetric modulated arc therapy? Acta Oncol. 2015;54(9):1501-1507. doi: 10.3109/0284186X.2015.1061213.

[12] Pillai M, Adapa K, Das SK, et al. Using Artificial Intelligence to Improve the Quality and Safety of Radiation Therapy. J Am Coll Radiol 2019;16:1267–72. doi: 10.1016/J.JACR.2019.06.001.

[13] Osman AFI, Maalej NM. Applications of machine and deep learning to patient-specific IMRT/VMAT quality assurance. J Appl Clin Med Phys 2021;22:20–36. doi:10.1002/ACM2.13375.

[14] Bedford JL, Hanson IM. A recurrent neural network for rapid detection of delivery errors during real-time portal dosimetry. Phys Imaging Radiat Oncol 2022;22:36–43. doi: 10.1016/J.PHRO.2022.03.004.

[15] Kimura Y, Kadoya N, Tomori S, et al. Error detection using a convolutional neural network with dose difference maps in patient-specific quality assurance for volumetric modulated arc therapy. Phys Med 2020;73:57–64. doi: 10.1016/J.EJMP.2020.03.022.

[16] Kimura Y, Kadoya N, Oku Y, et al. Error detection model developed using a multi-task convolutional neural network in patient-specific quality assurance for volumetric-modulated arc therapy. Med Phys 2021;48:4769–83. doi: 10.1002/MP.15031.

[17] Wolfs CJA, Canters RAM, Verhaegen F. Identification of treatment error types for lung cancer patients using convolutional neural networks and EPID dosimetry. Radiother Oncol 2020;153:243–9. doi: 10.1016/J.RADONC.2020.09.048.

[18] van Elmpt WJ, Nijsten SM, Schiffeleers RF, et al. A Monte Carlo based three-dimensional dose reconstruction method derived from portal dose images. Med Phys. 2006;33(7):2426-2434. doi:10.1118/1.2207315.

[19] Akiba, T., Sano, S., Yanase, T., Ohta, T., Koyama, M. Optuna: A Next-generation Hyperparameter Optimization Framework. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining 2019;2623–2631. doi: 10.1145/3292500.3330701.

[20] Lin TY, Goyal P, Girshick R, He K, Dollar P. Focal Loss for Dense Object Detection. IEEE Trans Pattern Anal Mach Intell 2020;42:318–27. doi: 10.1109/TPAMI.2018.2858826.

[21] Ladbury C, Li R, Danesharasteh A, et al. Explainable Artificial Intelligence to Identify Dosimetric Predictors of Toxicity in Patients with Locally Advanced Non-Small Cell Lung Cancer: A Secondary Analysis of RTOG 0617. Int J Radiat Oncol Biol Phys. 2023;117(5):1287-1296. doi: 10.1016/j.ijrobp.2023.06.019.

[22] Wolfs CJA, Verhaegen F. What is the optimal input information for deep learning-based pre-treatment error identification in radiotherapy?. Phys Imaging Radiat Oncol. 2022;24:14-20. doi: 10.1016/j.phro.2022.08.007

[23] Kim HE, Cosa-Linan A, Santhanam N, et al. Transfer learning for medical image classification: a literature review. BMC Med Imaging 2022;22(1):69. doi: 10.1186/s12880-022-00793-7.

Adaptive Radiotherapy Dose Prediction on Head and Neck Cancer Patients with a Multi-Headed U-net Deep Learning Architecture

Hui-Ju Wang, Austen Maniscalco, David Sher, Mu-Han Lin, Steve Jiang, and Dan Nguyen

Medical Artificial Intelligence and Automation Laboratory, University of Texas Southwestern Medical Center, Dallas, Texas, USA
Department of Radiation Oncology, University of Texas Southwestern Medical Center, Dallas, Texas, USA

Abstract Adaptive radiotherapy (ART) optimizes radiation treatment based on the changing anatomy and tumor during the treatment course. Current deep-learning prediction models do not fully utilize pre-existing ART information, which can result in a loss of physician/patient-tailored intents. To address this, we introduce a Multi-Headed U-net (MHU-net) that incorporates intents and compares it to a standard single-headed U-net (SHU-net). The MHU-net was trained to predict adaptive session dose distributions using data from the adaptive session as the primary head input and pre-plan as the secondary head input. We found high agreement between ground truth dose and MHU-net, resulting in a statistically significant reduction in MSE from 0.0048 (SHU-net) to 0.0038 (MHU-net) on 10 test patients (p-val = 0.006). In conclusion, our study demonstrates the feasibility of incorporating a pre-plan as the physician's preference for the patient to tailor the dose prediction further.

++++

1 Introduction

In the field of radiation therapy (RT) for head and neck (H&N) cancer, advancements such as intensity-modulated radiation therapy (IMRT) and Volumetric Modulated Arc Therapy (VMAT) have significantly enhanced treatment precision, creating highly conformal dose distributions on tumors [1,2]. However, these innovations often rely on plans derived from initial planning scans, overlooking factors that could accommodate structural and spatial changes during treatment. Studies have shown that patients often experience anatomical changes, encompassing factors like weight loss, muscle atrophy, or normal tissue edema, as well as tumor shrinkage, normal tissue reduction, and positional shift of certain structures throughout their treatment course [3,4]. These changes introduce uncertainties in physician-defined margins around target volumes, leading to larger irradiated volumes, heightened toxicity, and the risk of insufficient dose delivery to organs at risk, thereby diminishing treatment effectiveness and increasing the potential for side effects [5]. In response to this challenge, the innovative treatment paradigm of **Adaptive Radiotherapy** (ART) has emerged. ART is a revolutionary approach in radiation therapy, enabling the continuous adaptation of patients' treatment plans to accommodate changing anatomical conditions and reoptimizing the plan by adjusting target structures and organs-at-risk structures throughout radiation therapy [6]. This ensures accurate dose delivery to the target while minimizing radiation exposure to healthy tissues. ART adaptation provides a more personalized and effective treatment strategy, one that is specifically designed to address and adapt to the minimal changes that occur during treatment. Despite ongoing ART progress, the emergence of artificial intelligence (AI) and deep learning (DL) has been a notable development in recent years. In radiation therapy, DL-based dose prediction has garnered significant attention due to its potential to accurately predict entire dose distributions [7,8,9]. DL models can learn intricate patterns and relationships from vast data sets, allowing them to estimate radiation doses received by different areas of a patient's anatomy. However, a critical gap remains: current DL models for H&N dose prediction do not incorporate valuable pre-existing information provided by ART. This omission can result in the loss of essential physician and patient-tailored intentions, which are crucial for personalized treatment.

To address this challenge, we propose to incorporate the physician and patient-tailored intents using a Multi-Headed input U-net (MHU-net) and compare its performance with a standard Single-Headed input U-net (SHU-net) to assess the impact of these features on dose prediction accuracy and effectiveness. In leveraging these advanced architectures, we aim to bridge the gap between deep learning models and the valuable pre-existing information provided by ART. The stander U-net architecture has gained popularity in medical imaging tasks, such as dose prediction, due to its ability to capture image features at both global and local [10]. The MHU-net advances the U-net design by integrating multiple input heads, each tasked with predicting specific aspects of the dose distribution, thereby incorporating pre-existing information from ART and enabling the transmission of physician and patient-tailored intents to the next adaptive session.

2 Materials and Methods

2.1 HN cancer imaging dataset

We acquired H&N patients who were treated with ART on the Varian Ethos machine. Each plan consisted of important components such as CT scans, delineated contours of the planning target volume (PTV), 42 organ-at-risks (OARs) of interest, and the final approved plan dose. We had a total of 43 patients with 95 radiotherapy plans, which were divided into three groups: 30 patients for model training, three for validation, and 10 for testing purposes. Our dataset consisted of patient data that included both pre-plans and adaptive session plans.

2.2 Deep learning model architecture

The structural design of our model utilized the widely popular U-net network architecture as a baseline, named SHU-net. It consisted of an encoder on the left half and a decoder on the right half of its architecture. Building upon the U-Net structure with repeated 3x3x3 convolutions, rectified linear unit activations, four pooling operations, and spatial dropout at a rate of 0.1. In-depth details of our proposed MHU-net. model is depicted in Figure 1. The MHU-net incorporated the adaptive session's contours and a signed distance map of each contour as input into its primary head. Additionally, it utilized pre-plan contours, signed distance maps, and dose distribution as input into its secondary head, which then incorporated the data into the primary U-net. Finally, we utilized adaptive session data to predict the adaptive dose distribution. The MHU-net was trained to predict the adaptive session's dose distribution over 1000 epochs using an ADAM optimizer, with a learning rate of 0.001, and the mean squared error (MSE) as the loss function.

2.3 Evaluation

From these 10 patients, their ART treatment plans were withheld from training as the test dataset for the evaluation of dose prediction performance between the SHU-net and MHU-net models. SHU-net and MHU-net dose prediction was compared against the ground truth dose distribution as originally generated during the patient's course of treatment. Mean absolute percent error, as shown in equation (1), evaluated the dose differences between a model's prediction and the ground truth.

$$Mean\ Absolute\ Percent\ Error = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{D_{GT}-D_{Predicted}}{D_{Highest\ Prescibtion}}\right| * 100 \quad (1)$$

Furthermore, we conducted a detailed evaluation of the highest dose PTV, including parameters such as D95%, D98%, D99%, and homogeneity as shown in equation (2).

$$Homogeneity = \frac{D_{2\%}-D_{98\%}}{D_{50\%}} \quad (2)$$

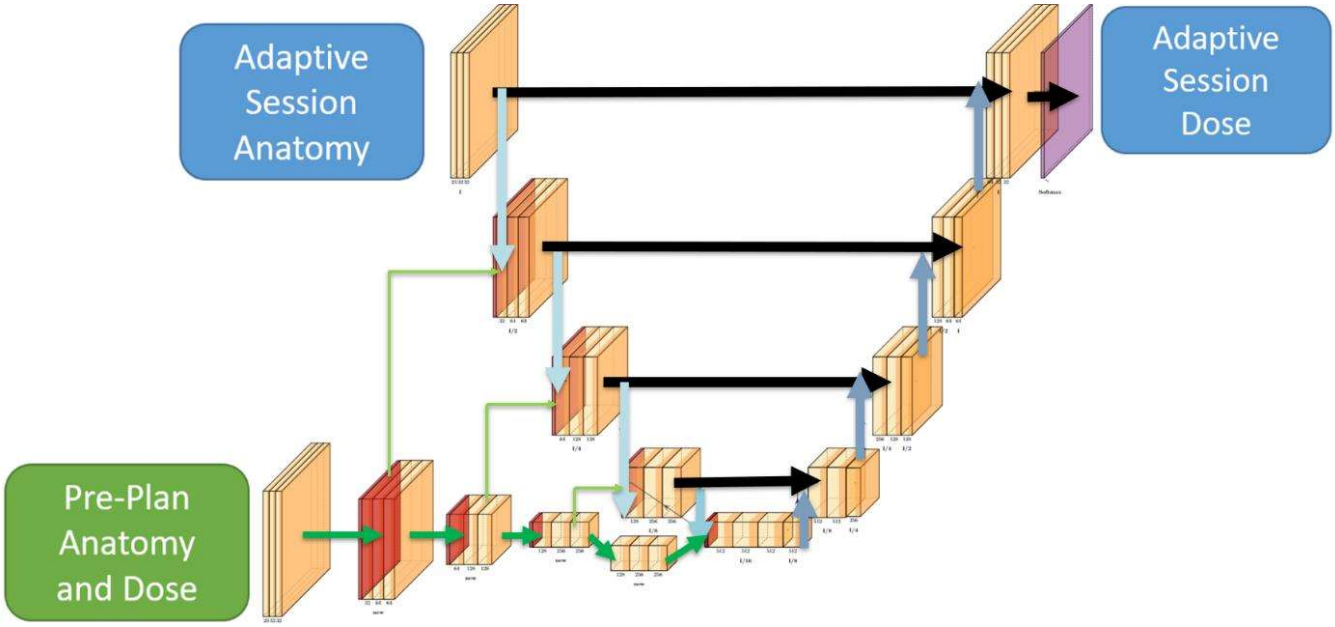


Figure 1: Multi-Headed U-net Architecture. The primary head processes adaptive session contours and signed distance maps during the contracting path to the bottleneck, indicated by light blue arrows. Simultaneously, the secondary head takes pre-plan contours, signed distance maps, and dose distribution as input, following the green arrows to the primary head and the bottleneck. The final step is to predict the adaptive session dose.

3 Results

3.1 Results of Dose Wash and Dose Volume Histogram

In Figure 2, an illustration of a dose wash from test patients is presented. The left columns display the ground truth clinical dose, while the subsequent columns exhibit the dose predictions generated by the SHU-net and the MHU-net, respectively. Upon careful analysis, it is evident that the MHU-net's predicted dose distributions closely correspond with the ground truth planned dose distributions, showcasing uniformity in various trade-offs between the PTV and OARs. Particularly noteworthy is the closer alignment observed between the predicted (dotted) and ground truth planned (solid) DVH lines for the PTV and OARs prediction in comparison to the SHU-net prediction line (dotted), as depicted in Figure 4. It is worth mentioning that the SHU-net forecasts lower doses on the inferior aspect of the brain and the left side of the shoulder compared to the MHU-net.

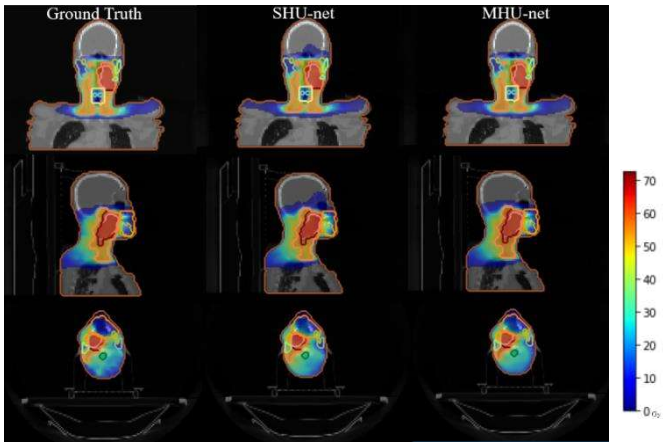


Figure 2: The dose washes of a representative patient from the test pool. The color bar represents the dosage in units of Gy.

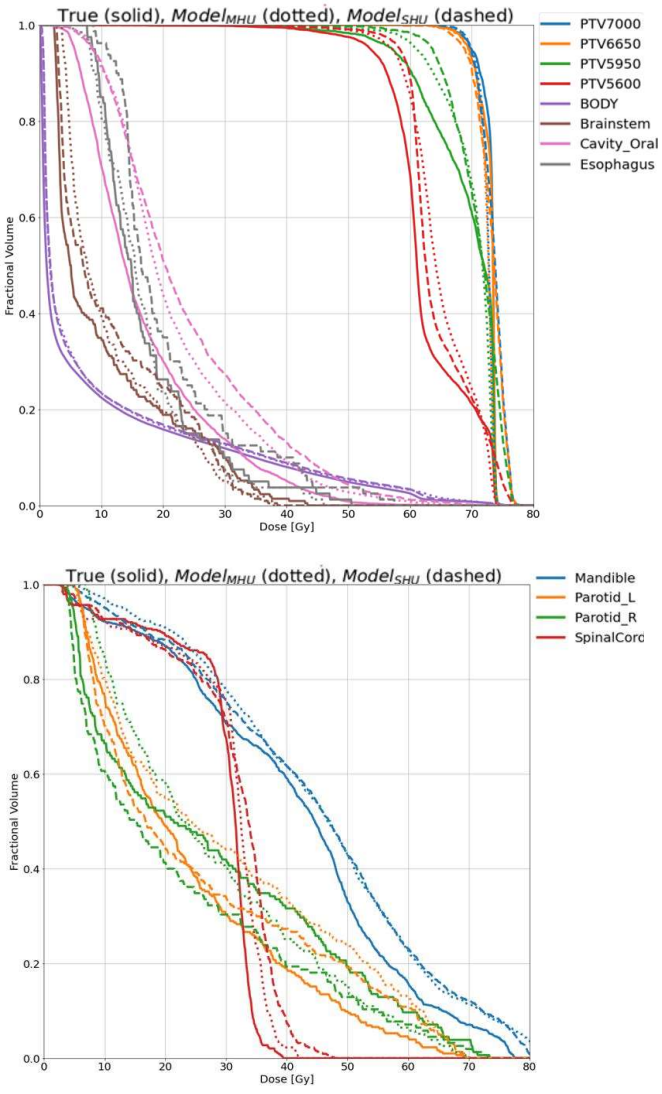


Figure 3: An example of DVH from a patient from the test data. The solid line represents the clinical ground truth dose, while the dotted and dashed lines represent the predicted dose from MHU-net, and the dashed line, respectfully.

3.2 Mean and Maximum Dose Error

According to Figure 4, the mean absolute percentage error analysis shows that out of the 11 structures examined, 8 of them demonstrated lower mean dose errors when predicted by the MHU-net (red bar) as compared to the SHU-net (blue bar). Moreover, the mean doses for these structures were within 6% of the prescription dose. This indicates a high level of accuracy and consistency in the dose prediction process. Similarly, as illustrated in Figure 3, the Absolute Percentage Error for the maximum dose ($\mathbf{D}_{max}$) shows that 8 out of the 11 structures predicted by MHU exhibited fewer errors in $\mathbf{D}_{max}$ error than those predicted by SHU models. The mean doses for these structures were within 7% of the prescription dose. This agreement between the predicted and prescribed doses highlights the effectiveness of the MHU-net in accurately estimating the radiation doses received by different regions of the patient's anatomy.

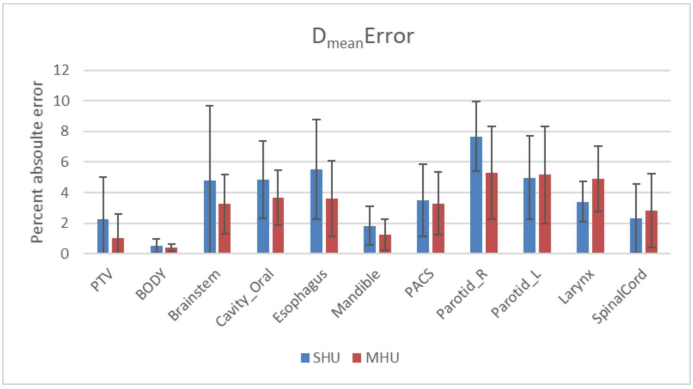


Figure 4: The absolute error of $\boldsymbol{D}_{mean}$ on the structures of interest. Error is reported as a percentage of the prescription dose



Figure 5: The absolute error of $\boldsymbol{D}_{max}$ on the structures of interest. Error is reported as a percentage of the prescription dose

Model Type	PTV D95	PTV D98	PTV D99	Homogeneity
Ground Truth	0.96±0.03	0.94±0.03	0.92±0.03	0.0013±0.00
SHU-net	0.92±0.02	0.88±0.02	0.82±0.02	0.0018±0.00
MHU-net	0.94±0.01	0.92±0.01	0.90±0.01	0.0015±0.00

Table 1: *Average PTV coverage and homogeneity for the ground truth and the prediction models.*

Table 1 shows that the MHU-net was able to achieve more similar PTV_{D95}, PTV_{D98}, and PTV_{D99} values with the ground truth than the SHU-net.

In our evaluation involving 10 test patients, we also observed a notable reduction in the mean squared error (MSE) when comparing the dose predictions of the SHU-net and the MHU-net. Specifically, the MHU-net achieved a significantly lower MSE of 0.0029 compared to the SHU-net's MSE of 0.0033, with a p-value of 0.0008 indicating statistical significance. This statistical finding also reflects a high level of agreement between the dose distributions predicted by the MHU-net and the ground truth dose.

4 Discussion

This study demonstrated a new DL-based model for 3D dose distribution prediction. To our knowledge, it has not been done with deep learning-based dose prediction for ART. The MHU-net is specifically proposed for ART, containing pre-plans accounted for the physician's treatment physician preferences and adaptive plans adjusted during treatment in response to anatomical changes for sustained treatment efficacy.

However, there are notable limitations to acknowledge. Firstly, the study's data size is constrained, involving only 43 eligible patients due to the novelty of the Varian Ethos treatment machine and ART software. This contrasts with typical deep-learning datasets, which often involve hundreds of thousands of data. Secondly, relying solely on the PTV, OARs, and CT as input data as input may constrain prediction performance. Improving the model's performance could be achieved by integrating additional information, such as various treatment modalities, beam energies, beam orientations, and field geometry. Lastly, the current dose prediction models are site-specific, necessitating restructuring and retraining for new target treatment sites with different OAR configurations. In future work, we aim to expand and enhance the model, addressing these limitations to enable implementation across multiple treatment sites.

5. Conclusion

Our proposed MHU-net demonstrates the viability of integrating a pre-plan that accounts for the physician's individualized preferences for each patient. This innovative approach goes beyond the SHU-net dose prediction model by taking into account the unique considerations and specifications set forth by the physician. By incorporating the clinician's preferences into the treatment planning process, the MHU-net enhances the personalization of dose predictions, ensuring that the proposed treatment regimen aligns more closely with the specific needs and characteristics of the patient. This personalized approach enhances treatment effectiveness, marking a significant advancement in tailoring medical care to individual patients in healthcare.

References

[1] Hong TS, Tomé WA, Chappell RJ, Chinnaiyan P, Mehta MP, and Harari PM 2005 “The impact of daily setup variations on head-and-neck intensity-modulated radiation therapy”. *International Journal of Radiation Oncology Biology Physics* 61(3) 779–788.

[2] Bhide S.A., Davies M., Burke K., McNair H.A. “Weekly volume and dosimetric changes during chemoradiotherapy with intensity-modulated radiation therapy for head and neck cancer: A prospective observational study”. *Int. J. Radiat. Oncol. Biol. Phys.* 2010;6:1360–1368. doi: 10.1016/j.ijrobp.2009.04.005.

[3] Trotti A 2000 “Toxicity in head and neck cancer: a review of trends and issues”. *International Journal of Radiation Oncology Biology Physics* 47(1) 1–12.

[4] Cooper JS, Fu K, Marks J, Silverman S 1995 “Late effects of radiation therapy in the head and neck region”. *International Journal of Radiation Oncology Biology Physics* 31(5) 1141–1164.

[5] Barker jr JL, Garden AS, Ang KK, et al. “Quantification of volumetric and geometric changes occurring during fractionated radiotherapy for head-and-neck cancer using an integrated CT/linear accelerator system”. *Int J Radiat Oncol Biol Phys*. 2004; **59**(4): 960-970.

[6] Wu Q, Chi Y, Chen PY, Krauss DJ, Yan D, Martinez “A 2009Adaptive Replanning Strategies Accounting for Shrinkage in Head and Neck IMRT”.75(3) 924–932.

[7] Fan J, Wang J, Chen Z, Hu C, Zhang Z, Hu W. “Automatic treatment planning based on three-dimensional dose distribution predicted from deep learning technique”. *Med Phys*. 2019; **46**(1): 370-381.

[8] Gronberg MP, Beadle BM, Garden AS, et al. “Deep learning–based dose prediction for automated, individualized quality assurance of head and neck radiation therapy plans”. *Pract Radiat Oncol*. 2023; **13**(3): e282–e291.

[9] Lee D, Zhang P, Nadeem S, Alam S, Jiang J, Caringi A, Allgood N, Aristophanous M, Mechalakos J, Hu YC. “Predictive dose accumulation for HN adaptive radiotherapy.” *Phys Med Biol*. 2020 Nov 27;65(23):235011. doi: 10.1088/1361-6560/abbdb8.

[10] Nguyen D, Jia X, Sher D, et al. “3D radiotherapy dose prediction on head and neck cancer patients with a hierarchically densely connected U-net deep learning architecture”. *Phys Med Biol*. 2019; **64**(6):065020.

Prediction of stress in radiation therapy patients using artificial intelligence and biological signals

Sangwoon Jeong[1], Hongryull Pyo[2,3], Won Park[2,3], Youngyih Han[2,3*]

[1] Department of Health Sciences and Technology, SAIHST, Sungkyunkwan University, Seoul, Korea
[2] Department of Radiation Oncology, Samsung Medical Center, Seoul, Korea
[3] School of Medicine, Sungkyunkwan University, Seoul, Korea

Abstract Patients undergoing radiation therapy can experience stress because of the fear of treatment. Stress causes muscles to contract, which can reduce the accuracy of the patient setup. In this study, we used biological signals to identify and artificial intelligence to predict stress during radiation therapy. Stress was calculated by analyzing biological signals measured before and during radiation therapy. We used various artificial intelligence models to verify the models optimized for stress prediction. Our findings indicate that over 90% of patients experience stress during treatment and artificial intelligence can predict this stress with over 80% accuracy. This study is pivotal for identifying patients requiring stress reduction before therapy, potentially enhancing the precision of cancer radiation therapy.

1 Introduction

A necessary procedure in radiation therapy involves the creation of an individual treatment plan to focus high-energy radiation on the tumor while sparing nearby critical organs, using simulation computed tomography images. For successful radiation therapy, the accurate positioning of the patient, identical to that in the treatment plan, is essential. If the patient's position does not match the treatment plan, the therapeutic effect is reduced and damage to normal tissues can occur [1]. Therefore, accurate beam alignment in accordance with the treatment plan to patients is critical for optimizing the therapeutic outcomes.

Accuracy of radiation treatment is influenced by both technological and human factors. Human factors, such as respiratory and gastrointestinal motion, shrinkage of the targeted tumor volume, and patient stress (anxiety) play significant roles. The inherent variability in human respiratory and digestive system movements can cause unpredictable displacements of the body and internal organs, which can be mitigated using techniques such as gating, tumor tracking, and image-guided radiation therapy. The treatment plan was modified using adaptive radiotherapy to accommodate the shrinkage of the target volume [2]. Although various methods to counteract these factors are used in daily patient treatment and are under development, there is a noticeable scarcity of research addressing the impact of patient stress on radiation therapy.

Evaluation of stress through biological signal monitoring is an emerging and pivotal field of medical research. Psychological stress triggers the sympathetic nervous system, leading to physiological changes such as increased heart rate (HR), blood pressure, breathing rate, and muscle stiffness [3]. Two of the most significant indicators in this field are photoplethysmogram (PPG) and electrocardiogram (ECG), both of which monitor changes in blood flow. Changes in blood flow are instrumental in determining heart rate variability (HRV), a key metric in stress evaluation [4]. The reliability and utility of HRV as a stress measure have been substantiated by comparison with traditional stress surveys. Moreover, the integration of HRV analysis into wearable technologies such as smartwatches, has opened new avenues for real-time, noninvasive stress monitoring [5]. Hence, HRV analysis is a promising alternative to survey-based methods that offer a more objective and continuous assessment of stress levels.

Stress in patients undergoing radiation therapy has been identified in survey studies [6]. It is observed that a majority of these patients experience heightened stress levels, particularly in the initial stages of their treatment. This underscores the need for effective stress management strategies. However, implementing universal stress reduction measures for all patients can be resource intensive and requires additional manpower and time. To address this challenge, we leveraged Artificial Intelligence (AI) techniques in conjunction with biological signal analysis to identify patients who are susceptible to stress during radiation therapy. Our approach involved training machine learning models on HRV data collected both before and during the treatment sessions. This study aimed to use before-treatment HRV data to predict the likelihood of patients experiencing significant stress during therapy sessions. This prediction enables us to tailor stress management interventions more effectively by focusing on those who need them the most.

2 Materials and Methods

2.1 Data acquisition and processing

The study protocol, including patient recruitment and data collection methods, was approved by the Institutional Review Board of the Samsung Medical Center (IRB number 2020-11-162). Prior to enrollment, written informed consent was obtained from all participants, confirming their voluntary participation and understanding of the study’s aims and processes. Our study prospectively enrolled patients who underwent radiation therapy for lung cancer. The recruitment period spanned from December 2020 to November 2023. Initially, 41 patients (123 case) were approached for participation in this study.

Data collection commenced with patients wearing a biological sensor (Laxtha, Ubpulse 360, Daejeon, Korea) upon arrival in the waiting room prior to receiving radiation

therapy. The sensor was positioned on the finger to ensure no interference during the treatment procedure. After a 10-minute acclimatization period, patients were escorted to the treatment room, where they continued to wear the sensor throughout their radiation treatment session. Upon completion of the treatment, the sensor was returned, and the collected PPG data were securely transferred to a dedicated computer system for analysis. Subsequently, the PPG data were segmented into two distinct phases for analysis: the before-treatment phase, captured while the patient was in the waiting room, and the during-treatment phase, recorded when the patient lay on the treatment couch. To account for potential HR elevations due to movement, we isolated 5 min of data following a 2-minute stabilization period in both the before- and during-treatment phases. From these phases, the HRV was computed by analyzing the intervals between successive PPG peaks. Pre-processing of the PPG data and subsequent HRV analyses were conducted using MATLAB R2020b (MATLAB, MathWorks, Natick, MA, USA) to ensure a standardized and reproducible methodology.

2.2 Stress features

Identification and accurate quantification of stress features are important for the assessment of stress levels using HRV analysis. In this study, we operationalized stress using a set of physiological markers derived from PPG signals. The second derivative of the PPG signal was used to pinpoint the heartbeat peaks, and HRV was calculated by measuring the intervals between these peaks. The selection of HRV-related stress features was based on a comprehensive literature review, identifying seven features consistently associated with physiological stress responses [7]. These features are summarized in Table 1. We employed these stress features to calculate the stress score (range: 0-100%) by observing changes before and during treatment.

Table 1. Summary of stress-related features based on heart rate variability.

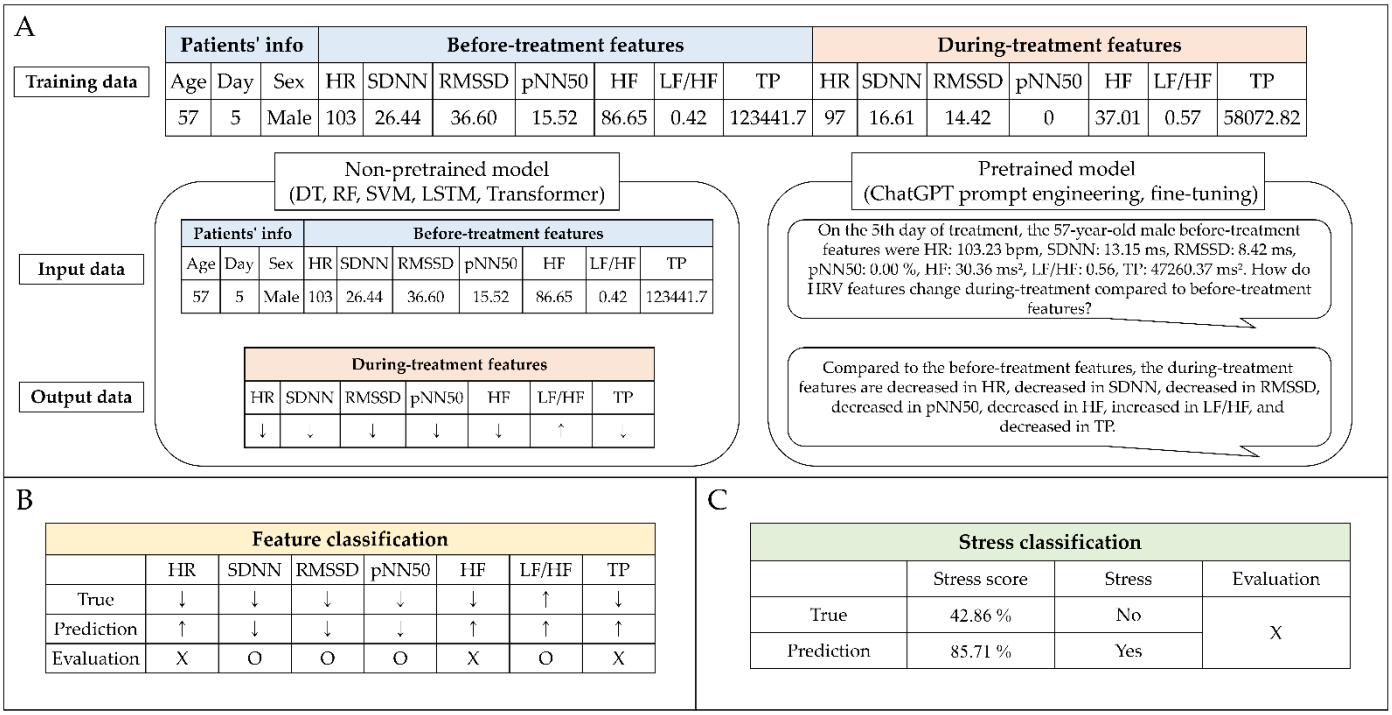
Features	Unit	Description	Stressful
HR	bpm	Average number of heart beats per minute	Increase
SDNN	ms	Standard deviation of NN intervals	Decrease
RMSSD	ms	Square root of the mean sum of squares of successive NN intervals differences	Decrease
pNN50	%	Percentage of successive NN intervals differing more than 50 ms	Decrease
HF	ms^2	Power of high-frequency range (0.15 – 0.4 Hz)	Decrease
LF/HF	ms^2	Ratio low-frequency range / high-frequency range	Increase
TP	ms^2	Total power of frequency range (0.004 – 0.4 Hz)	Decrease

2.3 Stress prediction

Predicting patient stress in the waiting room before treatment is crucial to enhance the accuracy of setting up patients for radiation treatment. This enables the early implementation of measures to reduce stress, potentially improving treatment efficacy. Non-pretrained and pretrained models were used for stress prediction. The non-pretrained model categories include decision tree (DT), random forest (RF), support vector machines (SVM), long short-term memory (LSTM), and transformer. The pre-trained models used were OpenAI's ChatGPT, which is based on a Large Language Model (LLM) and enables prompt engineering and fine-tuning. Prompt engineering involves the strategic design of input prompts to elicit the desired responses from an LLM, whereas fine-tuning refers to the process of adjusting an LLM's parameters on a specific dataset to improve its performance for particular tasks. The non-pretrained models were assessed using 10-fold cross-validation to evaluate their ability to handle three different input datasets. These datasets included treatment day, age, sex, and seven stress features identified before treatment. The model outputs were designed to classify the predicted changes in stress features during treatment (Figure 1). Subsequently, the top three input datasets from the performance of the non-pretrained models were selected for further analysis with the pretrained models. The pretrained model was evaluated against a representative one-fold out of a 10-fold cross-validation of the non-pretrained model. Therefore, the pretrained and non-pretrained models compared the results of the one-fold dataset. The pretrained models performed prompt engineering in GPT-3.5 and GPT-4.0 and fine-tuning in GPT-3.5-turbo-1106.

2.4 Evaluation

To assess the predicted stress features during treatment, we adopted two analytical approaches: feature classification (multi-label) and stress classification (binary). Feature classification utilized the raw output from our models to evaluate prediction accuracy across multiple labels. The key metrics included the Exact Match Ratio (EMR) and standard classification metrics such as accuracy, recall, precision, and F1 score, providing a holistic view of the models' performance. The feature classification result is calculated as a stress score but the stress classification is classified as "yes (> 50%)" or "no (< 50%)" based on the criterion of a stress score of 50%. The effectiveness of the stress classification was quantified using accuracy, recall, precision, and F1 score.



B

Feature classification	HR	SDNN	RMSSD	pNN50	HF	LF/HF	TP
True	↓	↓	↓	↓	↓	↑	↓
Prediction	↑	↓	↓	↓	↑	↑	↑
Evaluation	X	O	O	O	X	O	X

C

Stress classification	Stress score	Stress	Evaluation
True	42.86 %	No	X
Prediction	85.71 %	Yes	

Figure 1. Stress prediction workflow using artificial intelligence. (A) The training process of non-pretrained model and pretrained model. (B) Feature classification evaluation process. (C) The stress classification evaluation process based on 50% of the stress score calculated using the results of feature classification.

3 Results

3.1 Features classification

Datasets 6, 7, and 8, which exhibited the highest scores for recall, accuracy, and EMR, respectively, in the non-pretrained model, were selected for the pretrained model evaluation. The pretrained model was evaluated only one-fold, and for an intuitive comparison. Across the three datasets, GPT3.5 with prompt engineering, did not achieve an accuracy exceeding 0.5. In contrast, GPT4.0 outperformed GPT3.5 in all evaluation metrics. In the type 8 dataset, the fine-tuned GPT3.5-turbo-1160 model exhibited the most impressive results, whereas its performance on the type 7 dataset was comparatively lower. When focusing on the type 7 dataset, the LSTM model achieved the highest accuracy and recall. However, for the type 8 dataset, GPT4.0 with prompt engineering emerged as a superior model in terms of EMR, precision, and F1 score.

Table 2. Non-pretrained and pretrained model features classification.

Type	Model	EMR	Accuracy	Recall	Precision	F1 score
6	DT	0.000	0.637	0.727	0.635	0.672
	RF	0.077	0.637	0.777	0.611	0.669
	SVM	0.077	0.670	0.799	0.646	0.701
	LSTM	0.154	0.681	0.895	0.552	0.655
	Trans-	0.077	0.505	0.552	0.332	0.360
	3.5 (P)	0.077	0.440	0.610	0.330	0.397
	4.0 (P)	0.000	0.615	0.746	0.665	0.674
	3.5 (F)	0.154	0.527	0.726	0.412	0.503
7	DT	0.077	0.582	0.688	0.593	0.632
	RF	0.154	0.626	0.768	0.585	0.649
	SVM	0.154	0.637	0.783	0.581	0.646
	LSTM	0.154	**0.703**	**0.907**	0.587	0.707
	Trans-	0.000	0.560	0.453	0.563	0.492
	3.5 (P)	0.077	0.396	0.538	0.358	0.416
	4.0 (P)	0.077	0.560	0.667	0.629	0.638
	3.5 (F)	0.000	0.484	0.593	0.512	0.538
8	DT	0.000	0.637	0.727	0.635	0.672
	RF	0.154	0.626	0.768	0.585	0.649
	SVM	0.077	0.648	0.780	0.599	0.660
	LSTM	**0.231**	0.692	0.781	0.635	0.685
	Trans-	0.000	0.560	0.375	0.571	0.451
	3.5 (P)	0.077	0.407	0.369	0.278	0.246
	4.0 (P)	**0.231**	0.659	0.742	**0.731**	**0.723**
	3.5 (F)	0.077	0.615	0.714	0.613	0.646

3.2 Stress classification

Stress classification was binary (Figure 1C). The feature classification result is calculated as a stress score but the stress classification is classified as “yes (> 50%)” or “no (< 50%)” based on the criterion of a stress score of 50%. Therefore, this differs from the feature classification results. As shown in Table 3, among the non-pretrained models, the LSTM using the type 7 dataset achieved the highest accuracy (0.846) in stress classification. For the pretrained models on the type 8 dataset, GPT4.0 with prompt engineering, demonstrated the highest accuracy at 0.769 and GPT3.5 with prompt engineering, had the highest recall at 0.800, but the lowest accuracy at 0.385. The RF and SVM models exhibited equivalent performances across the three types of datasets. The DT and transformer models were the most effective for the type 8 dataset.

Table 3. Non-pretrained and pretrained model stress classification.

Type	Model	Accu-racy	Recall	Preci-sion	F1 score
6	DT	0.615	0.375	0.429	0.400
	RF	0.769	0.400	0.571	0.471
	SVM	0.769	0.400	0.571	0.471
	LSTM	0.692	0.444	0.500	0.471
	Trans-	0.538	0.571	0.400	0.471
	3.5 (P)	0.385	0.600	0.300	0.400
	4.0 (P)	0.615	0.500	0.444	0.471
	3.5 (F)	0.615	0.375	0.429	0.400
7	DT	0.615	0.250	0.400	0.308
	RF	0.769	0.400	0.571	0.471
	SVM	0.769	0.400	0.571	0.471
	LSTM	**0.846**	0.364	**0.667**	0.471
	Trans-	0.538	0.571	0.400	0.471
	3.5 (P)	0.462	0.500	0.333	0.400
	4.0 (P)	0.385	0.200	0.167	0.182
	3.5 (F)	0.462	0.167	0.200	0.182
8	DT	0.692	0.333	0.500	0.400
	RF	0.769	0.400	0.571	0.471
	SVM	0.769	0.400	0.571	0.471
	LSTM	0.769	0.400	0.571	0.471
	Trans-	0.692	0.000	0.000	0.000
	3.5 (P)	0.385	**0.800**	0.333	0.471
	4.0 (P)	0.769	0.300	0.600	0.400
	3.5 (F)	0.615	0.250	0.400	0.308

4 Discussion

In the analysis presented in Table 2, for the type 6 and type 7 datasets, the LSTM model continued to outperform the others in terms of EMR, accuracy, and recall. However, in the type 8 dataset, both GPT4.0 and LSTM demonstrated superior performance in EMR, achieving a score of 0.231. While LSTM led to accuracy and recall, GPT4.0 excelled in precision and F1 score. GPT3.5 displayed the lowest performance across all indicators in these datasets, with GPT3.5-turbo-1160 achieving an accuracy of 0.615 for the type 8 dataset.

Considering all models, including non-pretrained and pretrained, the LSTM model demonstrated robust performance across all evaluation indices and datasets, making it the most suitable for feature classification during treatment. In scenarios where implementing a machine learning model is challenging, the pretrained GPT4.0 model, particularly with the type 8 dataset, emerged as the most appropriate choice.

Stress classification classifies "yes (> 50%)" or "no (< 50%)" based on the criterion of stress score of 50% (Table 3). In stress classification, the LSTM of the type 7 dataset classified stress effectively with an accuracy of 0.846. The RF and SVM exhibited a stability of 0.769 accuracy across all datasets. For the pretrained model, GPT4.0 showed an accuracy of 0.769 in the type 8 dataset that included all data, but in the type 7 dataset, all pretrained models failed to exceed the accuracy of 0.5. As with feature classification, LSTM was the best among all models for stress classification, with GPT4.0 being superior for the type 8 dataset. GPT4.0 is suited for predictions using diverse patient information, whereas LSTM is recommended because of its stability in scenarios with limited information. Datasets 6, 7, and 8 used for comparison in Tables 2 and 3 contain the treatment days. The treatment day is important information for stress prediction. The performance of the non-pretrained model was similar across the three dataset types. However, the pretrained model performance was the best in the type 8 dataset, which included age, sex, and treatment day, and the worst in the type 7 dataset, in which age was omitted. Stress prediction using a pre-trained model may be better when using all available patient information.

5 Conclusion

This study marks a significant advancement in cancer radiation therapy by employing AI and biological signals for stress prediction. This highlights the potential use of these tools in identifying patients who may require psychological support prior to treatment. This study suggests the need for more diverse biological signals and patient groups to enhance our understanding of the effects of stress on radiation therapy.

References

[1] Van Herk, M. Errors and margins in radiotherapy. In Proceedings of the Seminars in radiation oncology, 2004; pp. 52-64.

[2] Lee, H.; Ahn, Y.C.; Oh, D.; Nam, H.; Noh, J.M.; Park, S.Y. Tumor volume reduction rate during adaptive radiation therapy as a prognosticator for nasopharyngeal cancer. Cancer Research and Treatment: Official Journal of Korean Cancer Association 2016, 48, 537-545.

[3] Grassi, G.; Vailati, S.; Bertinieri, G.; Seravalle, G.; Stella, M.L.; Dell'Oro, R.; Mancia, G. Heart rate as marker of sympathetic activity. Journal of hypertension 1998, 16, 1635-1639.

[4] Mohan, P.M.; Nagarajan, V.; Das, S.R. Stress measurement from wearable photoplethysmographic sensor using heart rate variability data. In Proceedings of the 2016 International Conference on Communication and Signal Processing (ICCSP), 2016; pp. 1141-1144.

[5] Dai, R.; Lu, C.; Yun, L.; Lenze, E.; Avidan, M.; Kannampallil, T. Comparing stress prediction models using smartwatch physiological signals and participant self-reports. Computer Methods and Programs in Biomedicine 2021, 208, 106207.

[6] Antoni, D.; Vigneron, C.; Clavier, J.-B.; Guihard, S.; Velten, M.; Noel, G. Anxiety during radiation therapy: a prospective randomized controlled trial evaluating a specific one-on-one procedure announcement provided by a radiation therapist. Cancers 2021, 13, 2572.

[7] Cinaz, B.; Arnrich, B.; La Marca, R.; Tröster, G. Monitoring of mental workload levels during an everyday life office-work scenario. Personal and ubiquitous computing 2013, 17, 229-239.

Habitat imaging based on voxel-wise GLCM joint energy clustering: prediction of disease-free survival in localized soft-tissue sarcoma.

Benoît Allignet[1,2], Benjamin Leporq[1], Amine Bouhamama[3], Frank Pilleul[3], Alexandra Meurgey[4], Gualter Vaz[5], Marie-Pierre Sunyach[2], Waisse Waissi[2], Olivier Beuf[1]

[1]CREATIS, CNRS UMR 5220, Inserm U1294, INSA-Lyon, Université Claude Bernard Lyon 1
[2]Department of Radiotherapy, Centre Léon Bérard, Lyon, France
[3]Department of Radiology, Centre Léon Bérard, Lyon, France
[4]Department of Biopathology, Centre Léon Bérard, Lyon, France
[5]Department of Surgical Oncology, Centre Léon Bérard, Lyon, France

Abstract We evaluated if disease-free survival (DFS) of localized soft-tissue sarcoma (STS) patients can be predicted by tumor habitats based on a single voxel-wise radiomics map. The GLCM joint energy map was calculated on fat suppressed T1-weighted contrast-enhanced MRI, performed before neoadjuvant radiotherapy. Habitats were individualized based on k-means clustering method. The median relative volume of habitat 1 (RVH1, reflecting high heterogeneity) was 56.1%, and was used for dichotomization of the sample. Between 2010 and 2020, 149 soft-tissue sarcoma patients were retrospectively included. In this real-life cohort, RVH1 group was independent of MRI acquisition parameters. After a median follow-up of 46.3 months, 4-year DFS was significantly shorter in RVH1-high compared to RVH1-low group (52.5% *vs* 71.5%, p=0.029). In multivariate analysis, RVH1-high and FNCLCC grade 3 remained significantly correlated with shorter DFS (both HR=1.8, p=0.046 and p=0.033, respectively). Habitat imaging based on voxel-wise radiomics may be a valuable tool to evaluate intra-tumor heterogeneity to predict prognosis and/or to personalize treatment.

1 Introduction

Soft-tissue sarcomas (STS) are rare ubiquitous neoplasms with poor prognosis. They present important inter-tumor heterogeneity, from indolent to highly aggressive disease, but also intra-tumor heterogeneity, associating various types of tissues.

Radiomics refers to the process of extracting multiple quantitative imaging biomarkers to use them for outcomes modelling and clinical decision making. This analytical approach usually summarises tumor characteristics into various 1-dimensional (i.e. scalars) features and perform a data-driven approach for outcome prediction. Data are available using various imaging technics applied to various indications, including MRI and STS [1].

New approaches are emerging to consider spatial profiling [2]. Voxel-wise radiomics generate radiomics filtered images and could be more interpretable than 1D radiomics. On the other hand, imaging data can be used to identify tumor sub-regions ("habitats") with different biological properties.

We here propose to combine these habitat imaging and voxel-wise radiomics approach to predict disease-free survival (DFS) in localized STS patients treated by neoadjuvant RT and surgery.

2 Materials and Methods

Patients and magnetic resonance imaging

We retrospectively included all adult patients treated by normofractionated neoadjuvant RT for a localized limb or truncal STS in Centre Léon Bérard (Lyon, France) from January 2010 to November 2020. Exclusion criteria were retroperitoneal STS, bone, Ewing and Ewing-like sarcoma, desmoid tumors, experimental (neo)adjuvant treatment, metastatic disease. Since MRI were performed in various centers, we focused on the only sequence common to all patients: the fat-suppressed T1-weighted gadolinium-enhanced sequence, acquired in 2 or 3 dimensions.

Segmentation and voxel-wise radiomics extraction

Anonymized images were processed on Slicer3D® software. In MRI, magnetic field inhomogeneity and MR coils sensitivity effect can generate a low frequency intensity non-uniformity, called image bias field, that can distort images. Bias field Image bias field was corrected by the retrospective histogram based N4ITK algorithm [3]. Tumors were manually segmented (BA, 5 years of experience). Voxel-wise radiomics were then extracted after z-score normalization, with a fixed bin size of 32 and a kernel radius of 3. B-spline interpolation and nearest neighbor interpolation were used to resample images and segmentation masks, respectively, to 1 mm isotropic voxel size. Radiomics extraction was performed using Python (v3.11.2) and reference packages PyRadiomics (v3.0.1), Numpy (v1.24.2), and SimpleITK (v2.2.1).

K-means clustering-based habitats

The grey-level co-ocurrence matrix joint energy (GLCM-JEnerg) is a measure of homogeneous patterns in the image:

$$GLCM_{JointEnergy} = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g}\left(p(i,j)\right)^2$$

with N_g the discrete number of intensity in the image and p(i,j) the normalized co-ocurrence matrix. We planned to focus on this 3D feature-map since it was previously described as repeatable and reproducible [4], and correlated with local recurrence in glioblastomas [5].

A k-mean clustering method was used on the GLCM-JEnerg map to separate various habitats according to their degree of local heterogeneity. For a number k of cluster (habitat) between 1 and 10, the centroids were calculated to minimize the sum of the squared Euclidean distance

differences between each value and the centroid of the cluster (*within cluster sum of square*, WCSS):

$$WCSS(k) = \sum_{j=1}^{k} \sum_{x_i \in cluster\ j} \left\| x_i - \bar{x}_j \right\|^2$$

with $\bar{x}_j$ the centroid of cluster j. The number of clusters for all maps was the optimal number of the majority, determined by the Elbow method.

The habitats were sorted by increasing centroid values (i.e. habitat 1 is the most heterogeneous one), and relative volume occupied by habitat 1 (RVH1) was recorded. The threshold used to dichotomize RVH1 was the overall cohort median value. All this clustering step was performed on MatLab (R2023b).

Clinical outcomes

Primary outcome was disease-free survival (DFS). Secondary outcomes included local recurrence-free survival (LRFS) and predictive factors of RVH1.

Statistical analyses

Categorical variables were reported as counts (percentages) and compared using Chi2 tests. Continuous variables were reported as median (interquartile range) and compared using T-tests or Welch tests (according to the normality of the distributions), or by ANOVA. Median follow-up was determined using the reverse Kaplan Meier method. Survival outcomes were estimated by the Kaplan-Meier method. Prognostic factors were explored in univariate and multivariate Cox proportional hazard models and included RVH1 and usual known factors: age ≥70 years, grade 3 according to fédération nationale des centres de lutte contre le cancer (FNCLCC), deep-seated tumor, size > 5 cm, undifferentiated pleomorphic sarcoma or myxofibrosarcoma histotypes. Two-sided p-value <0.05 was considered significant. Statistical analyses were performed using R software version 4.2.3 (R Foundation for Statistical Computing, Vienna, Austria).

3 Results

We retrospectively included 149 STS patients. Their clinical characteristics and the MRI acquisition parameters are summarized in the Table 1.

Table 1. Patients' characteristics

Characteristics	Overall (n=149)
Age	61 (48-73)
Sexe	
Male	72 (48.3)
Female	77 (51.7)
Deep-seated lesion	138 (92.6)
Location	
Extremities	118 (79.2)
Trunk	31 (20.8)
Histotype	
UPS	52 (34.9)
Myxoid liposarcoma	23 (15.4)
Myxofibrosarcoma	22 (14.8)
Dediff./pleo. liposarcoma	17 (11.4)
Synovial sarcoma	14 (9.4)
Others	21 (14.1)
FNCLCC grade	
Grade 1-2	89 (62.7)
Grade 3	50 (35.2)
Neoadjuvant chemotherapy	40 (26.8)
Number of cycles of NAC	4 (3-6)
Radiotherapy technique	
Tomotherapy	67 (45)
VMAT	39 (26.2)
3D-CRT	43 (28.9)
Total dose	50 (50-50)
Overall treatment time	36 (35-39)
Volumes of radiotherapy (cc)	
GTV	263 (115-500)
CTV	957 (457-1387)
PTV	1183 (599-1741)
Surgical margins	
R0	123 (82.6)
R1	26 (17.4)
Magnetic field of the MRI	
1.5 T	118 (79.2)
3.0 T	31 (20.8)
Dimension of acquisition	
2D	81 (54.4)
3D	68 (44.6)
Type of fat-suppressed T1-CE sequence	
Gradient echo sequence	63 (42.3)
Flip angle	11 (10-11)
Echo time	2.46 (2.39-2.46)
Repetition time	6.38 (5.55-6.79)
Spin echo sequence	86 (57.7)
Echo time	11 (9.5-13)
Repetition time	586.2 (531.25-719.5)
Method of fat suppression	
Fat-water decomposition	83 (55.7)
Fat saturation	60 (40.3)
Inversion-recovery imaging	6 (4)

Note:
Continuous variables: median (IQR)
Categorial variables: n (%)

Characteristics	Overall (n=149)

Abbreviations: UPS, undifferentiated pleomorphic sarcoma; dedif./pleo., dedifferentiated/pleomorphic; FNCLCC, fédération nationale des centres de lutte contre le cancer; NAC, neoadjuvant chemotherapy; VMAT, volumetric modulated arctherapy; T1-CE, T1-weighted contrast-enhanced sequence

The optimal number k of habitats was 4 for most tumors (n=111; k=3 for n=37 and k=5 for n=1), with a median centroid of habitat 1 of 0.021 (IQR 0.016-0.029).
The median RVH1 was 56.1% (49-65) (Figure 1). The RVH1 group was not related to the MRI magnetic field strength, type of sequence, dimension of acquisitions (all p>0.96) nor to the method of fat suppression (p=0.23). It was not either related to histological subtype (p=0.09), nor FNCLCC grade 3 (p=0.2), tumor absolute volume (p=0.17) or predictive of the percentage of viable cells on surgical specimen (p=0.9).

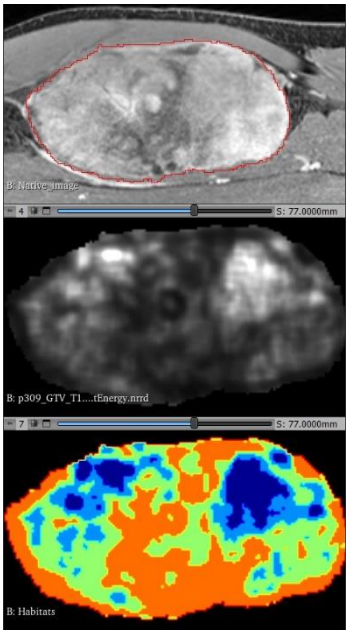


Figure 1. Native T1-weighted gadolinium enhanced Dixon-water image (up) with its GLCM -Joint Energy corresponding map (middle) and the habitat maps (down). The gross tumor volume is segmented in red. The habitat 1 to 4 are depicted from orange to dark blue, respectively.

After a median follow-up of 46.3 months (95%CI 33.1-70.1), 4-year DFS was 71.5% and 52.5% in RVH1-low and RVH1-high group, respectively (Figure 2), with 4-year LRFS of 77.3% and 65%, respectively (Figure 3). In univariate analysis, only grade 3 was significantly correlated with LRFS (HR 2.3, 95%CI 1.2-4.2, p=0.01). In both uni- and multivariate analyses, DFS was significantly related to FNCLCC grade 3 and RVH1 (HR 1.8, 95%CI 1.1-3.1, p=0.033; HR 1.8, 95%CI 1.01-3.1, p=0.046, respectively).

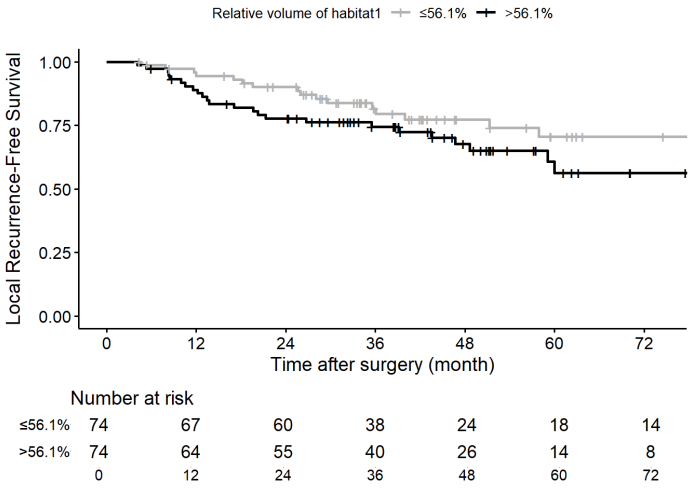


Figure 2. Local recurrence-free survival according to relative volume of habitat 1

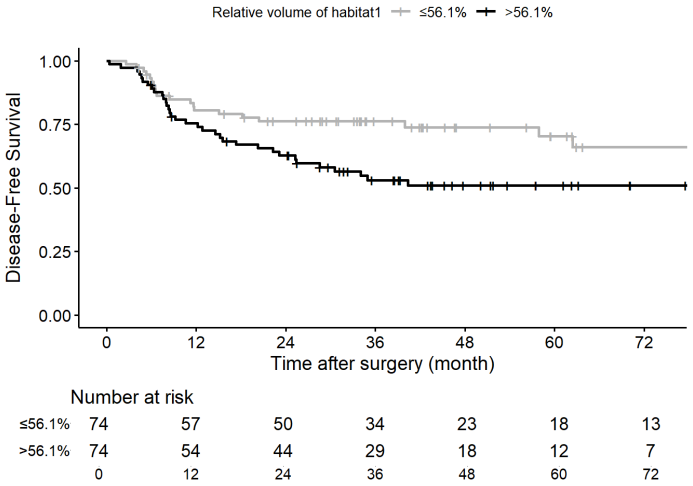


Figure 3. Disease-free survival according to relative volume of habitat 1

As a sensitivity analysis, we divided the cohort into 3 groups according to the terciles of the RVH1 distribution. The DFS tended to shorten with the increase of RVH1, with 4 year-DFS of 75% *vs* 60.9% *vs* 50.8% (Figure 4).

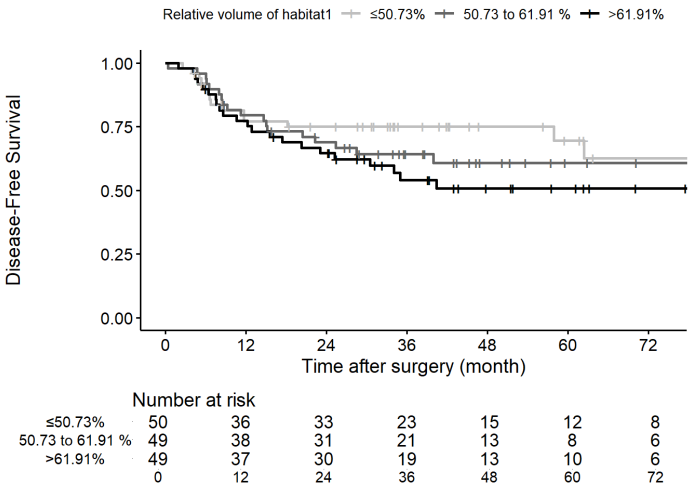


Figure 4. Disease-free survival according to relative volume of habitat 1, divided into 3 groups.

4 Discussion

We here report an original methodology of habitats imaging based on fat suppressed T1-weighted contrast-enhanced sequence voxel-wise radiomics. GLCM-JEnerg RVH1-low group present a significantly longer DFS, with a clinically meaningful difference.

This highlights the potential interest of voxel-wise radiomics either to identify tumor subclones or to predict patient prognosis.

Nevertheless, some refinements of the method may be performed. First, the repeatability of the habitats may be verified, for example by adding Rician noise in native images and process them back. As a retrospective analysis, the reproducibility of the method when performing other MR acquisitions cannot be confirmed. In contrast, an external cohort may be used to reproduce and validate the whole processes and results.

MR acquisition parameters were variable between patients and could thus introduce bias, particularly using vowel-wise radiomics [4]. Nevertheless, various steps limited this: intensity non-uniformity was corrected, native images were standardized, and a clustering was individually performed. Furthermore, the whole process was not data- but hypothesis-driven, with interpretable intermediate results, and is applied on real-world data acquired on an important cohort of patients, with sufficient follow-up, presenting rare disease managed in an expert center.

5 Conclusion

In this real-world cohort of localized STS, habitat imaging based on a single voxel-wise radiomic map was independently correlated with DFS. This kind of approach could be a valuable tool to integrate intra-tumor heterogeneity into RT personalization.

Acknowledgements

With financial support from ITMO Cancer of Aviesan within the framework of the 2021-2030 Cancer Control Strategy, on funds administered by Inserm, and from the committee 69 of the Ligue Nationale Contre le Cancer. This work was also performed within the framework of the the LabEx PRIMES (ANR-11-LABX-0063) as well as SIRIC LyriCAN+ grant (INCa INSERM DGOS 12563).

References

[1] Spraker MB, Wootton LS, Hippe DS, Ball KC, Peeken JC, Macomber MW, et al. MRI Radiomic Features Are Independently Associated With Overall Survival in Soft Tissue Sarcoma. Adv Radiat Oncol 2019;4:413–21. DOI: 10.1016/j.adro.2019.02.003

[2] Napel S, Mu W, Jardim-Perassi B V., Aerts HJWL, Gillies RJ. Quantitative imaging of cancer in the postgenomic era: Radio(geno)mics, deep learning, and habitats. Cancer 2018;124:4633–49. DOI: 10.1002/cncr.31630

[3] Tustison NJ, Avants BB, Cook PA, Zheng Y, Egan A, Yushkevich PA, et al. N4ITK: Improved N3 bias correction. IEEE Trans Med Imaging 2010;29:1310–20. DOI: 10.1109/TMI.2010.2046908.

[4] Bernatowicz K, Grussu F, Ligero M, Garcia A, Delgado E, Perez-Lopez R. Robust imaging habitat computation using voxel-wise radiomics features. Sci Rep 2021;11. DOI: 10.1038/s41598-021-99701-2

[5] Chougule T, Gupta RK, Saini J, Agrawal S, Gupta M, Vakharia N, et al. Radiomics signature for temporal evolution and recurrence patterns of glioblastoma using multimodal magnetic resonance imaging. NMR Biomed 2022;35. DOI: 10.1002/nbm.4647

A Deep Learning-Driven Framework for Automatic Target Volume Localization and Delineation in Spinal Metastases Stereotactic Body Radiotherapy

Zi Yang[1], Mahdieh Kazemimoghadam[3], Gregory Szalkowski[1], Cynthia Chuang[1], Lei Wang[1], Lianli Liu[1], Scott Soltys[1], Erqi Pollom[1], Elham Rahimy[1], Hao Jiang[4], David Park[2], Amit Persad[2], Yusuke Hori[2], Jie Fu[1], Ignacio Romero[1], Laszlo Zalavari[1], Mingli Chen[3], Weiguo Lu[3] and Xuejun Gu[1]

[1]Department of Radiation Oncology, Stanford University School of Medicine, Palo Alto, CA, USA
[2]Department of Neurosurgery, Stanford University School of Medicine, Palo Alto, CA, USA
[3]Department of Radiation Oncology, University of Texas Southwestern Medical Center, Dallas, TX, USA
[4]NeuralRad LLC, Madison, WI, USA

Abstract Stereotactic body radiotherapy (SBRT) for spinal metastases requires accurate target delineation to achieve optimal plan quality. However, the delineation of target volume is a labor-intensive process and subject to inter- and intra- observer variations. To address the clinical needs and improve the spine SBRT workflow, we propose a deep learning-driven framework for the auto-localization and delineation of the target volume for spine SBRT, using multi-modal imaging. Consisting of a vertebrae segmentation network and a target volume segmentation network, the proposed workflow can localize the lesion on images based on clinical chart diagnosis, and further generate clinical target volumes (CTV) with a Dice score of 0.86 on the independent testing set. Additionally, this framework is supported by a comprehensive web platform, allowing efficient patient database management and user interactions such as review and revise. Our proposed framework can be a promising tool to streamline the spine SBRT treatment workflow.

1 Introduction

The spine is the most common site for bony metastases, with up to 40% of cancer patients developing spinal metastases[1]. As cancer patients' life expectancy increases, so does the incidence of spinal metastases, making their management increasingly challenging and demanding[2]. Spine stereotactic body radiotherapy (SBRT) has emerged as a promising treatment regimen, delivering highly conformal radiation to spinal lesions while minimizing exposure to surrounding healthy tissue. Rapidly adopted for spine metastases management[3], SBRT has demonstrates high local control with minimal neural or systemic toxicity[4,5].

Like other radiotherapy (RT) regimens, spine SBRT requires a potent tumoricidal dose for effective tumor control, balanced against minimizing tissue-damaging doses to reduce complications in organs at risk (OARs), such as radiation myelopathy and vertebral compression fractures (VCFs). In current practice, creating an optimal treatment plan for spine SBRT is particularly challenging due to the close proximity of the spinal cord and the overlapping of the vertebrae with the target area. Precise target volume delineation is therefore an essential step in achieving a s high-quality treatment plan.

To aid in this process, consensus guidelines for spine SBRT target volume have been published to standardize the target delineation[6]. Yet, the manual process of defining treatment target volumes remains a time-consuming and laborious task, heavily reliant on the clinician's expertise. This process is prone to considerable variations among different clinicians, as several studies have shown common intra- and inter-observer variations in defining target volumes for spine SBRT[7-8]. The complex structure of the spine and the large variability in patient anatomy add to the difficulty in accurately localizing and delineating the target volume. Multi-modal imaging including CT and MRIs are often utilized together to provide sufficient information about the lesion target. After the localization of the gross tumor volume (GTV), clinical target volume (CTV) can be subsequently contoured by incorporating different anatomical regions (such as vertebral body, pedicles, spinous process, etc.) based on the location of the GTV per consensus guidelines.

With advancements in artificial intelligence (AI) technology, particularly in mathematical algorithms and computing technologies, AI has had a transformative impact on the field of radiation therapy[9-10]. However, despite

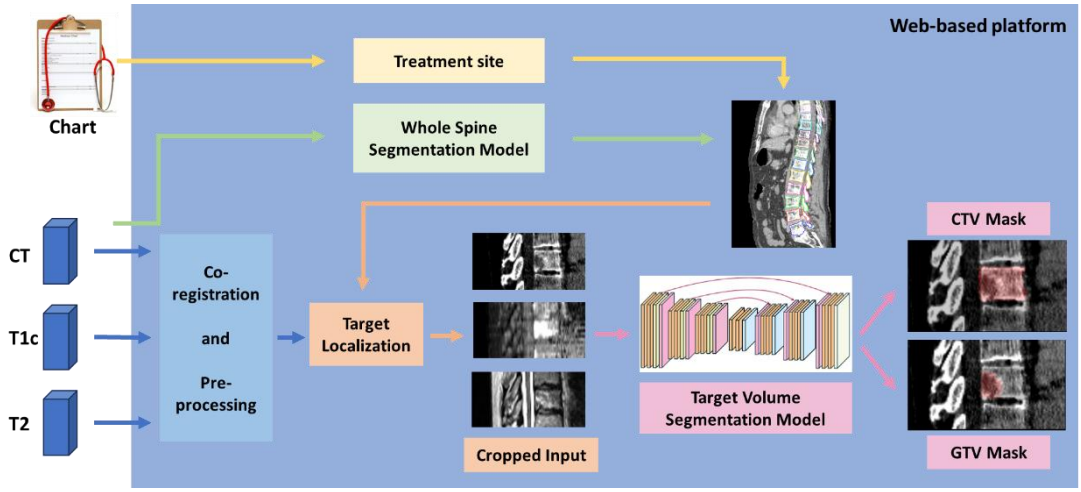


Figure 1 The overall design of the proposed framework for spine SBRT treatment target localization and delineation.

that various studies have been proposed on automating spine segmentation, they primarily focus on vertebrae segmentation of the entire spine and are not yet directly applicable to clinical spine SBRT workflow[11-12]. The automation of target volume localization and delineation remains an urgent need to improve the spine SBRT treatment workflow.

In this study, we propose a deep learning (DL)-driven framework to address the clinical need. The proposed framework not only enables AI-based vertebrae segmentation and labeling as well as GTV, CTV localization and delineation, but also maintains a comprehensive database for managing spine SBRT DicomRT data. In additional, this proposed framework is supported by a web-based platform to enable user interaction to facilitate clinical workflow.

2 Materials and Methods

The patient data used in this study were retrospectively collected from our institutional CyberKnife spine database. Only patients with spinal metastasis limited to one vertebral level and located at thoracic or lumbar spine are selected in this study. CTVs and GTVs that were contoured by physicians per consensus guidelines are used as ground truth in the study. Initially, a total of 48 spine SBRT patients were collected. Forty-two of these 48 patients met the inclusion criterion with a total of 56 treatment site. We further randomly divided these 56 cases into 3 groups: 40 for the DL model training, 8 for model validation and 8 for independent testing. The data stratification ensures that multiple treatment sites of one patient remain in the same group.

Figure 1 demonstrates the design of our proposed framework that can be summarized into three steps: 1) Spine segmentation and labeling, 2) target localization, and 3) Target volume segmentation. For each case, the corresponding framework input consists of CT, T1ce (T1 contrast enhanced), and T2 MRIs. The MRI images are co-registered and resampled into the same resolution and spatial coordinate as the CT.

In the first step, after importing the multi-modal images, the whole spine will be segmented from the input CT images using a TotalSeg-based DL model embedded in the framework platform[13]. Specifically, detailed spine vertebrae are auto segmented and labeled according to the spine anatomy.

Next, in the target localization step, the diagnosis regarding which vertebra the lesion is located at is then retrieved from clinical charts, e.g. “T03”. Then the corresponding vertebra mask generated from the above whole spine segmentation process will be used to localize the volume of interest (VOI) and to crop the multi-modal images into smaller crops with the size of 120×118×60 centered at the targeted vertebra, utilizing the center of mass (CoM) of the corresponding vertebra. The crop size is the same among all cases and is chosen to enclose the entire target vertebra with small margins added in all direction.

In the final step, the cropped multi-modal images are used as the 3-channel input to our segmentation model, which uses a 3D nnU-Net as the backbone, to generate the GTV and CTV segmentation simultaneously[14]. The model consists of a contracting path for context capture, featuring successive layers of 3D convolutions and pooling, where feature channels increase with each layer. A bottleneck connects to an expanding path for precise localization, which includes 3D up convolutions and concatenations with corresponding layers from the contracting path. This design, with skip connections between down-sampling and up-sampling layers, ensures detailed feature mapping.

The experiments were implemented using the Pytorch library with Python environment on Windows 11 64x, Intel Xeon processor CPU with 64 GB RAM, and NVIDIA GeForce RTX 3090 GPU. The proposed model parameters are optimized for 250 epochs through minimizing the Dice loss function (formula 1), which compared the prediction with manually contoured masks. Other model parameters and configurations, e.g., batch size, are determined using the nn-UNet framework. The validation set was used for tuning the model hyperparameters, while the independent testing set is a holdout set that is not involved in training at all to test the final model performance.

$$loss = 1 - \frac{2\sum_i^k P_i G_i}{\sum_i^k (P_i)^2 + \sum_i^k (G_i)^2} \quad (1)$$

P_i and G_i represent the predicted segmentation and ground truth manual contour with K voxels, respectively.

The entire framework is supported by a platform with web-based interface under our institutional network[15], as shown in the figure 2. Once importing the data from the web-browser, the platform will automatically execute the following procedures. In addition, the platform enables the visualization of the entire process and allows the user interaction for modification and inspection, meanwhile maintaining a patient database for improved data management.



Figure 2 The web-based interface of the proposed Spine SBRT data management and target delineation framework.

3 Results

We evaluate the performance of our proposed target localization and segmentation framework on both the validation set and the independent testing set using multi-modal images. The averaged Dice similarity coefficient (DSC), and the averaged 95th percentile Hausdorff

distance (HD95) are used to quantitatively assess the segmentation performance of our model. DSC is computed to measure the degree of overlap between the model prediction and the ground truth CTV masks (formula 2). In addition, HD95 (formula 3) measures the greatest value of all the distances from a point on model's segmentation output to the closest point on the ground truth contour set.

$$DSC = \frac{2|P \cap G|}{|P| + |G|} \quad (2)$$

$$HD = \max\left(h(S_P, S_G), h(S_G, S_P)\right) \quad (3)$$

$$h(S_P, S_G) = max_{p_P \in S_P} min_{p_G \in S_G} \|p_P - p_G\| \quad (4)$$

P and G represent the predicted segmentation and the ground truth volumes, respectively. S_P represents the surface of the predicted CTV segmentation, and the S_G represents the surface of the ground truth CTV contour. The terms p_P and p_G are the points in contour surfaces S_P and S_G, respectively. The DSC and HD95 of all 8 independent testing cases are listed in the table 1.

Table 1 Dice score and 95th percentile Hausdorff distance (HD95) of the segmentation results on the independent testing set.

Testing	GTV		CTV	
Case	DSC	HD95 (mm)	DSC	HD95 (mm)
1	0.20	20.06	0.88	4.03
2	0.63	5.12	0.89	6.63
3	0.76	4.19	0.89	2.93
4	0.61	3.15	0.92	2.00
5	0.76	2.93	0.92	2.23
6	0.65	6.91	0.90	2.00
7	0.08	38.25	0.79	6.82
8	0.29	15.13	0.70	12.85
Mean	0.50	11.97	0.86	4.93

Ablation studies on the impact of imaging modalities are also investigated by varying the input channels from all 3 modality (CT, T1ce and T2) to one channel only (CT). The mean DSC and HD95 scores of both the validation set and testing set are listed in the table 2.

Table 2 Mean Dice score (DSC) and 95th percentile Hausdorff distance (HD95) of the segmentation results with varied imput modalities.

Model	GTV				CTV			
	Val		Test		Test		Val	
	DSC	HD 95	DSC	HD 95	DSC	HD 95	DSC	HD 95
CT, T1ce, T2	0.43	13.14	**0.50**	**11.97**	**0.85**	**4.17**	**0.86**	**4.93**
CT, T1ce	**0.49**	**10.07**	0.47	13.30	0.85	4.48	0.86	5.47
CT	0.36	13.62	0.46	16.15	0.83	5.00	0.80	7.82

4 Discussion

Automatic delineation of spine SBRT target volumes can greatly alleviate the workload for neurosurgeons and radiation oncologists. Presently, these professionals are tasked with the manual delineation of treatment target volumes that involves reviewing multiple imaging modalities in tandem for each patient. Additionally, automated processes offer the advantage of greater consistency in target volume determination. By adhering to same clinical guidelines, these methods can minimize the intra- and inter-observer variability that typically arises from human interpretation, thereby enhancing the overall quality of treatment. In the current spine SBRT practice, GTV is the intact lesion, and CTV is derived from the intact lesion following the consensus guideline[6]. However, currently there is no available tool for the fully automated spine target volume delineation.

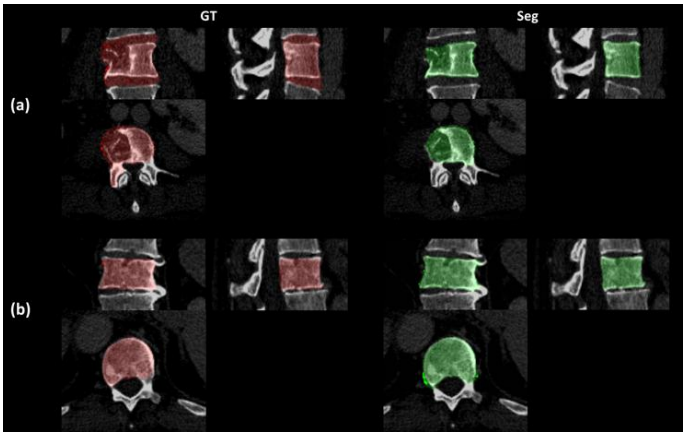


Figure 3 Examples of the CTV segmentation (Seg, in green) results comparing with the ground truth (GT, in red) contour on CT images.

Our proposed framework aims to streamline the spine SBRT treatment workflow by automating target localization and delineation facilitated by whole spine vertebrae segmentation and labeling. It is designed to mimic physicians' contouring practice to localize target position first then followed with a DL model to segment out the CTV by learning the rule-based expansion, to maximize the agreement between the model's prediction and current clinical contouring guidelines. In addition, the web-based platform interface design and the database management further improves the efficiency of the proposed framework. As shown from table 1, though with limited training data, the results on the unseen testing set suggest high agreement between the model prediction and physicians' manual CTV contours, with an averaged DSC and HD95 of 0.86 and 4.93 mm, respectively, indicating the promising performance of the proposed framework. On the other hand, the GTV segmentation accuracy is relative inferior, which is mainly caused by two limitations: 1) unlike CTVs that are usually defined by the rules, GTVs can vary a lot while our dataset is too small to learn such variations; 2) The quality and resolution of MRIs in this study is not optimal, which can greatly influence the GTV accuracy. Figure 3.b displays an example of the segmentation and the ground truth (the testing case #1) It can be found that the auto-segmentation agrees well with the manual contour. We also plot the worse testing case (case #8) in the figure 3.a. From the images we can see that in this case the ground truth CTV has a relatively rare expansion exceeding the vertebral body. Due to the lack of

samples of this type in the training data, the model couldn't learn this special CTV expansion and inclusion rule.

We also investigated the use of different imaging modalities in the spine SBRT CTV/GTV segmentation. From table 2 we can tell that, the use of T1c MRI can greatly help to improve the target delineation accuracy, while the contribution from T2 MRI images is relatively small. Our results agree with the current clinical practice, where T1c and CT images are often utilized for identify the lesion while T2 is usually used for spinal cord identification.

Additionally, the importance of target localization was also investigated. As shown in the figure 4.a, we shifted the target localization to include an extra adjacent vertebra in the model input for a testing case. Since CTV is usually derived from the lesion following certain inclusion rules, and the lesion in this case is very small, including one extra vertebra confused the model to generate a wrong CTV segmentation (in red) in the adjacent vertebra. The ground truth CTV mask is displayed in blue for comparison. Instead, the same patient but with the correct target localization, the DL model can easily generate accurate CTV segmentation (red mask in the figure 4.b) that agrees well with the ground truth.

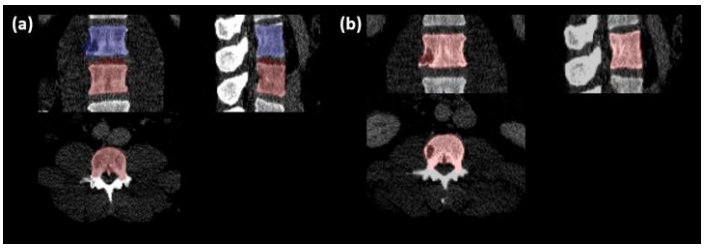


Figure 4 The impact of inaccurate target localization in the CTV detection and segmentation accuracy.

As a preliminary study, our work still holds several limitations. First, the data size of this study is very limited, which can greatly restrict the model ability. Currently, we are expanding the dataset to further improve the performance. Secondly, in the currently framework, lesion vertebra location was manually fed into the platform by extracting the key words from the clinical diagnosis chart following the institutional naming convention (e.g. "T03"), to enable the localization of the target in the CT and MRI images. As the ongoing work, we are automating this step with natural language processing (NLP) to further streamline the proposed framework.

5 Conclusion

In this study, we proposed a deep learning-driven framework to enable the automatic localization and delineation of the GTV and CTV in spine SBRT using multi-modal imaging. The preliminary results demonstrated promising result for the accurate target volume delineation. In addition, the framework is supported by a comprehensive web platform offering whole spine and OARs segmentation and efficient data management. It can be a promising tool to streamline the current spine SBRT workflow.

References

[1] Wong, David A., Victor L. Fornasier, and I. A. N. MacNab. "Spinal metastases: the obvious, the occult, and the impostors." Spine 15.1 (1990): 1-4.

[2] Massicotte, E., Foote, M., Reddy, R., & Sahgal, A. (2012). Minimal access spine surgery (MASS) for decompression and stabilization performed as an out-patient procedure for metastatic spinal tumours followed by spine stereotactic body radiotherapy (SBRT): first report of technique and preliminary outcomes. Technology in cancer research & treatment, 11(1), 15-25.

[3] Lutz, S., Balboni, T., Jones, J., Lo, S., Petit, J., Rich, S. E., ... & Hahn, C. (2017). Palliative radiation therapy for bone metastases: Update of an ASTRO Evidence-Based Guideline. Practical radiation oncology, 7(1), 4-12.

[4] Yamada, Y., Bilsky, M. H., Lovelock, J., ... & Fuks, Z. (2008). High-dose, single-fraction image-guided intensity-modulated radiotherapy for metastatic spinal lesions. International Journal of Radiation Oncology* Biology* Physics, 71(2), 484-490.

[5] Husain, Z. A., Sahgal, A.,... & Ryu, S. (2017). Stereotactic body radiotherapy for de novo spinal metastases: Systematic review: International Stereotactic Radiosurgery Society practice guidelines. Journal of Neurosurgery: Spine, 27(3), 295-302.

[6] Cox, B. W., Spratt, D. E., Lovelock, M., Bilsky,... & Yamada, Y. (2012). International Spine Radiosurgery Consortium consensus guidelines for target volume definition in spinal stereotactic radiosurgery. International Journal of Radiation Oncology* Biology* Physics, 83(5), e597-e605.

[7] Cox, B. W., Spratt, D. E., Lovelock, ... & Yamada, Y. (2012). International Spine Radiosurgery Consortium consensus guidelines for target volume definition in spinal stereotactic radiosurgery. International Journal of Radiation Oncology* Biology* Physics, 83(5), e597-e605.

[8] Giaj-Levra, N., Figlia, V., Cuccia, F, ... & Alongi, F. (2021). Reduction of inter-observer differences in the delineation of the target in spinal metastases SBRT using an automatic contouring dedicated system. *Radiation Oncology*, *16*(1), 1-6.

[9] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 25.

[10] Russakovsky, O., Deng, J., Su, H.,... & Fei-Fei, L. (2015). Imagenet large scale visual recognition challenge. International journal of computer vision, 115, 211-252.

[11] Vania M, Mureja D, Lee D. Automatic Spine Segmentation from CT Images Using Convolutional Neural Network via Redundant Generation of Class Labels. J Comput Des Eng (2019) 6(2):224–32.

[12] Kolařík M, Burget R, Uher V, Říha K, Dutta MK. Optimized High Resolution 3D Dense-U-Net Network for Brain and Spine Segmentation. Appl Sci (2019) 9(3):404.

[13] Wasserthal, J., Breit, H. C., Meyer, M. T, ... & Segeroth, M. (2023). Totalsegmentator: Robust segmentation of 104 anatomic structures in ct images. Radiology: Artificial Intelligence, 5(5).

[14] Isensee, F., Jaeger, P. F., Kohl, S. A., Petersen, J., & Maier-Hein, K. H. (2021). nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods, 18(2), 203-211.

[15] Yang, Z., Liu, H., Liu, Y., Stojadinovic, S., Timmerman, R., Nedzi, L., ... & Gu, X. (2020). A web- based brain metastases segmentation and labeling platform for stereotactic radiosurgery. *Medical physics*, *47*(8), 3263-3276.

Safeguarding Clinical AI-Based Auto-Segmentation: Open-Source Models as Quality Assurance Against Catastrophic Failures

Daniel Alexander, PhD[1] and Rafe McBeth, PhD[1]

[1]Department of Radiation Oncology University of Pennsylvania, Philadelphia, PA

Abstract This purpose of this work is to propose the use of an open source whole body CT segmentation model as a secondary independent check for validating outputs from a commercial auto-segmentation tool in radiation therapy planning. Evaluating uncertainty and comparing contours in 52 prostate cancer cases, we observe robust overall reliability with notable challenges in rectum contouring. Agreement is strong between the independent and clinical models for bladder and femoral heads (>0.9 DSC for >97% of patients), with moderate performance for prostate. Advocating for an automated secondary model check during contour generation, we emphasize the need for a verification of model output for the continued quality assurance of the planning workflow, aiming to enhance accuracy and confidence in AI-assisted radiation therapy planning.

1 Introduction

Anatomical delineation of organs-at-risk (OARs) on a volumetric image is a critical step in the workflow of generating treatment plans in radiation therapy. Conventionally this task is performed by physicians or trained treatment planners and is a time-intensive process. With the recent rise in popularity in the use of deep learning-based auto-segmentation solutions to reduce the time and resource burden on clinical staff, as well as to improve the consistency and standardization of clinical contouring, there is a need for robust verification of the output of these deep learning models.

It is understood that variability in contours is one of the largest sources of uncertainty in radiation therapy, due to both intra- and inter-observer variability, as well as contouring error[1, 2]. There have been proposed methods to perform quality assurance for traditional human-driven OAR contouring processes primarily by comparing clinical segmentations against a historical knowledge base [3, 4]. With the rise in AI-based tools in radiation therapy, including auto-segmentation, there have been separate suggestions for corresponding quality assurance methods [5, 6]. In this study, we suggest an open source whole body CT segmentation model based on the popular nnU-Net architecture, TotalSegmentator, to function as a secondary, independent model to cross-check the output of a commercial deep learning-based auto-segmentation tool. [7, 8]. Similar to secondary, independent dose calculation algorithms used in clinical practice to identify gross errors in monitor unit calculations, we believe that independent deep learning models trained on unrelated datasets can be used as a second check on clinically-deployed AI solutions to serve a similar purpose.

2 Methods

Planning images and structure sets for a cohort of 52 patients who underwent external beam radiation therapy for prostate cancer at out institution were anonymously exported from our treatment planning system database for contour analysis and inference. All planning CTs for this cohort had been submitted to our commercial auto-segmentation tool (ART-Plan, Therapanacea, Paris, France) as part of the standard clinical workflow, and these structure sets were exported alongside the planning CTs in DICOM format. All DICOM data was converted into NifTi format to facilitate inference by the pre-trained TotalSegmentator model and analysis. Five clinical structures were identified to evaluate the performance of the independent model: bladder, rectum, prostate, and right and left femoral heads.

2.1 Assessing prediction uncertainty

In order to evaluate the uncertainty in contour predictions from TotalSegmentator, the CT data for each patient was manipulated via translations along each axis of +4 and +8 voxels. Theses translated image volumes were submitted to the pre-trained model along with the original image volume, and resulting predictions were processed with the reversed translations of -4 and -8 voxels along each axis, respectively. This resulted in seven unique predictions for each structure per patient. Each of the predictions from the translated image volumes was compared against the prediction from the original image volume by calculating Dice similarity coefficients (DSC) for each pair. Additionally, in order to assess the uncertainty in the model output as a function of anatomical direction, a composite prediction was formed by summing the binary masks from the seven unique predictions, and the gradients along the x, y, and z-dimensions were calculated as shown in Figure 1. For each dimension, line profiles were evaluated through each point in the plane perpendicular to the gradient direction, the positive and negative peaks were identified, and the peak heights and widths were aggregated. The average width of the positive and negative peaks were interpreted as uncertainty in the border of the prediction along that axis; for example, for peak gradient widths along the z-axis, positive and negative average peak widths corresponded to uncertainty in superior and inferior borders, respectively.

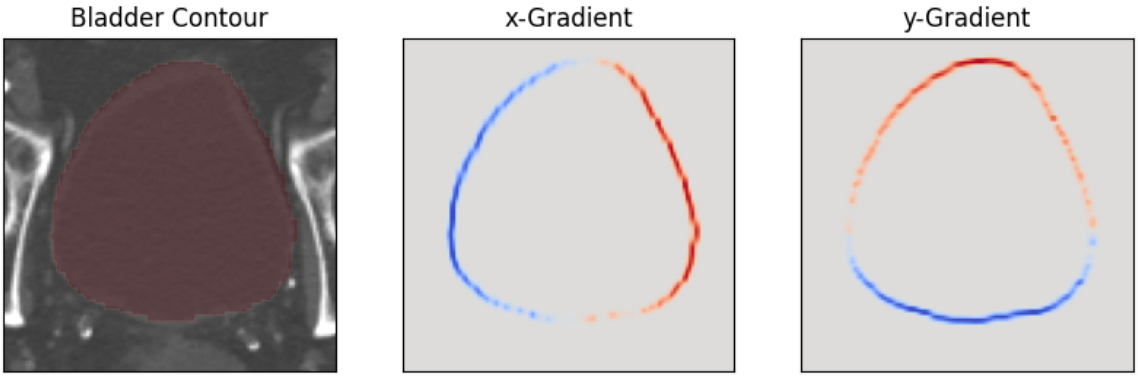


Figure 1: Analysis of composite mask gradients along different axis. Left: Bladder contour in red overlaid on CT image slice. Center: Resulting x-gradient of composite bladder prediction. Right: Resulting y-gradient of composite bladder prediction.

2.2 Comparison of clinical and independent contours

Baseline predictions on original planning image volumes from the independent model were compared with the clinical structures generated by the commercial system for the five structures listed above. DSC and Hausdorff distance metrics were calculated between the structure pairs for each patient, and aggregate data was analyzed.

3 Results

3.1 Prediction uncertainty

Table 1 shows the mean and standard deviation of the DSSC values for each structure, averaged across the patient cohort.

Structure	Average Mean	Average Std Dev
Bladder	0.987696	0.004095
Femur_Head_L	0.981661	0.006421
Femur_Head_R	0.981555	0.006311
Prostate	0.965240	0.009722
Rectum	0.960647	0.013215

Table 1: Average Mean and Standard Deviation Across All Patients for Each Structure.

The worst performing structure across seven independent predictions was the rectum contour, while the best performing was bladder; however, all structures were fairly stable at DSC > 0.95 and a standard deviation of < 1.5%. Furthermore, Table 2 shows the gradient widths, a surrogate for prediction uncertainty, along each anatomical direction for the five evaluated contours. Similarly, the rectum shows the greatest uncertainty overall.

3.2 Inter-system contour comparison

Figure 2 highlights a comparison of the contours produced by the clinically deployed auto-segmentation tool and the independent secondary model. Calculated Hausdorff distances for the evaluated contours were 36.9 ± 29.6, 11.3 ± 3.8, 3.0 ± 1.0, 3.0 ± 0.7, and 5.3 ± 2.5 mm for the rectum, prostate, right femoral head, left femoral head, and bladder, respectively (Median ± IQR). Corresponding DSC values were 0.66 ± 0.15, 0.76 ± 0.07, 0.97 ± 0.01, 0.97 ± 0.01, and 0.96 ± 0.02, respectively. For the bladder and femoral heads, the success rate was >97% for a 0.9 DSC criteria, while the prostate contours has worse agreement, with a 0.65 criteria yielding a 95% success rate. Again, the rectum comparison showed the worst agreement with only 86% of patient contours meeting a DSC threshold of 0.5.

4 Discussion

The uncertainty quantification performed on the TotalSegmentator predictions highlights the fairly robust reliability of the model. However, due to the somewhat arbitrary nature of the labels produced based on the paradigm in the training data, especially with the rectum structure, there continued to be disagreements with the clinical autocontouring tool. Clinical deployment of a pre-trained secondary model would necessitate thoughtful postprocessing of the output to align with the clinical standards in place.

We propose a suggested workflow to enhance the clinical integration of a "second check" quality assurance approach for AI based contouring. In this envisioned workflow, a database of benchmark metrics would be established to determine action and tolerance levels for contour agreement based on statistical process control [9]. After deployment, the secondary model would be automatically invoked at the time of clinical contour generation. The contours produced by the commercial deep learning-based auto-segmentation tool would be simultaneously cross-verified by the independent model and evaluated in the context of the preestablished action level. If necessary, the contours would then be made available to the planner and physicist for visual examination prior to the planning phase. This proactive approach aims to catch potential inaccuracies and uncertainties early in the workflow, providing an additional layer of assurance and reinforcing the collaborative role of clinicians in the AI-assisted

Structure	Left	Right	Anterior	Posterior	Superior	Inferior
Bladder	2.046	2.021	2.103	2.143	2.296	2.271
Femur_Head_L	2.217	2.431	2.197	2.287	2.549	2.620
Femur_Head_R	2.490	2.207	2.197	2.262	2.504	2.601
Prostate	1.975	1.978	2.046	2.022	2.064	2.149
Rectum	2.400	2.320	2.232	2.350	3.162	2.719

Table 2: Average width (in voxels) of peaks in the gradients of the composite masks for each structure, related to uncertainty in the prediction.

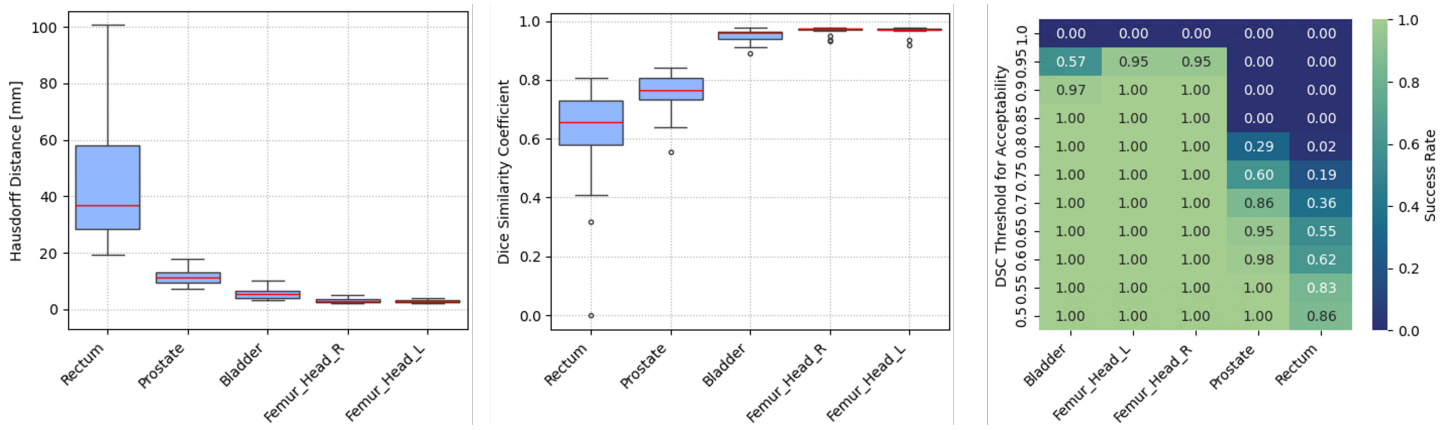


Figure 2: Quantitative comparison of contours generated from a clinically deployed AI autocontouring tool and an open source secondary independent model. Left: Distribution of Hausdorff distances across the patient cohort for each structure. Center: Distribution of DSC values across the patient cohort for each structure. Right: Success rates for each structure, defined for a given structure as the fraction of patients for which the DSC between the two structure predictions is above the threshold of acceptability.

radiation therapy planning process.

5 Conclusion

This study addresses the critical need for robust verification in the context of deep learning-based auto-segmentation solutions for radiation therapy planning. With the increasing reliance on AI tools to streamline and standardize the contouring process, a secondary independent model can serve as an cross-check on the output of an otherwise automated tool. The comprehensive evaluation conducted on a cohort of 52 prostate cancer patients provides insights into both the strengths and challenges of this approach.

The assessment of prediction uncertainty through manipulation of CT data and subsequent analysis of model outputs reveals a generally robust reliability of the model. Despite its overall performance, discrepancies highlight the nuanced nature of aligning model outputs with clinical standards. This emphasizes the importance of thoughtful postprocessing when deploying pre-trained secondary models in a clinical setting.

In conclusion, our study contributes to the evolving landscape of AI applications in radiation therapy by introducing a potential verification strategy. The proposed workflow and insights into prediction uncertainty provide a foundation for further refinement and implementation of AI-driven contouring solutions, with the goal of ultimately reliability of radiation therapy planning in the era of deep learning.

References

[1] M. G. Jameson, L. C. Holloway, P. J. Vial, et al. “A review of methods of analysis in contouring studies for radiation oncology”. *Journal of medical imaging and radiation oncology* 54.5 (2010), pp. 401–410.

[2] S. K. Vinod, M. G. Jameson, M. Min, et al. “Uncertainties in volume delineation in radiation oncology: a systematic review and recommendations for future studies”. *Radiotherapy and Oncology* 121.2 (2016), pp. 169–179.

[3] M. Altman, J. Kavanaugh, H. Wooten, et al. “A framework for automated contour quality assurance in radiation therapy including adaptive techniques”. *Physics in Medicine & Biology* 60.13 (2015), p. 5199.

[4] C. B. Hui, H. Nourzadeh, W. T. Watkins, et al. “Quality assurance tool for organ at risk delineation in radiation therapy using a parametric statistical approach”. *Medical physics* 45.5 (2018), pp. 2089–2096.

[5] M. Claessens, C. S. Oria, C. L. Brouwer, et al. “Quality assurance for AI-based applications in radiation therapy”. *Seminars in radiation oncology*. Vol. 32. 4. Elsevier. 2022, pp. 421–431.

[6] M. Claessens, V. Vanreusel, G. De Kerf, et al. “Machine learning-based detection of aberrant deep learning segmentations of target and organs at risk for prostate radiotherapy using a secondary segmentation algorithm”. *Physics in Medicine & Biology* 67.11 (2022), p. 115014.

[7] F. Isensee, P. F. Jaeger, S. A. Kohl, et al. “nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation”. *Nature methods* 18.2 (2021), pp. 203–211.

[8] J. Wasserthal, H.-C. Breit, M. T. Meyer, et al. “Totalsegmentator: Robust segmentation of 104 anatomic structures in ct images”. *Radiology: Artificial Intelligence* 5.5 (2023).

[9] T. Pawlicki, M. Whitaker, and A. L. Boyer. “Statistical process control for radiotherapy quality assurance”. *Medical physics* 32.9 (2005), pp. 2777–2786.

Automatic target segmentation and beam setup for electron skin radiotherapy

Hendrik Westendorp[1], Miranda Beumer[2], Tristan van Heijst[1], and Kelvin Ng Wei Siang[3, 4]

[1]Isala, Department of Radiotherapy, Zwolle, The Netherlands
[2]Saxion University of Applied Sciences, Department of Applied Physics, Enschede, The Netherlands
[3]Erasmus MC Cancer Institute, University Medical Center Rotterdam, Department of Radiotherapy, The Netherlands
[4]Holland Proton Therapy Center, Department of Medical Physics & Informatics, Delft, The Netherlands

Abstract Accurate superficial target segmentation on a CT scan is challenging task, since the 2D segmentations must accurately follow the 3D surface curvatures projected onto orthogonal planes. For radiation oncologists this is a considerable time investment.
We present a technique to automatically segment a target from a wire placed on the patient's skin using a CT scan. This method furthermore allows for automatic beam and couch angle determination as well as placement of the treatment plan reference point as defined by fiducial markers.
Multiple wire materials were considered (Al, Cu, Sn) to define the CTV borders on the skin surface; from these, Sn proved to be superior with respect to detectability on CT.
A phantom study was performed to assess the properties of the wires. The performance of the detection and segmentation algorithm were tested and evaluated on 25 patients. The algorithm correctly detected all fiducial makers, wires and other objects.

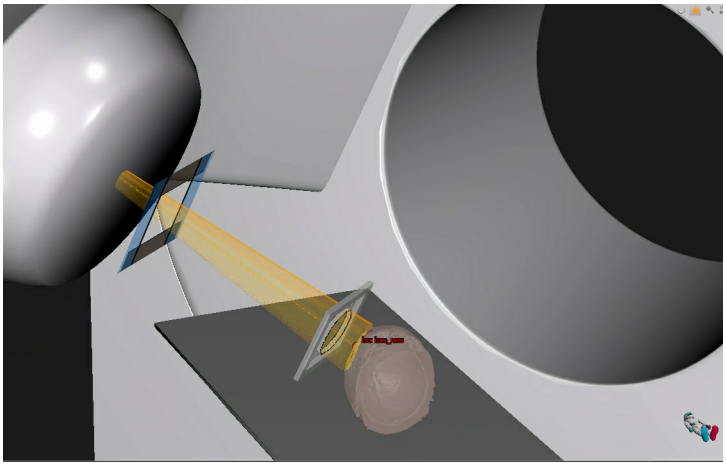


Figure 1: Example of an automatically determined optimal beam and couch angle.

1 Introduction

Electron radiotherapy is an established treatment for non melanoma skin cancer [1]. In recent years, Monte Carlo electron planning has become widely available to accurately calculate dose for electron treatments.
The transition from classic, water tank measurement based, isodose atlases to the current (eMC) practice has a considerable impact on the clinical workflow. In the classical workflow, the radiation oncologist defines the visible target area (either the gross target volume (GTV) or clinical target volume (CTV)) on the skin of the patient using a marker. This area is then used to select the inlay for the electron applicator, either from existing templates or customized through casting for the electron applicator. The beam and couch angles are then defined visually by approximating the beam direction with the projected normal of the target. Due to the difficulty of defining these angles visually, the setup angles can vary per treatment fraction.
Treatment planning using eMC requires high quality Computed Tomography (CT) image data to accurately calculate the dose for the specific anatomy of the patient. The radiation oncologist places a wire on the skin to define the CTV boundary, so the CTV at the surface is visible at the CT. However, accurate segmentation of superficial target volumes on a CT is often challenging and time-consuming for radiation oncologists, who have to segment the CTV on 2D planes while adhering to the 3D skin surface curvatures projected onto the orthogonal planes. Moreover, these CT images often show poor contrast between target and healthy tissues, which increases the difficulty of accurate target segmentation.
In this work, we developed a method to automatically determine the treatment area, starting with a wire around the visible lesion. Using this method, we were able to accurately determine the reference point in the CT dataset, as defined by fiducial markers. Furthermore, using the detected wire on the CT, we could easily determine the optimal beam and couch angle for the treatment.

2 Materials and Methods

2.1 Wire properties

Multiple wires were first tested for robust detectability on a CT scan. The wires should be non-toxic for the human skin and should be easily shaped by the radiation oncologist so to follow the often, highly irregular target surfaces, e.g. in the nasal region.
Copper (Cu), aluminium (Al) and tin (Sn) were considered candidates. Table 1 shows the properties of the wires that were measured. CT scans were obtained with 1 mm thick slices with matrices of 512^2 or 1024^2 pixels. Since a large fraction of non melanoma skin radiotherapy patients have lesions on the skull, the wire should show sufficient contrast with high density bone tissue.
We developed a anthropomorphic skull phantom to test the contrast of the wires. High density plaster ($\pm$ 1500 HU)

Material	Z	Density (g/cm^3)	Diameters (mm)	
Al	13	2.70	0.56	0.90
Cu	29	8.96	0.50	1.06
Sn	50	7.26	0.30	0.50

Table 1: Wires of multiple densities and atomic number were considered. For each material two diameters were used. Note the approximate doubling of atomic number.

mimicked cortical bone ($\pm$ 1300 HU). A thin layer of silicone casting rubber ($\pm$ 140 HU), represented surrounding skin tissue ($\pm$ 100 HU). The contrast of the wire with respect to bone tissue was determined using equation 1

$$\text{contrast} = \frac{I_{\text{wire}} - I_{\text{skull}}}{I_{\text{skull}}}. \tag{1}$$

Artefacts were visually assessed.

2.2 Wire detection

A stepped, region growing wire detection algorithm was integrated in the treatment planning system (TPS, RayStation 12A, RaySearch, Sweden) using Python 3.8. As a standard procedure, the external contour of the patient is outlined on the CT in the TPS. A stepwise, region growing algorithm proved to be compatible with the functionality of the TPS. An initial lookup volume for the wire extended isotropically 2 mm inside and 20 mm outside the external contour.
In the lookup volume, the highest densities ($>$ 2300 HU) were contoured using the gray level threshold functionality of the TPS. An isotropic expansion of this structure of 5 mm formed a next lookup volume. In this volume a threshold of 1800 HU generated an expansion for the wire. The new structure was also limited to 2 mm within the external contour. In the last step, a 3 mm expansion of the resulting structure served as a lookup volume for voxels with a density $>$ 1150 HU. This resulting volume was limited within 2 mm of the external contour. A final expansion and contraction of 6 mm closed remaining small holes.
The stepped, region growing threshold algorithm identified all high metal objects near the skin surface, such as fiducial markers, wires and other objects (e.g. metallic dental implants). Each object has to contain voxels with a density $>$ 2300. Non metal object do not contain densities in the order of 2000 HU. Consequently, the stepped detection process prevented the high density cortical bone to be misidentified as metallic structure, while still identifying connected voxels with a lower density, because of the partial volume effect, as part of the high density structure.
The single high density structure needed to be split in separate parts, with one metalic item each. The high density structure (near the skin) was transformed to a triangle mesh. In this representation connected indices of the triangle points form the separate parts of the structure and can be split using this connections.

Categorie	V_{mesh} [cm^3]	V_{filled}	Z-position [cm]
Fiducial marker	< 0.03	> 0.7	$-0.5 < Z < 0.5$
Wire	-	< 0.1	-
Other object	-	-	-

Table 2: Criteria to assign a part of the high density structure to a category.

Finally, the parts were divided in three categories. To categorize parts, we made use of the concept of filled volume, V_{filled}) (equation 2). Where r_{max} is the maximum distance to the Center of Mass (COM) of all points belonging to a connected part.

$$\begin{aligned} V_{\text{filled}} &= \frac{V_{\text{mesh}}}{\frac{4}{3}\pi r_{max}^3} \\ &\text{where} \\ V_{\text{mesh}} &= \sum_{t \in T} \frac{\mathbf{o}_t \cdot (\mathbf{v}_{t,1} \times \mathbf{v}_{t,2})}{6}. \end{aligned} \tag{2}$$

T represents the set of triangles. The vector from the first point of triangle t to the origin is represented by $\mathbf{o}_t$ and the vectors from the first point to the two other points by $\mathbf{v}_{t,1}$ and $\mathbf{v}_{t,2}$. In table 2 the criteria for the assignment to a category are defined.

2.3 Isocenter, beam and couch angle

The points of the triangle mesh representation of the wire served as input for the calculation of the gantry and couch angle. First, a plane was fitted through the points using the least–squares method as shown in equation 3

$$\begin{aligned} ax + by + c = z \rightarrow Ax = B \\ \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \mathbf{n} = \left(A^T A\right)^{-1} A^T B. \end{aligned} \tag{3}$$

The vector $\mathbf{n}$ defines the normal to the fitted plane and was expressed as a beam and couch angle by expressing the vector in spherical coordinates, correcting for the patient scan position. Two solutions were found; the solution that had a couch angle between 100° and 260° was selected.
Subsequently, the COM of the wire was set as the entrance point of the beam. The isocenter was identified and placed on the surface of the patient by searching for a drop in HU values, resulting in a source skin distance of 100 cm. The HU values were taken by the outward pointing $\mathbf{n}$, crossing the COM. A threshold of -250 HU was taken, where a value of 0 was assigned if below the threshold and 1 if above. The surface of the patient was identified as the furthest gradient step. Figure 1 shows an example of an automatically determined isocenter (where the center of the beam crosses), beam and couch angle.

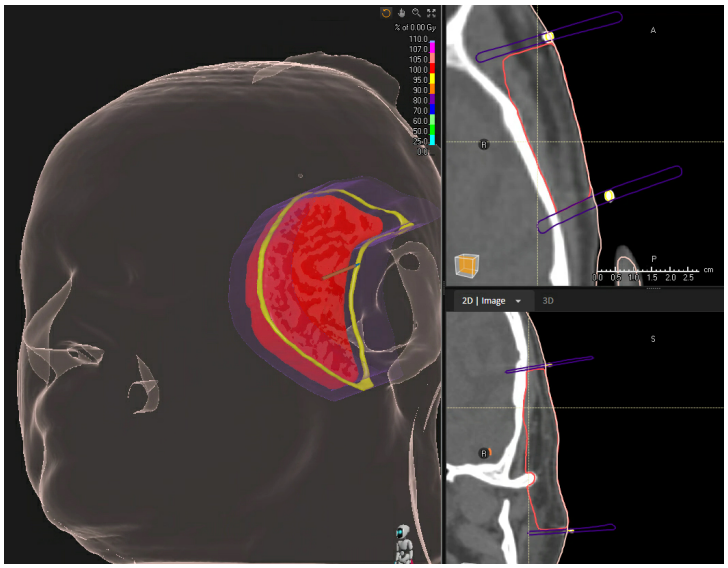


Figure 2: The wire (yellow) is expanded (purple) in the beam direction (arrow) and marks the lateral border of the CTV (red). The CTV depth is prescribed at 1 cm, limited to bony structures.

Finally, the reference point could be determined by taking the COM of the fiducial markers. The first detected fiducial marker was taken as the reference point. In case of two fiducial markers, the most anterior marker defined the lateral position, and the most posterior determined the height. With three located fiducial markers, the mean of the two closest in height defined the height, the other lateral position. The average of all markers was taken as the inline position.

2.4 Target segmentation

The radiation oncologist defined the CTV on the skin by placing the wire around the lesion. The wire was segmented using the algorithm of paragraph 2.2. Together with the optimised beam and couch angles the wire contour served as input for target segmentation.
An expansion of the wire in the direction of the beam marked the surface border of the target area. The radiation oncologist prescribed the depth of the target, eventually limited by organs at risk (OAR). Some OARs such as bone and eyes can be automatically segmented by the TPS. Using standard TPS functionality (margin operations), the CTV was obtained. An example of the target definition is shown in figure 2.

3 Results

The phantom study on Al, Cu and Sn showed that Al did not exceed the density of (cortical) bone, rendering it ineligible for further study. Cu and Sn showed good contrast in the phantom study.
The algorithm developed was tested on 25 patients treated previously at our clinic. All objects were correctly categorized by the algorithm (table 3). Cu wires were occasionally incorrectly identified as two parts (Patients 5, 7, 12 and 17). All Sn wires detected were closed.

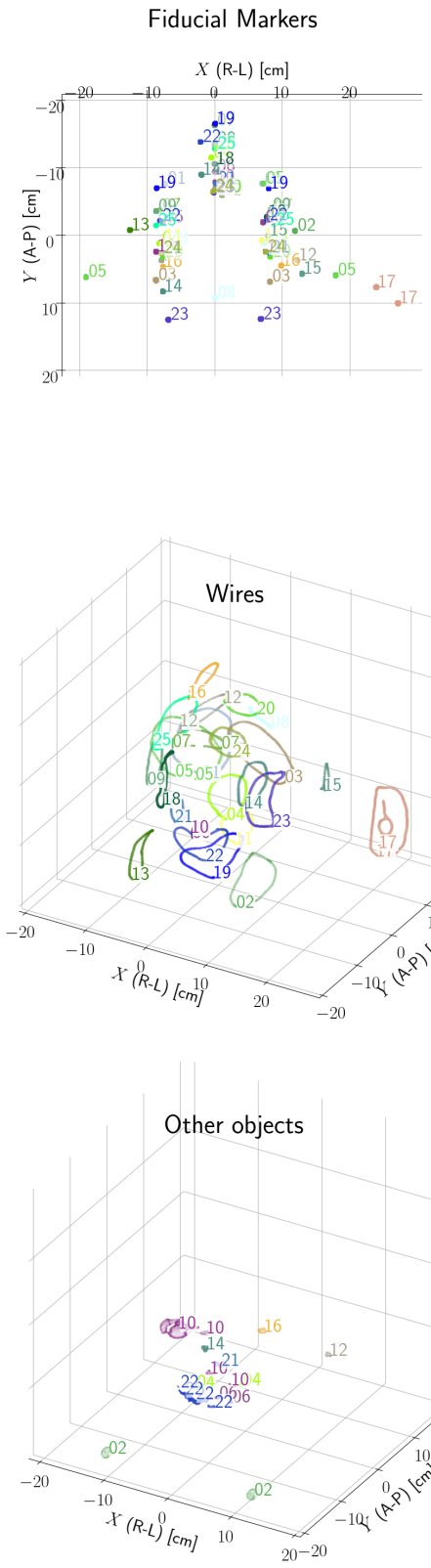


Figure 3: The algorithm correctly detected all fiducial makers, wires and other objects in the set of 25 cases.

Patient	Material	Number of objects detected			Other object type
		Wires	Fiducial markers	Other objects	
1	Cu	1	3	-	
2	Cu	1	1	2	Support structure
3	Cu	1	3	-	
4	Sn	1	3	2	Dental
5	Cu	2	3	-	
6	Cu	1	3	2	Dental
7	Cu	2	3	-	
8	Cu	1	3	-	
9	Cu	1	3	-	
10	Cu	1	3	4	Dental
11	Cu	1	3	-	
12	Cu	2	3	1	Dental
13	Sn	1	1	-	
14	Sn	1	2	1	Dental
15	Cu	1	2	-	
16	Sn	1	3	1	Dental
17	Cu	2	2	-	
18	Sn	1	2	-	
19	Sn	1	3	-	
20	Sn	1	3	-	
21	Sn	1	3	1	Dental
22	Sn	1	3	4	Dental
23	Cu	1	3	-	
24	Sn	1	3	-	
25	Sn	1	3	-	

Table 3: Number of objects identified in each category; the majority of the other objects identified dental

Objects that were not wire or fiducial marker were mostly mostly dental. In one case, the TPS included the couch and supporting structures in the external contour. Parts of this support were categorised as other, metallic, object.
In figure 3 all categorised objects are shown. The target segmentation of the CTV gave consistent results. In some cases the depth of the CTV was adjusted according to anatomical information. In all 25 patient scans, the reference point was correctly identified.

4 Discussion

The introduction of eMC dose calculation has greatly changed the treatment planning and delivery of electron radiotherapy for superficial lesions.Target definition on a CT scan has become more precise but less straightforward. We presented a method to automatically segment the target. This method takes intuitively the beam as a starting point and generates a target region in the direction of the beam. The proposed algorithm proved to be reliable in all 25 clinical situations.
Cu showed limitations as a closed wire was split into two parts in 4 of 14 cases. Sn proved to be the optimal material for the wire detection and segmentation of the CTV. Besides the excellent detectability on CT, the material can easily be shaped to conform to complex geometries of the CTV. Sn wire is easily available for limited cost.

The automatic detection and segmentation process creates CTV contours where the radiation oncologist can evaluate on and adjust where necessary. This saves considerable time in the clinical workflow process, which enables a more efficient workflow. The next step of this work will evaluate the accuracy of the CTVs created by this process with input from our radiation oncologists.
The presented auto segmentation methodology and algorithm can be extended to automate the definition and segmentation of a standard or patient specific bolus on the CT image. Methods to design [2] and construct [3] a patient specific bolus have been earlier proposed.

5 Conclusion

We have presented a robust method to automatically and accurately detect the wires and segment the CTV for electron radiotherapy. Optimal beam and couch angles are determined using our algorithm. The Sn wire performs better than Cu wire on its detectability on a CT.

References

[1] S. Benkhaled, D. Van Gestel, C. Gomes da Silveira Cauduro, et al. “The State of the Art of Radiotherapy for Non-melanoma Skin Cancer: A Review of the Literature”. Frontiers in Medicine 9 (June 2022). DOI: 10.3389/fmed.2022.913269.

[2] S. Su, K. Moran, and J. L. Robar. “Design and production of 3D printed bolus for electron radiation therapy”. Journal of Applied Clinical Medical Physics 15.4 (July 2014), pp. 194–211. DOI: 10.1120/jacmp.v15i4.4831.

[3] R. A. Canters, I. M. Lips, M. Wendling, et al. “Clinical implementation of 3D printing in the construction of patient specific bolus for electron beam radiotherapy for non-melanoma skin cancer”. Radiotherapy and Oncology 121.1 (Oct. 2016), pp. 148–153. DOI: 10.1016/j.radonc.2016.07.011.

Comparative Analysis of Deep Learning Models for Flap Segmentation in Head and Neck Computed Tomography Images: A Generalization Study

Abir Fathallah[1], Alice Blache[2], Zacharia Mesbah[3], Romain Modzelewski[3], Juliette Thariat*[4], and Mathieu Hatt*[1]

[1]LaTIM, INSERM, UMR 1101, University Brest, Brest, France
[2]Department of Radiation Oncology, CHU Amiens, Amiens, France
[3]LITIS - EA4108 - Quantif, Rouen, France
[4]Department of Radiation Oncology, centre François-Baclesse, Caen, France
*contributed equally

Abstract Segmentation of flaps plays a pivotal role in the treatment planning of radiotherapy for patients with head and neck cancer. Accurate and robust flap segmentation could significantly enhance the accuracy and safety of radiotherapy. In this work, a comparative study was carried out using two distinct models for fully automated medical image segmentation: the nnUNet framework, leveraging the U-Net convolutional neural network architectures, and MedSAM, founded on transformers. Both model were previously trained and validated using a well-curated GORTEC (Groupe d'Oncologie Radiothérapie Tête Et Cou) prospective multicentric cohort. The present work focuses on the external validation of trained models in a "real-life" external cohort. As it was already observed in the previous study, the nnUNet outperformed MedSAM (Mean Dice score of 55 versus 74). The performance of nnUNet was lower than it was in the test set of the previous study , although the drop in performance is mostly due to the fact that the method works well on a majority of cases but completely fails in some studies that suffer from artifacts or with flaps with previously unseen or rare locations/characteristics. Our future work will focus on improving the performance of the segmentation in these cases.

1 Introduction

Head and neck cancer (HNC) constitutes a complex oncological domain where therapeutic interventions, particularly radiation therapy (RT), require accuracy and robustness to reach an optimal balance between therapeutic efficacy and sparing of healthy tissues in organs at risk (OAR). The efficacy of RT is contingent upon the accurate delineation of anatomical structures. Flaps, comprising diverse tissues within the intricate head and neck region, exhibit inherent morphological intricacies, therefore posing a challenge for conventional segmentation methodologies applied to Computed Tomography (CT) three-dimensional (3D) images.

The requirements for accurate and robust flap segmentation is underscored by its direct impact on the comprehensive analysis of dose-volume effects, thereby influencing treatment planning and outcome prognostication. Traditional segmentation paradigms manifest inadequacies in addressing the intricate morphological variations characterizing flaps, prompting the exploration of advanced computational methods, particularly those rooted in Deep Learning (DL) frameworks.

The application of DL is widespread in the field of medical imaging, particularly in the context of RT planning. Defining target volumes and identifying OAR during RT planning is a complex task, given that the native anatomy undergoes modifications due to tumor resection, resulting in significant tissue defects and post-operative changes such as inflammation, edema, fibrosis, and the presence of surgical clips.

For an accurate assessment of the causality of radiation-induced deterioration in the functional outcomes of reconstructive flap surgery, it is imperative to achieve fully automatic flap delineation. This process aims to address issues such as flap atrophy and fibrosis. However, the delineation of flaps in CT 3D images poses a novel, challenging, and time-consuming task for radiation oncologists.

Recent advancements in DL have demonstrated superior performance in the segmentation of complex 3D images compared to traditional computer vision techniques [1]. While segmentation models are increasingly prevalent in the analysis of CT or magnetic resonance imaging (MRI) scans, their primary application has been in discriminating the presence or absence of tumors or cancers [2].

This study endeavors to contribute to the field by presenting segmentation and evaluation techniques specifically tailored for flap analysis in the context of reconstructive surgery and RT. The focus is on addressing the challenges associated with flap delineation and evaluating the performance of the proposed techniques.

2 Materials and Methods

2.1 Datasets

Two distinct datasets were relied upon. The first one served the dual purpose of training and testing our segmentation models. It contained DICOM-RT data (CT images and manual contours) pertaining to poRT planning, sourced from a GORTEC (Groupe d'Oncologie Radiothérapie Tête Et Cou)) prospective study. A total of 152 patients, recruited across 33 centers between 2019 and 2021, were included in the study, with the exclusion of 4 patients due to the missing CT slices. The remaining 148 patients' CT and RT structure sets encompassed tumor volumes and OAR information. The delineation of ground truth flap contours adhered to established guidelines [3] and was conducted on Raysearch station

v8.0. Importantly, the contouring process was executed independently by two experts, who were blinded to patient characteristics and outcomes.
The second imaging dataset exclusively served as an external testing set, for evaluating the generalization capabilities of the models. This dataset comprised poRT planning DICOM-RT data sourced from Amiens CHU, involving 70 patients. Similar to the first dataset, the creation of ground truth flap contours followed the delineation guidelines [3] and was performed on Raysearch station v8.0. The contouring process in this instance was also conducted independently by two experts unaware of patient characteristics and outcomes during the procedure. Because this dataset was sourced from routine clinical practice data, it can be considered a "real life" dataset, in contrast to the more curated data of the first dataset, collected in a prospective study.

2.2 nnUNet Segmentation Model

The nnU-Net framework was employed for the training of a pre-configured generic U-Net [4] on our first (GORTEC) dataset [5]. A random partitioning of the dataset into training and test sets was performed using an 80/20 split, resulting in 120 patients for training and 28 patients for testing. All images were uniformly resampled to a voxel size of $1.17\times1.17\times2mm^3$, selected as the median voxel size within the dataset. Resampling was carried out using 3rd-order spline interpolation for images and nearest-neighbor interpolation for Ground Truth masks.
Data augmentation techniques encompassed the application of nnU-Net augmentation methods, including random flips, affine transformations (rotation, translation, and/or scaling applied to the volume), elastic deformation (deformation along a grid mesh) of both CT scans and segmentation maps, brightness and contrast adjustments, as well as image quality reduction.
Standard parameters of the nnUnet framework were utilized.

2.3 MedSAM Segmentation Model

The Medical Segment Anything Model (MedSAM) [6]is derived from the open-source Segment Anything Model (SAM) [7], developed by Meta AI. MedSAM is applied to medical image segmentation. It adeptly processes both two-dimensional (2D) and three-dimensional (3D) images, with the caveat that 3D images are processed as independent 2D slices.
In our efforts to fine-tune MedSAM for the specific task of delineating flaps in CT images, we adhered to the network architecture delineated in SAM [7]. This architecture comprises an image encoder and a mask decoder. The image encoder [8] utilizes positional encoding [9] to map the input image into a high-dimensional image embedding space. Following this, the mask decoder integrates the image embedding and prompt features through cross-attention [10].
The network employed in our investigation is built upon the transformer architecture [10], incorporating a Vision Transformer (ViT)-derived image encoder designed to extract features from images. Additionally, a mask decoder was implemented to generate segmentation outcomes and confidence scores, making use of the image embedding and output token. Striking a balance between segmentation efficacy and computational efficiency, we opted for the foundational ViT model as the image encoder.
During the data pre-processing phase, the dataset was partitioned with the same split as for the nnU-Net training and evaluation, relying on an 80/20 split ratio, resulting in distinct training and test sets encompassing 120 and 28 patients, respectively. The initialization of the model involved employing the pre-trained MedSAM model in conjunction with the ViT-Base model. All trainable parameters within the image encoder and mask decoder underwent updates throughout the training process.
The chosen loss function constituted an unweighted summation of Dice loss and cross-entropy loss, a combination proven to exhibit robustness across diverse segmentation tasks [11]. Optimization of the network was achieved through the implementation of the AdamW optimizer [12], configured with parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$, an initial learning rate set at $1e^{-4}$, and a weight decay of 0.01. The training regimen utilized a batch size of 160, and data augmentation was omitted. Employing Quadro P400 GPUs, the model underwent training for 2000 epochs, with the final model selected based on the last checkpoint.

2.4 Evaluation

Both trained models were then applied without any fine-tuning or changes in the parameters, on the second dataset from the CHU Amiens. Quantitative evaluation of performance was reported using the mean and median Dice score, as well as individual patient scores. In addition, the behavior of the algorithm with respect to image characteristics, such as the size and location of flaps, or the presence of metal artifacts, was qualitatively assessed.

3 Results

This section will present both quantitative and qualitative outcomes for the comparative assessment of MedSAM and nnUNet models in the context of automated flap segmentation.

3.1 Quantitative results

It is important to emphasize that both models were exclusively trained using the first dataset and evaluated the second dataset as an external validation set. In the initial training and evaluation of both models on the first dataset, the nnU-Net outperformed the MedSAM model (0.74 vs 0.55 mean Dice

and 0.87 vs 0.63 median Dice). In assessing the generalizability of segmentation models, the mean and median Dice scores, based on the external Amiens dataset, are presented in Table 1. MedSAM demonstrated a mean Dice coefficient of 0.35, whereas nnUNet exhibited a higher mean Dice coefficient of 0.62. Furthermore, nnUNet surpassed MedSAM in terms of the median Dice coefficient, with values of 0.66 and 0.37 respectively. Both methods exhibited a decrease of performance when applied to this new "real life" dataset (from mean Dice of 0.74 to 0.62 and median Dice from 0.87 to 0.66).

Model	mean DSC	median DSC
nnUNet	0.62	0.66
MedSAM	0.35	0.37

Table 1: Dice scores of nnU-Net and MedSAM on Amiens dataset.

MedSAM yielded an average DSC with a mean and median value of 0.35 and 0.37, respectively. This performance was deemed inadequate in comparison to the outcomes achieved by nnN-Unet. An examination of the MedSAM results exposed significant variations, with certain images exhibiting satisfactory delineation (DSC: 0.55-0.75), while others displayed inaccuracies (DSC < 0.10-0.20). This discrepancy suggests that the absence of 3D modeling particularly penalized the efficacy of MedSAM in the challenging task of 3D flap delineation. Figure 1 describes the observed behavior, illustrating the distribution of Dice scores calculated for each of the 70 patients (the entire Amiens dataset) according to the nnUNet and MedSAM models.

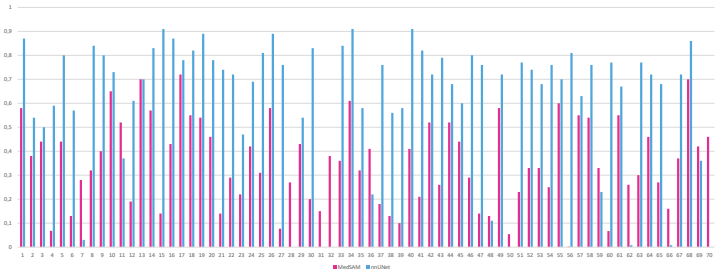


Figure 1: Distribution of Dice scores for the entire Amiens dataset according to nnUNet (blue) and MedSAM (pink) models

3.2 Qualitative results

Similarly as for the behavior of the algorithms in the first dataset, some variability was also observed in the performance, as a majority of cases were well segmented and a minority displayed non-satisfactory delineation or complete failures, which could be mostly attributed to the rare size/location of the flap and/or the presence of moderate to severe artifacts as illustrated in Figure 2.
The results depicted in Figure 3, present consecutives slices

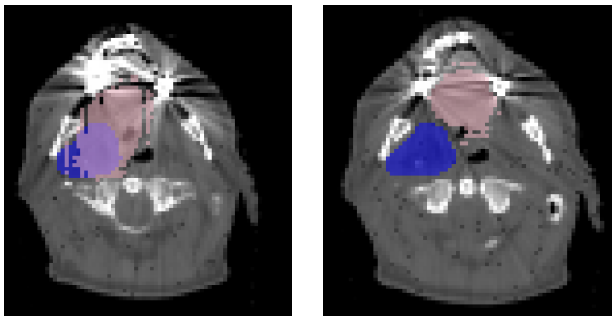


Figure 2: Example of segmentation results on training dataset (two consecutive slices) produced by MedSAM model in the presence of metal artifacts.

from one patient, showcase the segmentation outcomes generated by the nnUNet and MedSAM models.

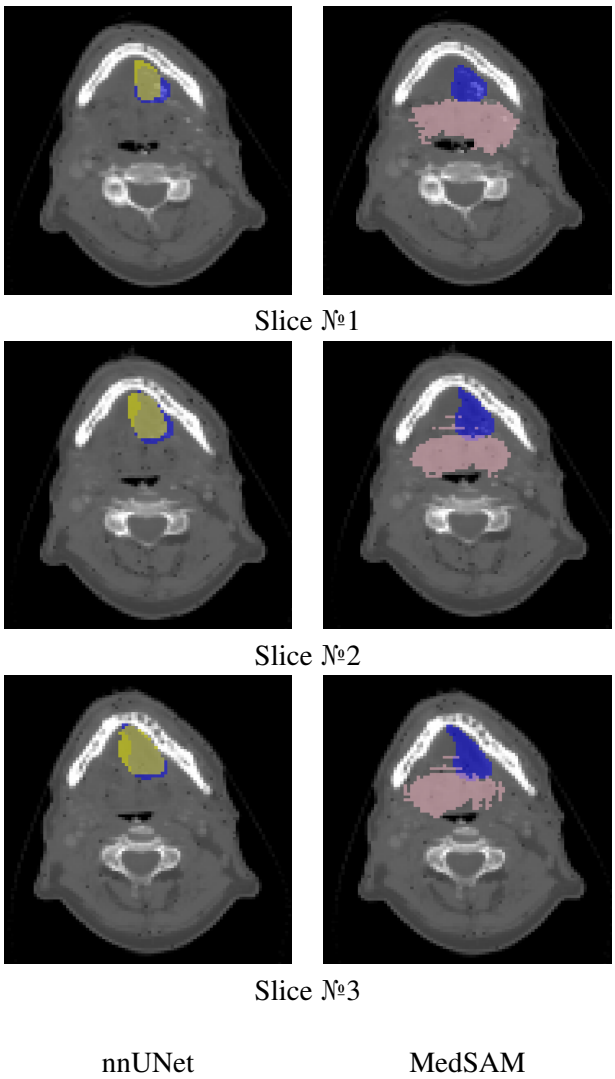
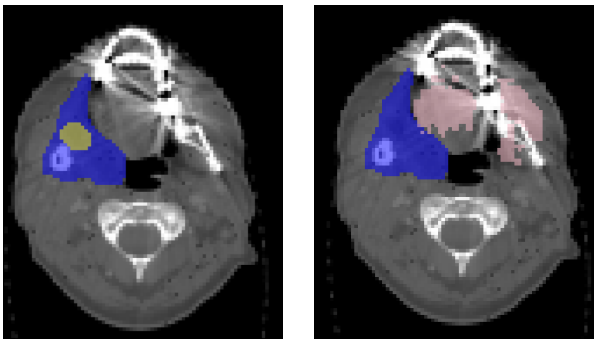


Figure 3: Example of segmentation results produced by nnU-Net & MedSAM models on consecutive slices (Ground truth in blue, nnUNet segmentation in yellow and MedSAM segmentation in pink).

In these representations, the ground truth is denoted in blue, nnU-Net segmentation in yellow, and MedSAM segmentation in pink. The visual evidence highlights the superior

performance of nnU-Net over MedSAM in the context of flap segmentation. The segmentation produced by nnU-Net exhibited a closer alignment with the ground truth compared to MedSAM, indicating a higher degree of concordance between nnU-Net's segmentation output and the actual delineation of structures.
The discerned generalization performance of nnUNet and MedSAM in the context of flap segmentation, particularly when confronted with dental metal artifacts, is visually illustrated by Figure 4.



nnUNet MedSAM

Figure 4: Example of segmentation results on Amiens dataset (slice 17) produced by nnUNet and MedSAM models in the presence of metal artifacts.

4 Discussion

nnU-net is specifically designed for medical image segmentation tasks, utilizing a contracting and expanding path to capture and refine features from medical images. It typically achieves higher accuracy in medical image segmentation tasks compared to SAM. SAM, with its general-purpose nature, can be applied to various tasks, including image segmentation, natural language processing, and code generation. It provides class-specific outputs, allowing it to label each segment with a corresponding class.
Our results confirm the previous findings that nnU-Net strongly outperforms MedSAM for automatic flap segmentation in CT images of postoperative HNC RT patients. However, we also found that the performance of the algorithm substantially decreased in this new "real life" evaluation dataset. However, this trend was not systematic in all the 70 cases, as 48 out of 70 had Dice scores between 0.6 and 0.9. The lower mean and median scores are therefore due to more cases with complete failures (flap not detected or confused with another structure, or due to metal artifacts).

5 Conclusion

Our external validation in a new, "real life" cohort of 70 patients, suggest the previously trained nnU-Net model performs with satisfactory accuracy in a majority of cases. However, there are, as it was the case for the training/validation step in the multicentric prospective study, some patients for which the method is unable to properly detect and/or delineate the flaps, either because the flaps is at a rare location or because of the presence of strong artefacts. Regarding the comparison between nnU-Net and MedSAM, our new evaluation confirms the superiority of nnU-Net, MedSAM being unable to properly account for the 3D nature of the data and to take into consideration the information between adjacent 2D slices processed independently. Our future work will focus on improving the performance on rare locations/characteristics of flaps and dealing with cases that present moderate or severe artifacts.

6 Acknowledgments

With financial support from ITMO Cancer of Aviesan within the framework of the 2021-2030 Cancer Control Strategy, on funds administered by INSERM

References

[1] F. Li, C. Wang, X. Liu, et al. "A composite model of wound segmentation based on traditional methods and deep neural networks". *Computational intelligence and neuroscience* 2018 (2018).

[2] T. Saba. "Recent advancement in cancer detection using machine learning: Systematic survey of decades, comparisons and challenges". *Journal of Infection and Public Health* 13.9 (2020), pp. 1274–1289.

[3] J. Le Guevelou, V. Bastit, P. Marcy, et al. "Flap delineation guidelines in postoperative head and neck radiation therapy for head and neck cancers". *Radiotherapy and Oncology* 151 (2020), pp. 256–265.

[4] O. Ronneberger, P. Fischer, and T. Brox. "U-net: Convolutional networks for biomedical image segmentation". *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer. 2015, pp. 234–241.

[5] F. Isensee, J. Petersen, A. Klein, et al. "nnu-net: Self-adapting framework for u-net-based medical image segmentation". *arXiv preprint arXiv:1809.10486* (2018).

[6] J. Ma, Y. He, F. Li, et al. "Segment anything in medical images". *Nature Communications* 15.1 (2024), p. 654.

[7] A. Kirillov, E. Mintun, N. Ravi, et al. "Segment anything". *arXiv preprint arXiv:2304.02643* (2023).

[8] A. Dosovitskiy, L. Beyer, A. Kolesnikov, et al. "An image is worth 16x16 words: Transformers for image recognition at scale". *arXiv preprint arXiv:2010.11929* (2020).

[9] M. Tancik, P. Srinivasan, B. Mildenhall, et al. "Fourier features let networks learn high frequency functions in low dimensional domains". *Advances in Neural Information Processing Systems* 33 (2020), pp. 7537–7547.

[10] A. Vaswani, N. Shazeer, N. Parmar, et al. "Attention is all you need". *Advances in neural information processing systems* 30 (2017).

[11] F. Isensee, P. F. Jaeger, S. A. Kohl, et al. "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation". *Nature methods* 18.2 (2021), pp. 203–211.

[12] I. Loshchilov and F. Hutter. "Decoupled weight decay regularization". *arXiv preprint arXiv:1711.05101* (2017).

An Automated Framework of Accurate Dominant Intraprostatic Lesion Delineation based on Multimodal Images for MRI-Guided Adaptive Prostate SBRT

Zirong Li[+1], Xiao Liu[+1], Qichao Zhou[*1], and X. Sharon Qi[*2]

[1]Department of Research Algorithms, Manteia Medical Technologies Co., Milwaukee, WI. USA
[2]Department of Radiation Oncology, University of California, Los Angeles, Los Angeles, CA. USA

Abstract Prostate cancer is the most common non-skin cancer for men. Despite of early diagnosis and intervention, the recurrence rate remains high after primary treatment. It was reported that, following external beam radiotherapy (EBRT), local recurrence is mostly dose-dependent and most frequently occur at the site of the dominant intraprostatic lesion (DIL). Dose escalation to the DIL via simultaneous integrated boost (SIB) may improve local tumor control. However, accurate delineation of DIL based on on-board MR images poses a great challenge for MRI guided prostate adaptive therapy. In this work, we propose an algorithm of edge enhancement confidence based on multimodal images, to simultaneously utilize the DIL position information of pre-treatment PET images and the alignment information derived from daily MR images, to accurately identify DIL areas. The effectiveness of the framework was evaluated on prostate cancer patients treated with SBRT on a low field MRI system . The obtained DIL confidence images, account for daily motion on daily MR images improving the accuracy and robustness oftarget delineation and dose escalation for adaptive radiotherapy.

1 Introduction

Prostate cancer(PC) is the most common non-skin cancer for men, representing 7.3% of all new cancers globally[1]. Approximately one in six men will be diagnosed with prostate cancer during their lifetime, and about one in 36 will die of prostate cancer, with an estimated 268,490 cases diagnosed in the United States in 2021[2, 3]. Despite early intervention, the recurrence rate of approximately 20-50% after primary treatment[4]. Following external beam radiotherapy (EBRT), local recurrence is most commonly observed in the dominant intraprostatic lesion (DIL) site[5]. Many clinical studies have demonstrated that dose escalation to the DIL via simultaneous integrated boost (SIB) may improve local tumor control[6].

However, accurate delineation of DIL poses a great challenge in computed tomography(CT). Many studies have used functional imaging to delineate DIL, such as single photon emission computed tomography, positron emission tomography/computed tomography(PET/CT), and Magnetic Resonance Imaging(MRI).PET has high specificity in identifying DILs and benign areas and has high sensitivity in identifying some very small DILs.

In recent years, MR-guided adaptive radiotherapy has developed rapidly, MR guided RT (MRgRT) has superior soft tissue contrast during image guidance or treatment planning. MRI-guided radiation therapy systems, ie, Viewray and Unity[7, 8], allow for MR imaging while patients in treatment position, thereby promoting precise dose delivery within the target area for adaptive radiotherapy. Although daily MR imaging can promote the difference between DIL and benign areas, accurate DIL delineation without PET images presents a major challenge.

Many studies have proposed semi-automatic segmentation algorithms[9]. Compared with automatic segmentation algorithms through neural networks, semi-automatic segmentation can introduce prior knowledge and does not require training. In addition to binary segmentation results, confidence can often be given. Sinop and Grady proposed a common framework based on semi-automatic segmentation algorithms. include graph cuts, and random walk algorithms for interactive segmentation[10]. However, these algorithms are not well adapted to 3D medical images and multi-modal scenes.

To solve the problem of accurately identifying DIL areas during the adaptation radiotherapy, we propose an automated framework based on multimodal reference images, which allows us to simultaneously utilize the DIL position information of offline PET images and the motion information of daily MR images, thereby Accurately identify DIL areas.

2 Materials and Methods

We propose a framework to calculate ROI confidence based on pre-treatment PET and daily MRI. The ROI confidence algorithm is adapted from the algorithm proposed by Liu and Li et al[11]. The original algorithm requires manual designation for a rough trimap, marking extremely few foreground voxels, many background voxels, and a large unknown region. Our adapted algorithm is designed to streamline the adaptive radiotherapy workflow. We implement an automatic confidence algorithm and achieved completely automatically through PET without requiring the manual designation of a rough trimap. Our proposed workflow is illustrated in Figure 1.

[+] The authors contributed equally to this work.
[*] Authors to whom correspondence should be addressed.

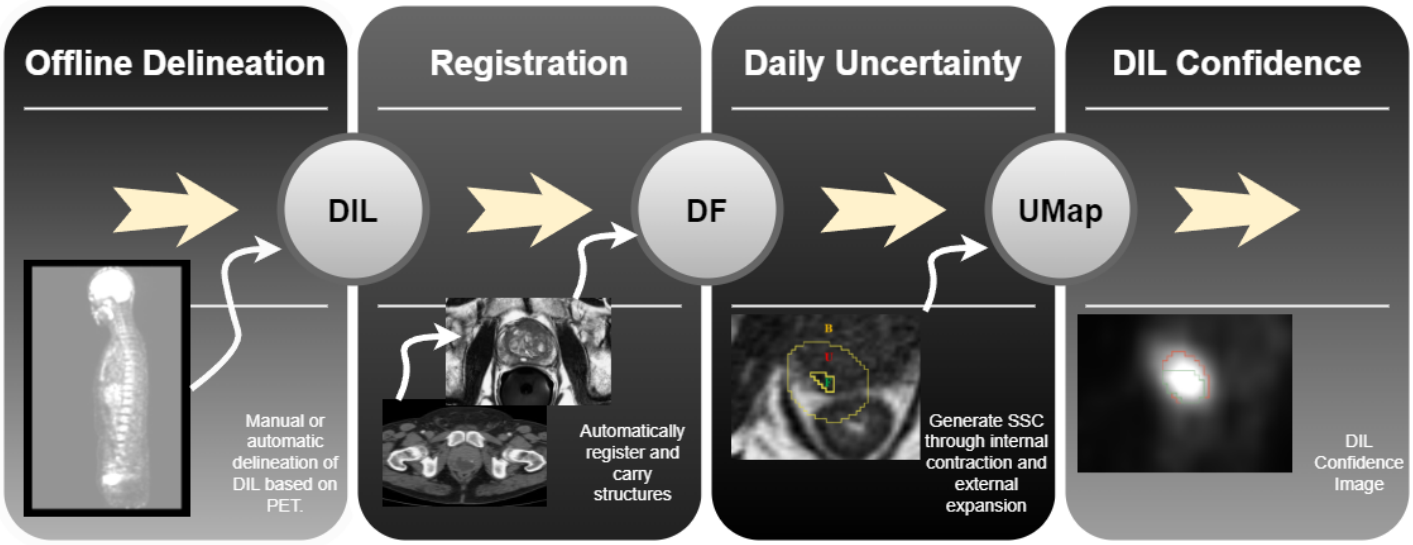


Figure 1: Framework of generating accurate DIL structure confidence based on multimodal images.

Specifically, on the patient's pre-treatment PET/CT images, the prostate structure on the CT is automatically delineated through an automated AI-segmentation tool Manteia AC. The DIL area is automatically segmented through the threshold in the corresponding PET area.

The segmented DIL structure is contracted and expanded by 50% in size in 3-dimension space. The contracted structure is defined as a certain area, and the shell formed by the expanded structure and the contracted structure is defined as an uncertain area. Defined areas in such a way as to cover possible motion in adaptive radiotherapy and the sum of certain and uncertain areas is defined as a structure with solution confidence (SSC). The confidence in the determined area is set to 1, the uncertain area is set to 0.1, and the rest is set to 0.

In the process of adaptive radiotherapy, the MRI of the day is obtained, and the pretreatment CT is deformed to the MRI of the day through deformable registration to obtain the deformation field. The SSC is superimposed on the daily MRI through the deformation field. After registration, the area with a confidence level of 1 in the SSC is positioned as a reference point in the algorithm, and the confidence level of the area with a confidence level of 0 is calculated by calculating the correlation matrix between 3-dimensional multi-modal pixels. The final DIL confidence images were determined based on offline PET and MRI.

We verified the effectiveness of the algorithm on retrospective prostate cancer patients treated on a low-field MR treatment system. Each patient underwent a pre-treatment PET-CT scan and daily MR scans before each treatment fraction.

3 Results

Figure 2(a) illustrates a typical patient of the registration results of CT and MR. Except for the large changes in the skin, the bony structures are completely aligned.

Figure 2(b) shows registration of PET to onboard MRIs. 2(c) DIL confidence image fusion MR. 2(d). DIL confidence image fusion PET.

Calculation speed is very important for adaptive radiotherapy. The calculation of a 3D multi-modal correlation matrix is very time-consuming. However, in the optimized algorithm, the confidence image calculation time of a typical patient is 2.13 seconds.

4 Discussion

We propose a framework that can automatically run and solve DIL confidence images. This framework allows input of pre-treatment PET/CT and DIL structure, as well as daily MRI, to solve the confidence image of DIL structure through a multi-modal 3D correlation matrix. The proposed framework shows promising to accurately delineate the DIL area for prostate adaptive radiotherapy, enabling safer for dose escalation to the DIL via SIB treatment delivery.

The proposed framework is novel in the following aspects. Compared with traditional semi-automatic segmentation algorithms, the proposed framework automates scaling the pre-treatment PET, which doesn't require manual intervention. Compared with deep learning-based algorithms, the framework does not require a large amount of data for training.

5 Conclusion

We propose a framework of utilizing multimodal images, based on pre-treatment PET and onboard treatment MRs, to accurately delineate DIL with confidence. The proposed framework enables the accuracy and reliability of dose escalation to the DIL for online prostate SBRT adaptive radiotherapy.

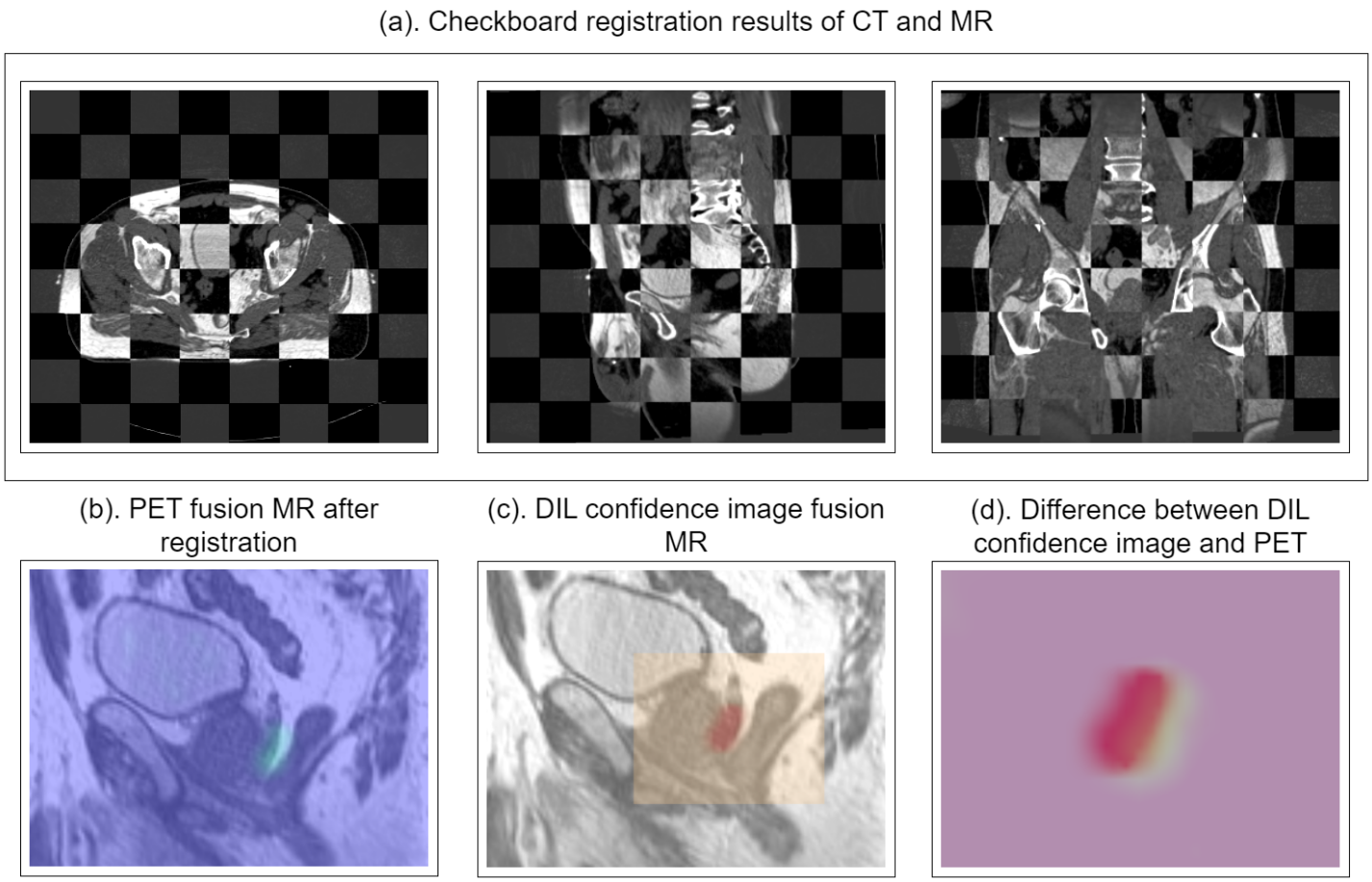


Figure 2: Multimodal calculation of DIL structure confidence results. (a). The registration result is a checkerboard pattern, from left to right, the horizontal plane, the sagittal plane, and the coronal plane. (b). PET fusion MR. The green area is PET. (c). DIL confidence image fusion MR. The red area is the DIL confidence, and the darker the red, the higher the confidence level. (d). DIL confidence image fusion PET, red is DIL confidence and green is PET.

References

[1] H. Sung, J. Ferlay, R. L. Siegel, et al. "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries". *CA: a cancer journal for clinicians* 71.3 (2021), pp. 209–249.

[2] M. W. Dunn and M. W. Kazer. "Prostate cancer overview". *Seminars in oncology nursing*. Vol. 27. 4. Elsevier. 2011, pp. 241–250.

[3] R. L. Siegel, K. D. Miller, H. E. Fuchs, et al. "Cancer statistics, 2021". *Ca Cancer J Clin* 71.1 (2021), pp. 7–33.

[4] N. D. Shore, J. W. Moul, K. J. Pienta, et al. "Biochemical recurrence in patients with prostate cancer after primary definitive therapy: treatment based on risk stratification". *Prostate Cancer and Prostatic Diseases* (2023), pp. 1–10.

[5] E. Arrayeh, A. C. Westphalen, J. Kurhanewicz, et al. "Does local recurrence of prostate cancer after radiation therapy occur at the site of primary tumor? Results of a longitudinal MRI and MRSI study". *International Journal of Radiation Oncology* Biology* Physics* 82.5 (2012), e787–e793.

[6] F. E. von Eyben, T. Kiljunen, A. Kangasmaki, et al. "Radiotherapy boost for the dominant intraprostatic cancer lesion—a systematic review and meta-analysis". *Clinical Genitourinary Cancer* 14.3 (2016), pp. 189–197.

[7] S. Mutic and J. F. Dempsey. "The ViewRay system: magnetic resonance–guided and controlled radiotherapy". *Seminars in radiation oncology*. Vol. 24. 3. Elsevier. 2014, pp. 196–199.

[8] D. Winkel, G. H. Bol, P. S. Kroon, et al. "Adaptive radiotherapy: the Elekta Unity MR-linac concept". *Clinical and translational radiation oncology* 18 (2019), pp. 54–59.

[9] X. Dong, J. Shen, L. Shao, et al. "Sub-Markov random walk for image segmentation". *IEEE Transactions on Image Processing* 25.2 (2015), pp. 516–527.

[10] A. K. Sinop and L. Grady. "A seeded image segmentation framework unifying graph cuts and random walker which yields a new algorithm". *2007 IEEE 11th international conference on computer vision*. IEEE. 2007, pp. 1–8.

[11] X. Liu, Z. Li, X. Qi, et al. "Objective Boundary Generation for Gross Target Volume and Organs at Risk Using 3D Multi-Modal Medical Images". *International Journal of Radiation Oncology, Biology, Physics* 117.2 (2023), e476.

Medical Image Segmentation Assisted with Clinical Inputs via Language Encoder in A Deep Learning Framework

Hengrui Zhao[1], Biling Wang[1], Deepkumar Mistry[1], Jing Wang[1], Michael Dohopolski[1], Daniel Yang[1], Weiguo Lu[1], Steve Jiang[1,2], and Dan Nguyen[1,2]

[1]Medical Artificial Intelligence and Automation (MAIA) Laboratory and Department of Radiation Oncology, , University of Texas Southwestern Medical Center, Dallas, TX 75390, USA
[2]Co-corresponding authors: steve.jiang@utsouthwestern.edu, dan.nguyen@utsouthwestern.edu

Abstract Medical image segmentation plays a crucial role in clinical imaging services, with automatic segmentation techniques being pivotal for enhancing diagnostic accuracy and saving time. Despite their importance, current auto-segmentation methods exhibit suboptimal accuracy, marked by a significant gap between automated results and the ground truth. Traditional methods rely heavily on the image data alone, while clinicians typically utilize additional contextual information beyond the images for accurate contouring of targets and organs. To bridge this gap, our study introduces an innovative approach that incorporates clinical inputs into the auto-segmentation model. We collected clinical data that could potentially alter the segmentation, translated it into meaningful text, and employed a transformer-based text encoder to convert these texts into distinct vectors. These vectors effectively scale the features in our auto-segmentation model, allowing it to incorporate and benefit from multi-modal clinical information. This methodology significantly enhances segmentation quality, as demonstrated in our application of prostate segmentation for radiation therapy, where the delineation of prostate target involves integrating further clinical findings and physician preferences. Our experiments focused on five key factors that influence the segmentation of prostate targets for radiation therapy, with results showing that our proposed method not only outperforms the baseline model lacking clinical input but also surpasses current state-of-the-art methods. This advancement in auto-segmentation technology holds the potential to revolutionize diagnostic processes, offering a more accurate and efficient tool for clinicians.

1 Introduction

Medical image segmentation is a critical task that classifies pixels to identify specific regions within medical images. While clinicians can perform segmentation manually, this process is time-consuming and requires specialized training. Hence, automatic segmentation has been proposed, particularly for CT and MRI images[1]. Although automatic segmentation is already used in various clinical applications, most methods still rely solely on image data.

Recent advancements in natural language processing[2, 3], notably the development of CLIP[4], have been significant. CLIP can embed text and image inputs into a shared vector space, potentially revolutionizing medical image auto-segmentation, which often requires information beyond the images themselves.

To enhance medical image segmentation using additional modalities, we propose a framework that utilizes both image and supplementary information. With variables identified as influential to the segmentation task, our process begins by collecting the data and translating the variables into concise, informative sentences. These sentences are then encoded using a transformer-based text encoder, producing a unique vector embedding for each input. Incorporating this encoded information into a standard image-based segmentation model, we achieve improved segmentation results.

We evaluated our method in a radiation oncology application: prostate segmentation for prostate cancer patients. Five variables were considered for their potential impact on segmentation during radiotherapy[5, 6]: slice thickness, IV contrast, spacer, physician, and MRI. These variables were converted into text, encoded by a pre-trained language model into a vector space, and then normalized and sampled for integration into the segmentation network. The sampled vectors scale features at multiple levels within the neural network, merging the text embedding with image features. This process culminates in decoding the hybrid image-text features to achieve segmentation.

Our method significantly outperforms the baseline neural network without text inputs, addressing the challenge of data scarcity that limits the efficacy of learning-based auto-segmentation. Typically, hundreds to thousands of images are required to train a deep neural network for medical image auto-segmentation. Insufficient data can severely compromise model accuracy. By integrating text data with image data, our method provides richer information, enhancing segmentation accuracy. We also present a controlled text analysis demonstrating how each variable influences segmentation model behavior.

In summary, our contributions are:

- A novel segmentation framework that incorporates medical inputs from multiple modalities into image-based segmentation.
- Validation of our method in prostate radiotherapy, demonstrating superior performance compared to baseline methods.
- Efficacy of our approach in multiple down-sampled datasets, showing optimal performance with limited data availability.

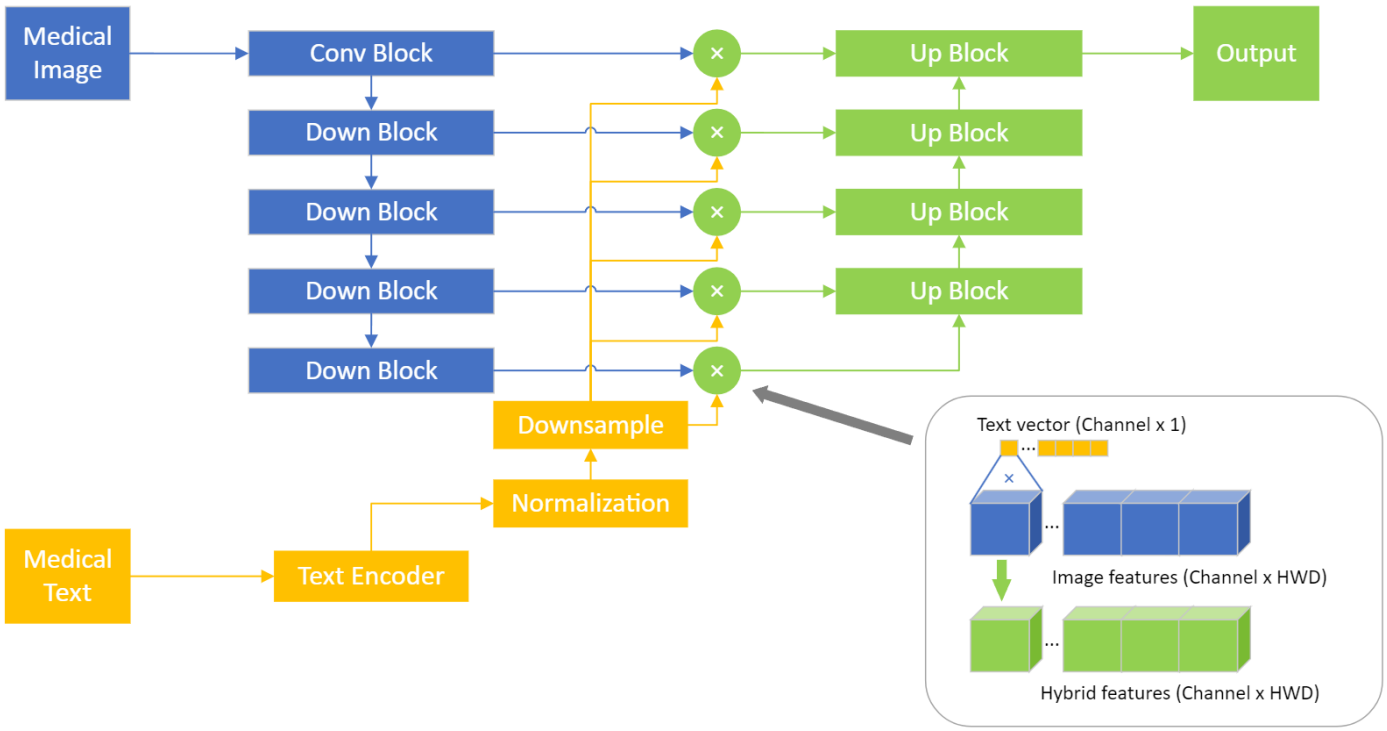


Figure 1: The proposed framework and details.

2 Materials and Methods

Joint framework of image and text input The proposed framework, as illustrated in Figure 1, operates on two primary inputs: a medical image and accompanying medical text. The medical image could be a standard CT or MRI image; in radiation oncology, 3D CT images are typically utilized for treatment planning. These images are processed using a U-Net image encoder, producing encoded image features across five different resolutions.

The second input is medical text. Clinical inputs are often initially in non-textual formats, necessitating translation into text. This process requires a deep understanding of clinical data to ensure that the resultant text is both distinct for different variables and amenable to encoding by a standard text encoder. Once we obtain well-translated text, it is encoded using a text encoder like the CLIP text encoder. The CLIP model, pretrained on a diverse range of internet datasets including medical data, is well-suited for encoding general medical-related texts. It converts each word of a sentence into a token, resulting in a token series that mirrors the sentence's length. This series is then transformed into a 512-value 1-D vector.

To integrate this vector into the deep learning network effectively, we normalize it to a mean of 0 and a standard deviation of 1. This normalization is crucial for stabilizing the training process, as it reduces variability in model accuracy across training sessions. Subsequently, the text vector undergoes down-sampling to align its dimensions with those of the image features. This involves applying stacked down-sampling layers and collecting the output of each layer. Each down-sampled text vector is then utilized to scale an image feature of a corresponding size, as depicted in Figure 1.B. The resultant scaled features, embodying both image and text characteristics, are decoded using a standard U-Net decoder to achieve the final segmentation.

The model is trained using cross-entropy and Dice loss, employing an Adam optimizer with an initial learning rate of 10^{-5}.

Application of prostate segmentation

In this section, we detail the implementation of our proposed framework in a prostate segmentation task. Current automatic segmentation tools for the prostate predominantly rely on image findings, yet their effectiveness has been shown to diminish over time. A recent study revealed a decline in pixel accuracy for newer patient images. This degradation in model performance is attributed to various time-sensitive variables that either alter the image or affect prostate segmentation. Five variables, in particular, are noted for their significant impact on model accuracy:

- **CT Slice Thickness**: Changes in CT slice thickness can alter the perceived prostate volume in pixel terms. A 3 mm slice thickness, for instance, reduces the number of slices containing the prostate.
- **Intravenous Contrast**: This contrast agent, injected venously, enhances body fluid contrast. It particularly increases urinary contrast in the bladder, clarifying the boundaries between the prostate and bladder.
- **MRI Imaging** When available, MRI images of the prostate region are used for more accurate delineation.

Variable	Values	Translation
CT slice thickness	1.5 mm, 2 mm,3 mm	The CT image slice thickness is 1.5/2/3 mm.
MRI availability	Yes, No	The patient has an/no MRI image.
Intravenous contrast	Yes, No	The patient has a normal/higher contrast in bladder.
Spacer hydrogel	Spacer regular, Spacer VUE, No	The patient has a no/an invisible/visible spacer hydrogel.
Physician	A, B, C, D, E, F, G, H, I, J	The prostate is contoured by physician X.

Table 1: The 5 variables of prostate segmentation and their possible values and translation.

Studies suggest that prostate delineation based on MRI is generally smaller than that based solely on CT.

- **Spacer Hydrogel**: This absorbable hydrogel, injected before imaging, comes in two types: Spacer VUE, which is visible in CT scans, and Spacer Regular, which is less visible. Their presence can affect the imaging of the prostate.
- **Physician Contouring**: Physicians contour the prostate using institutional guidelines and personal preferences, leading to delineations that reflect various personal styles.

We incorporated these five variables into the prostate delineation process, guiding our framework's segmentation approach. Each variable was translated into sentences summarizing its effect on prostate segmentation, as outlined in Table 1. These sentences were then inputted into the text encoder of our framework.
For practical application, we collected a clinical dataset of 729 patients from a cancer hospital.

3 Results

3.1 CLIP-UNet has a higher performance than U-Net

We compared our proposed model with a baseline model and a state-of-the-art (SOTA) method for medical image segmentation. The results are presented in Table 2, where five models were tested on the same test set comprising 82 patients. MIS-FM, a foundational model for medical image segmentation, was trained on the same training set. U-Net (16) represents a typical 3D U-Net with 16 initial channels. Given that 3D models are substantially larger than 2D models, the model size can be a limiting factor in some research environments. A 16-channel model provides a balanced trade-off between size, speed, and performance. Similarly, U-Net (32) is a 3D U-Net with 32 initial channels, equivalent to the full-scale model described in the original paper. The last two columns detail our proposed method: CLIP-UNet with 16 and 32 initial channels, corresponding to the U-Net configurations.
CLIP-UNet demonstrated significantly higher accuracy compared to both the 16-channel and 32-channel U-Net models. It also surpassed MIS-FM in terms of accuracy. Particularly,

Model/Performance	Test Dice	Model Size
MIS-FM	84.02%	526 MB
U-Net (16)	83.43%	12.7 MB
U-Net (32)	84.43%	50.6 MB
CLIP-UNet (16)	86.20%	12.7 MB
CLIP-UNet (32)	86.40%	50.6 MB

Table 2: Performance of proposed framework, baseline model and SOTA method

CLIP-UNet with 16 initial channels outperforms U-Net by a large margin of 2.77%. Additionally, the text input for CLIP-UNet can be pre-encoded before testing, as there is a limited variety of texts. Once all possible texts are encoded, the text encoder within the model becomes unnecessary; only the encoded text embeddings are needed. Consequently, the model size of CLIP-UNet matches that of the corresponding U-Nets with the same channel count. Furthermore, the computational cost of CLIP-UNet during inference is also comparable to that of U-Net.

3.2 CLIP-UNet is better than U-Net with fewer data

The proposed segmentation framework utilizes a greater number of medical inputs compared to the baseline U-Net, given the same number of patients. Consequently, CLIP-UNet shows promise for achieving superior performance even with a smaller patient cohort for training. In our experiment, we created multiple sub-datasets from the full dataset, training both CLIP-UNet and U-Net on each. The findings are detailed in Table 3. The complete dataset comprised 729 training patients, with the Dice scores for CLIP-UNet (32) and U-Net (32) being 86.49 and 84.43, respectively. Subsequently, we randomly selected 700 patients from the original 729 to form a training subset, applying both models to this subset. This process was repeated by further sampling down to 600, and progressively to 100 patients, each time training both models anew. The evaluation set of 98 patients and the test set of 82 patients remained constant across all models. The Dice scores for models trained on both the complete dataset and the subsets are presented in Table 3.
As the volume of training data decreases, CLIP-UNet demonstrates remarkable resilience against performance decline. Notably, with as few as 300 training cases, CLIP-UNet's

Training data size	U-Net	CLIP-UNet
100	71.86	74.70
200	81.32	78.69
300	80.91	84.48
400	82.19	85.08
500	82.87	85.90
600	84.99	86.15
700	84.36	86.34
729	84.43	86.40

Table 3: Performance of CLIP-UNet and U-Net with different number of training patients. The result is the tested Dice score in percents.

performance still holds at 84.48%, surpassing that of U-Net trained on the full dataset. This outcome suggests that the paired medical text provides substantial guidance for accurate prostate segmentation.

3.3 Analysis of variables with controlled text guidance

This experiment evaluates the trained CLIP-UNet (32) using the complete training dataset, but with controlled texts during testing phase only. We manipulated each variable to indicate its correctness, for instance, altering the statement from "The patient has an MRI image." to "The patient has no MRI image." To transition each variable from correct to incorrect, we altered them to their subsequent possible value, such as changing 'Physician D' to 'Physician E', or 'invisible spacer' to 'visible spacer'. We analyzed each variable in two scenarios: one where a single variable was incorrect while the others were correct, and another where a single variable was correct while the others were incorrect. These variations were tested with the CLIP-UNet (32) on a test set comprising 82 patients.

We use the CLIP-UNet with accurate text input as a baseline and compares all other test results against it. It is observed that the UNet's performance is 1.97% inferior to that of the CLIP-UNet. Furthermore, when all variables are incorrect, the CLIP-UNet's accuracy decreases by 2.56% compared to when variables are correct. This suggests that the improved accuracy of the CLIP-UNet is heavily reliant on the accuracy of the variables in the texts. Comparing the baseline to scenarios with a single incorrect variable, it appears that Slice Thickness, IV Contrast, and Physician information exert a minor influence on the test results, suggesting these variables might be less critical. However, even with all other variables being incorrect, correct Slice Thickness improved accuracy from -2.56% to -2.18%. Likewise, accurate physician information enhanced accuracy from -2.56% to -2.36% under the same conditions. The impact of IV Contrast on accuracy was minimal, regardless of the correctness of other variables.

4 Discussion & Conclusion

Medical image segmentation is a longstanding and multifaceted topic, with numerous applications and methodologies. Since the rise of Convolutional Neural Networks (CNNs) in the 2010s, they have become the predominant architecture in most automated segmentation models for medical imaging. However, other segmentation methods, such as those based on image registration, continue to be widely used in clinical settings. Despite their effectiveness, both CNNs and image registration methods primarily focus on image input alone. This approach differs from how physicians typically delineate clinical targets, which often involves considering additional clinical information from medical systems or diagnostic findings to achieve accurate segmentation of the desired quality. Consequently, integrating clinical inputs beyond mere imaging into automated segmentation is increasingly favored by clinicians. Leveraging recent advances in machine learning, we have incorporated various clinical inputs into an auto-segmentation model based on CNNs.

Our proposed model demonstrates higher accuracy compared to baseline CNN models, yielding segmentation results that more closely align with the gold standard. Furthermore, the model performs well even with limited data resources, a common limitation for traditional segmentation models. Uniquely, our model tailors its output segmentation based on the provided clinical inputs, allowing clinicians to adjust the results adaptively according to different requirements, such as the preferences of the approving physician. This high level of accuracy and adaptability promises an enhanced auto-segmentation experience for clinicians.

References

[1] O. Ronneberger, P. Fischer, and T. Brox. "U-net: Convolutional networks for biomedical image segmentation". *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer. 2015, pp. 234–241.

[2] A. Vaswani, N. Shazeer, N. Parmar, et al. "Attention is all you need". *Advances in neural information processing systems* 30 (2017).

[3] A. Dosovitskiy, L. Beyer, A. Kolesnikov, et al. "An image is worth 16x16 words: Transformers for image recognition at scale". *arXiv preprint arXiv:2010.11929* (2020).

[4] A. Radford, J. W. Kim, C. Hallacy, et al. "Learning transferable visual models from natural language supervision". *International conference on machine learning*. PMLR. 2021, pp. 8748–8763.

[5] C. V. Dinh, P. Steenbergen, G. Ghobadi, et al. "Magnetic resonance imaging for prostate cancer radiotherapy". *Physica Medica* 32.3 (2016), pp. 446–451.

[6] L. I. Karsh, E. T. Gross, C. M. Pieczonka, et al. "Absorbable hydrogel spacer use in prostate radiotherapy: a comprehensive review of phase 3 clinical trial published data". *Urology* 115 (2018), pp. 39–44.

Influence of Electron Transport Algorithm on Monte Carlo Treatment Planning

G. Mora[1], A. Eldib[2], J Li[2] and Charlie Ma[2]

[1]Instituto de Biofísica e Engenharia Biomédica, Faculdade de Ciências da Universidade de Lisboa, Campo Grande, 1749-016 Lisboa, Portugal
[2]Department of Radiation Oncology, Fox Chase Cancer Center, 333 Cottman Avenue, Philadelphia, PA 19111, USA

Abstract

Purpose: To investigate the effect of the global electron cutoff energy (ECUT) on the calculation accuracy of Monte Carlo (MC) based Treatment Planning. Systematic studies on heterogeneous phantoms are done to determine how ECUT affects the dose uncertainty and consequently the dose volume histogram (DVH) and dose mass histogram (DMH).

Methods: Square phantoms made of different density materials are included in this study. Dose calculations (6MV-IMRT with two beams) are performed using the MCSIM code with the same number of histories and two different ECUT settings (521 keV and 700 keV). Dose DVH and DMH are calculated using MCSHOW code.

Results: The uncertainty on the average dose calculated for low density region (0.001 g/cm^3) is about 2% with ECUT=521 keV and 9% with ECUT=700 keV, respectively. For regions with densities of 0.3 g/cm^3, the uncertainty is about 1% in both 521 keV and 700 keV calculations. We observe differences up to 20% in the Total Target - DVH between 521 keV and 700 keV while the DMH values differed by less than 5% between the two ECUT settings.

Conclusions: DVHs for targets including large (about 40%) low-density regions may be significantly affected by the ECUT value used with the Monte Carlo algorithm. In contrast, DMH is not affected by high ECUT in the treatment plan evaluation, and thus we can use high ECUT and DMH to improve the efficiency and accuracy of Monte Carlo based Treatment Planning.

1 Introduction

Previous studies [1,2,3] have been reported the influence of uncertainties on DVH. The (ECUT) equal to 700 keV is generally accepted as an adequate parameter for MC dose plan calculations. The use of 700 keV versus 521 keV reduces the time of dose calculation, however, it can affect the calculation accuracy of Monte Carlo Treatment Planning. In particular, the accuracy of dose values calculated in low density regions may be very sensitive to the ECUT parameter. Since the DVH histograms are built by the sum of voxel doses of the given volume, the uncertainties of dose values will affect the DVHs, and consequently may affect the evaluation of the plan.

In the present work we aim to study the effect of ECUT on the calculation accuracy of Monte Carlo Treatment Planning. Systematic studies on heterogeneous phantoms were done to determine quantitatively, how ECUT affects the dose uncertainty [4,5] and consequently the DVH and DMH.

2 Material and Methods

Monte Carlo dose calculations are performed in tissue-like material phantoms to investigate the combining effect of density and electron transport parameter (ECUT) on DVH and DMH.

Four square phantoms of different densities are designed, and dose calculations are performed for each phantom using MCSIM code [4,5] and two different ECUTS, 521 keV and 700 keV. Each phantom (*see Figure 1*) with dimensions of 10x10x10 cm^3 includes one target and the target itself has two components (T1 and T2) with equal dimensions of 4x4x4 cm^3. T2 is made of the same tissue-like material with density (1 g/cm^3) in the four different phantoms while T1 is made of air material but different density (0.001 g/cm^3, 0.1 g/cm^3,0.2 g/cm^3 and 0.3 g/cm^3) in each phantom.

6MV-IMRT dose calculations are performed for each phantom using the same number of histories (ten million) and the same simulation parameters except ECUT.

MCSHOW [6] was used to display isodose distributions on top of the phantom and plot the DVHs for T1, T2 and total target (T1+T2). MCSHOW includes a new feature which allows to plot DMH for T1, T2 and total target (T1+T2).



		Phantom 1 Density 0.001	Phantom 2 Density 0.1	Phantom 3 Density 0.2	Phantom 4 Density 0.3
	Volume (cc)	Mass (gm)	Mass (gm)	Mass (gm)	Mass (gm)
Target1	256	0.3	25.6	51.2	76.8
Target2	256	256	256	256	256
Tot Target	512	256.3	281.6	307.2	332.8

Figure 1: Square phantom subdivided into T1 and T2

3 Results

a)

ρ g/cm³			700 keV Uncertainty on dose				
T1	T2	T1+T2	Max T1	Max T2	Av T 1	Av T 2	Av T1+T2
0.001	1	0.50	76.45	1.08	9.20	1.12	5.16
0.1	1	0.55	5.8	1.45	1.46	1.12	1.29
0.2	1	0.60	3.77	1.07	1.25	1.12	1.18
0.3	1	0.65	2.86	1.08	1.17	1.12	1.14

Table 1a, b: Uncertainties on maximum (Max) and average (Av) dose values calculated for T1 and T2 of four different density phantoms a) ECUT= 700keV and b) ECUT=521keV

b)

ρ g/cm³			521 keV Uncertainty on dose				
T1	T2	T1+T2	Max T1	Max T2	Av T 1	Av T 2	Av T1+T2
0.001	1	0.50	21.06	1.48	1.9	1.1	1.5
0.1	1	0.55	3.75	1.39	1.2	1.0	1.1
0.2	1	0.60	1.57	1.04	1.1	1.0	1.1
0.3	1	0.65	1.69	1.05	1.0	1.0	1.0

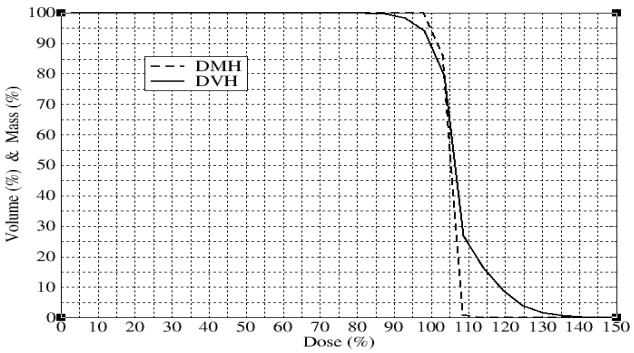


Figure 2: DVH and DMH calculated for the air-like density phantom and ECUT=700 keV

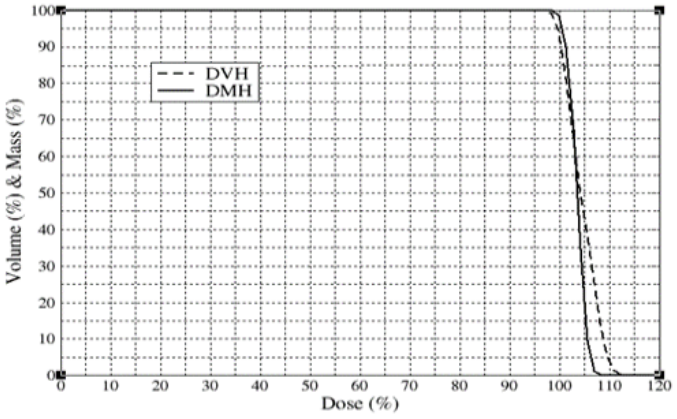


Figure 3: DVH and DMH calculated for the air-like density phantom and ECUT=521 keV

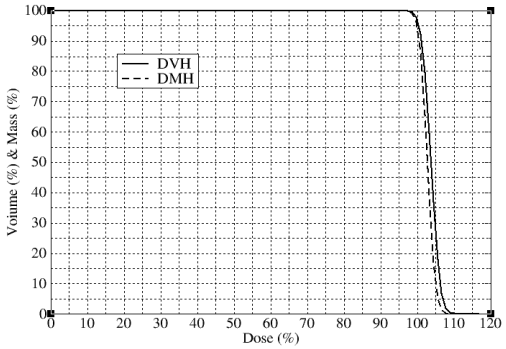
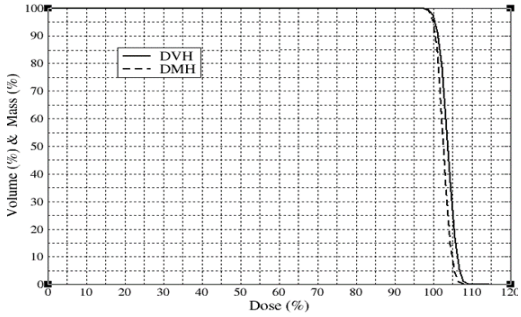


Figure 4a,b: DVH and DMH calculated for the phantom with density of 0.1 g/cm^3 a) upper: ECUT=700 keV and b) lower: ECUT=521 keV



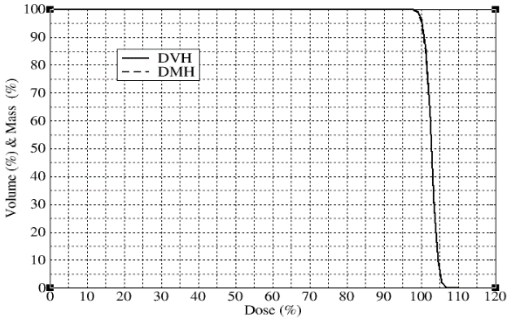
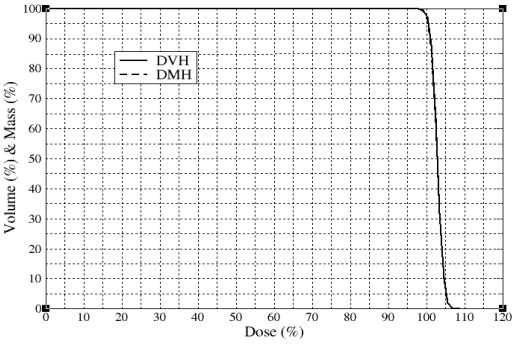


Figure 5a,b: DVH and DMH calculated for the phantom with density of 0.3 g/cm^3 a)upper: ECUT=700 keV b) lower: ECUT=521 keV



4 Discussion

In above Tables, we see that the uncertainty on the average dose values calculated for T1 (0.001 g/cm^3) is about 2% for calculation with ECUT=521 keV and 9% for that with ECUT=700 keV. Increasing the density of T1 to 0.3 g/cm^3 decreases the uncertainty to about 1% in both 521 keV and 700 keV calculations. Table 1a) shows that the uncertainty on the values of $D_{average}$ obtained by 700 keV calculation for the Total Target of air-like density phantom is about 5% while those values for 521 keV calculation (see Table 1b) have an uncertainty lower than 2%. The dramatic effect of the density on the uncertainty of the maximum dose is shown in both tables. And this effect is strongly accentuated by ECUT. Varying ECUT from 521 to 700 keV increases the uncertainty on maximum dose (air tissue phantom) by about 55%.

Figure 2 shows the comparison of DVH and DMH calculated for the air-like density phantom and ECUT=700 keV. The value of DVH-maximum dose is about 10% lower than prescribed dose while the minimum dose is about 20% greater than prescribed dose. In contrast, the DMH plot exhibits good shape.

The DVH and DMH plots determined in 521 keV calculation are presented in Fig 3. In the upper region of the plot, we do not observe differences between DVH and DMH. However, in the lower region (for less than 50% of the volume) differences up to 5% are observed between the two plots. Our results indicate that DVH may not be the good solution for the low-density regions even when the calculation is done with ECUT equal to 521 keV.

In Figures 4 a) and b), we see that differences between DVH and DMH calculated for the phantom with density of 0.1 g/cm^3 are not significant for both cases, 700 keV and 521 keV calculation. Figure 5 a) and b) show that DVH and DMH plots match perfectly.

5 Conclusion

Our results indicate that the ECUT value does not affect the DVH plotted from Monte Carlo dose distributions calculated in regions with density equal or greater than 0.3 g/cm^3.
DVHs for targets including large (about 40%) low-density regions may be significantly affected by the ECUT value used with the Monte Carlo algorithm. In contrast, DMH is not affected by high ECUT in the treatment plan evaluation, and thus we can use high ECUT and DMH to improve the efficiency and accuracy of Monte Carlo based Treatment Planning. Based on the results of the present work, real clinical cases were selected and included in the ongoing studies.

References

[1] J Sempau, A F Bielajew. Towards the elimination of Monte Carlo statistical fluctuation from dose volume histograms for radiotherapy treatment planning, Phys Med Biol, 2000. DOI:10.1088/0031-9155/45/1/310

[2] C M Ma, J S Li, S B Jiang, T Pawlicki1, et al "Effect of statistical uncertainties on Monte Carlo treatment planning", Phys Med Biol, 2005 DOI 10.1088/0031-9155/50/5/013

[3] G. Mora, A. Eldib, J S li, C M Ma, "Investigating the accuracy of Monte Carlo based treatment planning of head and neck patients", WORLD CONGRESS ON MEDICAL PHYSICS AND BIOMEDICAL ENGINEERING. JUNE 3-8, 2018, Prague, Czech Republic

[4] C M Ma , J S Li, T Pawlicki et al "A Monte Carlo dose calculation tool for radiotherapy treatment planning", Phys Med Biol, 2002, pp:1671-89. DOI: 10.1088/0031-9155/47/10/305.

[5] C M Ma, E Mok, A Kapur, T Pawlicki et al "Clinical implementation of a Monte Carlo treatment planning system", Med Phys, 1999. DOI: 10.1118/1.598729

[6] I. Kawrakow, MCSHOW-A visualization Tool for Monte Carlo dose distributions

This work is supported by "Fundação para a Ciência e a Tecnologia "FCT under Strategic Programme UIDB/00645/2020 and UIDP/00645/2020.

GMC [gímık] then and now: (r)evolution of a one-variable proton Monte Carlo dose algorithm.

N. Depauw[1], M. Williams[1], B. Clasie[1], K. Jee[1], T. Madden[1], and H. Kooy[1]
[1]Francis H. Burr Proton Therapy Center, Department of Radiation Oncology,
Massachusetts General Hospital (MGH), Boston MA 02114, USA

Abstract GMC [gímık] is a one-variable Monte Carlo dose algorithm for proton therapy. Its accuracy and performance were presented at the ICCR in Melbourne in 2013. Here, we present the current state of GMC after porting to a GPU architecture. Elevated from the original version, we demonstrate the impressive speed gain while conserving the system's accuracy. Features now include dose-averaged LET and BioDose computations, support for PBS and SOBP fields, as well as RBE modeling based on the McNamara model with arbitrary alpha-beta ratios. The DICOM data processing and loading steps for transparent integration in the clinical workflow take the major part of the total running time, while the computation runs in order of seconds, regardless of patient size, number of beams, prescription, etc. A live demo may be performed.

1 Introduction

GMC [gímık] was first introduced at the XVII International Conference on the use of Computers in Radiation therapy in Melbourne in 2013 [1]. At the time, this one-variable Monte Carlo dose algorithm was implemented on a CPU and shown to provide fast and accurate dose calculation for proton therapy beams, with the promise of exponential speed gain through porting on the GPU architecture [2].
The code has been ported to a GPU implementation and now includes dose-averaged LET_d, BioDose [3], and RBE based on the MacNamara model [4] with default alpha-beta ratios of 2 and 10 – one can also specify any other arbitrary alpha-beta ratio value. Here, we report on the accuracy and performance of this updated GMC version.

2 Materials and Methods

GMC is designed de-novo to reproduce clinical proton physics observables, including dose (to water) and LET_d, and to reduce the computational and physics models to their minimum.
A set of simple plans was generated directly in our treatment planning system (TPS), Astroid [4] which is based on a pencil beam algorithm for dose calculations [5]. These were intended to test the accuracy of the physics observables at the macro level. Clinical patient plans were then recomputed given various parameters (a subset is provided in Table 1) for verification purposes. The timing of the various operations (data loading, data preparation, computations, data presentation, and saving) was aggregated for performance evaluation.
Computations were performed on an individual Dell Precision 7820 workstation with an Intel Xeon P family I7 processor and an NVIDIA RTX A5000 GPU card (Ampere architecture-based with 24GB GDDR6 memory).

3 Results

Physical expectations and GMC calculations show excellent agreement at the macro level, i.e. single energy depth dose and scattering power [6]. This is then demonstrated by recomputing a 1-liter 2-GyRBE reference dose sphere, which is provided in Figure 1.

In Figure 2, we further illustrate the accuracy of the BioDose calculation – which depends on dose and LET_d – based on the same reference sphere simulation.

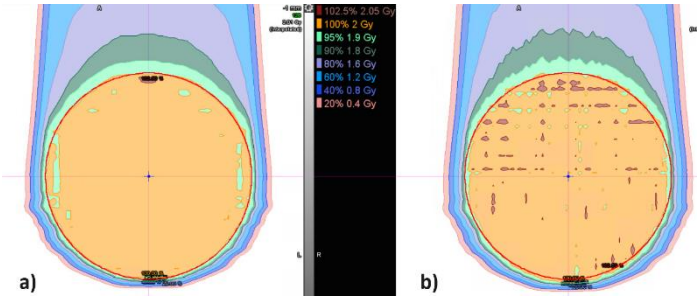


Figure 1: Dose distribution for a 1-liter 2-GyRBE reference dose sphere a) generated in Astroid and b) recomputed in GMC.

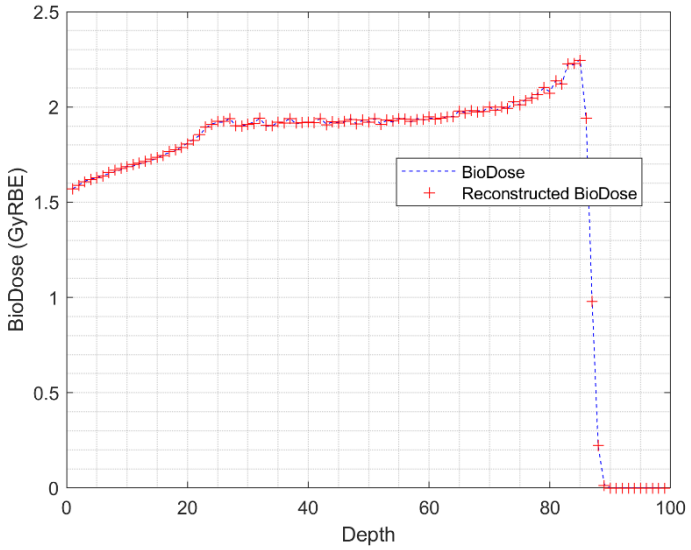


Figure 2: The GMC output BioDose cube is compared to the reconstruction from the base quantities: (Dose/1.1) × (1 + LET × 0.03).

Overall, excellent agreement is observed in terms of both dose distribution and dose-volume histograms for the clinical patient plans. An arbitrarily chosen patient case is depicted in Figure 3, comparing Astroid and GMC. In this CSI case, a slight underdose of the target is observed around the brain due to the improved scattering modelization with the Monte Carlo engine.

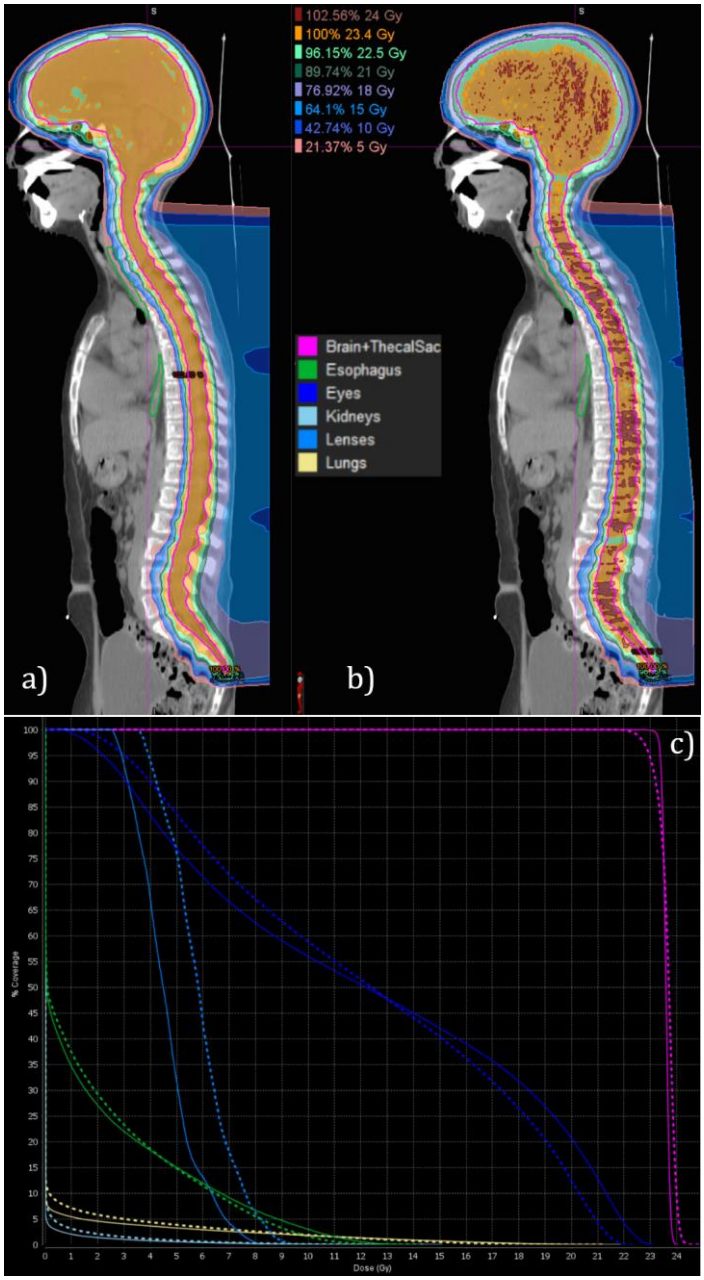


Figure 3: Dose distributions for a CSI plan a) generated in Astroid and b) recomputed in GMC. The dose-volume histograms are provided for both computation in panel c – Astroid (solid lines) vs GMC (dash lines). This plan is based on a common 1.8 GyRBE dose per fraction, it uses five large fields with apertures and thick range shifters. It consists of a total of 615×1e^9 protons.

Table 1 provides the total GMC running time, which includes all the steps from data retrieval to data saving, for a subset of six representative patient cases. Computation-only (proton transport and post-processing) timing is also provided. The computations here included the auxiliary quantities (LET_d, BioDose, and RBE).

Table 1: GMC simulation run times for a subset of Astroid clinical patient plans representative of the planning parameters used in the clinic. All computations (and total) times include the auxiliary quantities (LETd, BioDose, and $RBE_{a/b=2 \text{ or } 10}$).

Treatment site	Planning parameters	Number of protons (Gp)	Total time (s)	Computation time (s)
Breast	1 Field, large size, 8 cm RS, No AP, highly hypo-fractionated	838	7.2	3.5
Brain	4 fields, small size, 3 cm RS, No AP	31	6.9	2.3
Clival	4 fields, small size, 8 cm RS, AP	77	6.8	3.1
CSI	5 fields, large size, 8 cm RS, AP	615	16.5	8.5
Prostate	4 fields, small size, no RS, AP	133	7.7	2.6
Shoulder	2 fields, medium size, 8 cm RS, AP, Metal rod	286	7.1	3.4

RS = range shifter, AP = apertures, Gp = 1e^9 protons, “large” field size refers to an open snout > 20 cm, and “small” fields are < 10 cm.

4 Discussion

In this work, excellent dosimetric accuracy is observed between the TPS and GMC. Here, we do not compare it against a general Monte Carlo engine. These are time-consuming and may take days to complete, especially as the number of treatment fields and the field size increase [6]. The CPU version of GMC was benchmarked against GEANT4/TOPAS in the past [1-2]. Similarly, many studies at our institution involved our various proton TPS (which all used the same pencil beam algorithm) and comparisons based on TOPAS [8-10]. Future TOPAS comparisons against GMC will be generated.

Future work involves the automation of plan recalculations as new proton DICOM plans are detected in our Picture Archiving and Communication System (PACS). This will be performed using a dedicated server with four NVIDIA RTX A5000 paired through NVLink Bridge for improved speed. All the dose computations in this work were based on an individual workstation; recomputation speeds and parallel plan processing are expected to improve further using the dedicated server.

In this work, dose computational times were provided inclusive of auxiliary quantities; the rational for this is the growing interest in using these variables in the general planning paradigm. However, it would be faster and less resource-intensive to perform the dose calculation only. This could be of interest for online adaptive planning.

5 Conclusion

GMC has been ported to a GPU architecture. It retains the original accuracy while adding novel features and exhibiting an incredible speed gain. Future work will involve improving integration for online adaptive therapy systems and pursuing open licensing to distribute the code for general use by the proton community.

References

[1] Depauw, N., B. Clasie, T. Madden, A. Rosenfeld, and H. Kooy. "GMC ['gimik]: a one-variable Monte Carlo dose algorithm for proton therapy." In Journal of Physics: Conference Series, vol. 489, no. 1, p. 012010. IOP Publishing, 2014. DOI: 10.1088/1742-6596/489/1/012010.

[2] Depauw, Nicolas. "A path towards adaptive proton pencil beam scanning therapy." PhD diss., School of Physics, Faculty of Engineering and Information Sciences, University of Wollongong, 2014.

[3] Unkelbach, J., Botas, P., Giantsoudi, D., Gorissen, B.L. and Paganetti, H., 2016. Reoptimization of intensity modulated proton therapy plans based on linear energy transfer. International Journal of Radiation Oncology* Biology* Physics, 96(5), pp.1097-1106.

[4] McNamara, A.L., Schuemann, J. and Paganetti, H., 2015. A phenomenological relative biological effectiveness (RBE) model for proton therapy based on all published in vitro cell survival data. Physics in Medicine & Biology, 60(21), p.8399.

[5] Kooy, Hanne M., Benjamin M. Clasie, Hsiao-Ming Lu, Thomas M. Madden, Hassan Bentefour, Nicolas Depauw, Judy A. Adams et al. "A case study in proton pencil-beam scanning delivery." International Journal of Radiation Oncology* Biology* Physics 76, no. 2 (2010): 624-630.

[6] Hong, Linda, Michael Goitein, Marta Bucciolini, Robert Comiskey, Bernard Gottschalk, Skip Rosenthal, Chris Serago, and Marcia Urie. "A pencil beam algorithm for proton dose calculations." Physics in Medicine & Biology 41, no. 8 (1996): 1305.

[7] Gottschalk, Bernard. "On the scattering power of radiotherapy protons." Medical physics 37, no. 1 (2010): 352-367.

[8] Faddegon, B., Ramos-Méndez, J., Schuemann, J., McNamara, A., Shin, J., Perl, J. and Paganetti, H., 2020. The TOPAS tool for particle simulation, a Monte Carlo simulation tool for physics, biology and clinical research. Physica Medica, 72, pp.114-121.

[9] Testa, M., Schümann, J., Lu, H.M., Shin, J., Faddegon, B., Perl, J. and Paganetti, H., 2013. Experimental validation of the TOPAS Monte Carlo system for passive scattering proton therapy. Medical physics, 40(12), p.121719.

[10] Perl, J., Shin, J., Schümann, J., Faddegon, B. and Paganetti, H., 2012. TOPAS: an innovative proton Monte Carlo platform for research and clinical applications. Medical physics, 39(11), pp.6818-6837.

[11] Depauw, N., Kooy, H.M., Daartz, J., Bussière, M., Batin, E., Madden, T., Williams, M., Schueman, J. and Clasie, B.M., 2022. Implementation of apertures in a proton pencil-beam dose algorithm. Biomedical Physics & Engineering Express, 8(2), p.025024.

Assessing neutron-induced carcinogenesis risk in high-energy radiation therapy: a Monte Carlo and track structure study using realistic neutron spectra and DNA damage clusters

Sachin Dev[1], Felix Mathew[1], and John Kildea[1]

[1]Department of Oncology, McGill University Health Centre, McGill University, Canada

Abstract High-energy photon radiation therapy (> 8 MeV) can lead to unintended patient exposure to secondary photoneutrons, posing a risk for iatrogenic cancers. Existing studies to evaluate the potential of photoneutrons to induce carcinogenesis have been performed solely with flat neutron spectra irradiating the ICRU-4 sphere. Therefore, to replicate a real-world cell irradiation scenario, our study is modeling an in-vitro cell irradiation geometry irradiated with realistic photoneutron spectra encountered in high-energy external-beam photon radiotherapy and aims to evaluate the resulting neutron-beam RBE. As a first step, we have quantified the secondary particle spectra in the ICRU-4 sphere for monoenergetic neutron beams of initial energies 1 keV, 1 MeV, and 10 MeV. An in-house nuclear DNA model and an algorithm for recording clustered DNA damage will be used to calculate DNA damage yields. The neutron RBE obtained from the simulation will be validated by performing in-vitro irradiations of B-lymphoblastoid cells with a 6 MV photon beam used as reference radiation and with the photoneutron spectra arising from a 15 MV linac.

1 Introduction

The use of ionizing radiation offers great benefits to society. However, radiation may be detrimental to human health. The use of radiation thus entails an associated non-zero risk. The unwanted photoneutron exposure to patients during high-energy (> 8 MeV) radiation therapy is one such example of a risk associated with the use of radiation [1, 2]. This unavoidable exposure of secondary neutrons can potentially cause patients to develop iatrogenic secondary cancers later in life [3]. The risk for inducing stochastic effects associated with neutrons is energy dependent and relatively higher in magnitude as compared to other types of radiation. The energy dependence of neutron-induced stochastic effects is typically quantified by neutron radiation weighting factors (w_R) provided by International Commission on Radiological Protection (ICRP 2003, 2007) [4, 5] and the neutron quality factors (Q) published by the United States Regulatory Commission (US NRC 2021) [6]. Although the factors w_R and Q follow similar energy dependence trends qualitatively, their magnitudes are highly discrepant. The wide variety of neutron RBE data used to derive the weighting and quality factor potentially explains this discrepancy. This study is an expansion of our Neutron-Induced Carcinogenic Effects (NICE) research group's efforts to trace the biophysical origin of neutron RBE [2, 7, 8, 9, 10]. Recent studies [11, 12] have used Monte Carlo simulation frameworks to investigate neutron RBE using neutron-induced DNA damage lesions with flat neutron spectra composed of multiple mono-energetic beams. However, we aim to use realistic neutron spectra measured previously by our group [2]. In the future, we aim to validate the RBE values obtained from Monte Carlo simulations with in vitro experiments. In this work, we describe how we obtained secondary species spectra in the ICRU-4 sphere when irradiated with monoenergetic neutron beams of initial energies 1 keV, 1 MeV, and 10 MeV. Simulations were repeated with different versions of the TOPAS Monte Carlo framework [13] and the obtained spectra were validated against the corresponding versions of Geant4 [14].

2 Materials and Methods

The four-component soft tissue-equivalent sphere recommended by the International Commission on Radiation Units and Measurements (ICRU) for use in radiation safety and dosimetry applications was modeled in TOPAS v3.6. The ICRU-4 sphere is 30 cm in diameter and made up of hydrogen, carbon, nitrogen, and oxygen. The ICRU-4 sphere approximates a human adult torso. Three scoring volumes of radius 1.5 cm each and made up of the same ICRU-4 material were centered along a single axis at radial distances of 0 cm, 7.5 cm, and 13.5 cm from the sphere's center, consistent with the previous study by our group [8] performed in Geant4. These scoring volumes have been referred to as inner, intermediate, and outer scoring volumes, respectively. The ICRU-4 sphere was irradiated with monoenergetic neutrons with initial kinetic energy of 1 keV, 1 MeV, and 10 MeV originating from a source positioned on the surface of the ICRU sphere.

The Lawrence Livermore data libraries, which include information for photons [15], electrons [16], and atomic relaxation [17], were employed to manage electromagnetic processes. Default models were integrated to address reactions of lesser dosimetric significance, such as free neutron decay (G4DecayPhysics) and the decay of radioactive isotopes (G4RadioactiveDecayPhysics). Hadronic elastic scattering and inelastic processes (fission, scattering, neutron capture) were described using Geant4 ParticleHP models. An extension was developed in TOPAS to correctly simulate the transport of thermal neutrons below 4 eV. An additional G4NeutronHPThermalScattering model was incorporated in the TOPAS extension which simulates thermal neutron (0 eV -> 4 eV) transportation. To make the G4NeutronHPThermalScattering data set work, the elemental hydrogen in the ICRU sphere was replaced by the

element Hydrogen in Water (TS_H_of_Water) by redefining it using elemental composition and the natural abundance of its isotopes. All the simulations were performed on Digital Research Alliance of Canada clusters for 10^9 histories. The spectra of each secondary species (protons, deuterons, alphas, carbon nuclei, nitrogen nuclei, oxygen nuclei, and gammas) produced were recorded for each monoenergetic neutron energy. The TOPAS v3.6 generated secondary species spectra were compared with Geant4 v10.06p03. All the plots were generated in the ROOT [18] data analysis framework.

After obtaining the secondary species spectra in the ICRU-4 sphere with monoenergetic neutron irradiation and matching the spectra with those in the Lund et al study [8], the next step is to simulate in vitro experimental geometry in TOPAS. Our group has measured photoneutron spectra arising from a 15 MV linac at different locations in the radiotherapy bunker [2]. The realistic photoneutron spectra will be used to irradiate in vitro geometry computationally and obtain secondary species spectra.

Our lab has built a geometric nuclear DNA model of a single cell using the TOPAS-nBio [19] framework and it is available under an open-source license [9]. The model consists of nucleotide base pairs created using six spheres to represent two nitrogenous bases, two deoxyribose molecules, and two phosphate groups. The nucleus has been constructed as a cubic grid to make the model computationally efficient. The nucleus-sized volume has 6.3 Gbp of DNA at a density of 13.3 Mbp/um^3 making it consistent with human nuclei. Our single-cell model will be stochastically exposed to secondary particle spectra at different dose levels: 0.0 Gy, 0.5 Gy, 1.5 Gy, and 3.0 Gy. The IAEA phase space files [20] will be used in TOPAS to simulate 6 MV linac beams and generate the transport of secondary particles and thus irradiate our single-cell model to different dose levels mentioned above.

The photoneutron RBE obtained from Monte Carlo simulations will be validated against an RBE value estimated experimentally. We are currently exposing human B-lymphoblastoid cells to three doses of 0.5 Gy, 1.5 Gy, and 3.0 Gy to 6 MV photons in vitro along with sham irradiation to 0.0 Gy. A linac is used to deliver the doses at a dose rate of 600 MU/min. We aim to perform in vitro cell irradiations with photoneutrons originating from 15 MV linac.

3 Results

The initial spectra of secondary gammas and charged particles were collected in all three scoring volumes of the ICRU-4 sphere for 1 keV, 1 MeV, and 10 MeV neutron sources. Figure 1 shows the secondary charged particle spectra produced in the inner scoring volume with a 1 MeV neutron source. In Figure 2, we present a comparison of electron spectra produced within the inner scorer by a 1 MeV neutron source. In this analysis, we used the "Pre-Step point" object of the Geant4 framework that allowed us to retrieve information about the particle, such as its kinetic energy, position, and direction, at the moment before it moves to the next point in the trajectory.

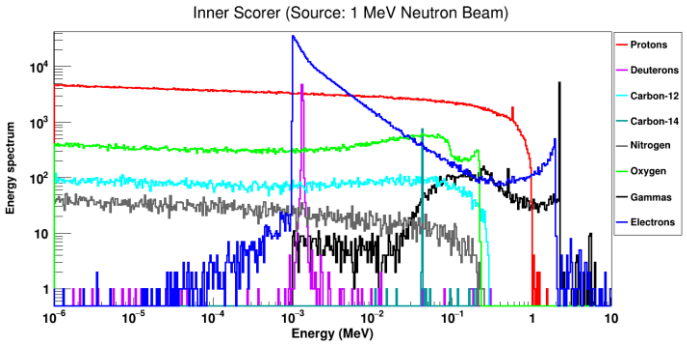


Figure 1: Secondary charged particle spectra in inner scoring volume obtained by 1 MeV neutron source in Topas v3.6

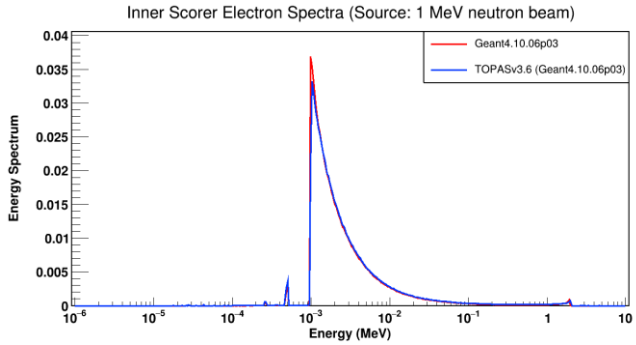


Figure 2. The electron spectra within the inner scoring volume, generated by a 1 MeV neutron source, are compared between Topas v3.6 and Geant4 v10.06.p03.

4 Discussion

For all incident monoenergetic neutron beams of 1 keV, 1 MeV, and 10 MeV, a strong agreement was observed in the secondary particle spectra produced in all three scoring volumes of the ICRU-4 sphere with Geant4 v10.06.p03. In this study, TOPAS v3.6 was used for simulations, which is a wrapper of Geant4 v10.06.p03. However, Lund et al [8] performed simulations in a different version of Geant4: v10.04.p02, yet all spectra obtained in TOPAS v3.6 matched well with Lund et al study [8] except for electrons and secondary gamma due to improved cross-sections for electromagnetic processes in Geant4 v10.06.p03. In addition, the simulations were also repeated with TOPAS v3.9 and corresponding Geant4.10.07p03, and a good agreement between the two was observed. The thermal neutron transportation physics lists were not prebuilt in TOPAS so an extension was developed using G4NeutronHPThermalScattering dataset to correctly simulate thermal neutron transportations. An important limitation of our single-cell model is that it considers the direct and indirect DNA damage induced by radiation only but does not apply cell repair models. Our research group is

currently developing an update to our custom DNA model by incorporating cell repair kinematics. We aim to expand this study by analyzing the impact of repair models on resulting neutron RBE.

5 Conclusion

We have generated secondary particle species produced by monoenergetic neutrons of initial energies 1 keV, 1 MeV, and 10 MeV by simulating the ICRU-4 sphere in TOPAS. We have developed an extension in TOPAS for correct simulation of thermal neutron transportation for the first time thus contributing to advance the capabilities of the TOPAS simulation application. Repeated simulations with different versions of TOPAS code and matching the results with corresponding Geant4 versions demonstrated the robustness of both our methodology and results. In the future, we aim to calculate realistic photoneutron RBE using our simulation application and validate it experimentally.

Acknowledgments

This research was supported by the Natural Sciences and Engineering Research Council (NSERC) Discovery Grant by Prof. John Kildea and by the NSERC CREATE grant in Responsible Health and Healthcare Data Sciences (SDRDS).

References

[1] Howell R M, Hertel N E, Wang Z, Hutchinson J and Fullerton G D 2006 Calculation of effective dose from measurements of secondary neutron spectra and scattered photon dose from dynamic MLC IMRT for 6 MV, 15 MV, and 18 MV beam energies Medical Physics 33(2) 360-368. DOI: 10.1118/1.2140119

[2] Maglieri R, Licea A, Evans M, Seuntjens J and Kildea J 2015 Measuring neutron spectra in radiotherapy using the nested neutron spectrometer Medical Physics 42(11) 6162-6169. DOI: 10.1118/1.4931963

[3] Friedman D L et al 2010 Subsequent neoplasms in 5-year survivors of childhood cancer: the Childhood Cancer Survivor Study JNCI: Journal of the National Cancer Institute 102(14) 1083–1095. DOI: 10.1093/jnci/djq238

[4] ICRP 2003 ICRP 92: Relative Biological Effectiveness (RBE), Quality Factor (Q), and Radiation Weighting Factor (wR) Annals of the ICRP 33(4)

[5] ICRP 2007 ICRP 103: The 2007 Recommendations of the ICRP Annals of the ICRP 37(2-4)

[6] US NRC 2021 Units of Radiation Dose Code of Federal Regulations 10(20.1004) 298-300

[7] Kildea J 2017 The Canadian Neutron-Induced Carcinogenic Effects Research Program: A research program to investigate neutron relative biological effectiveness for carcinogenesis with a particular focus on secondary (by-product) neutrons in high-energy radiation therapy Radiation environment and medicine: covering a broad scope of topics relevant to environmental and medical radiation research 6(2) 55–61

[8] Lund C M, Famulari G, Montgomery L and Kildea J 2020 A microdosimetric analysis of the interactions of mono-energetic neutrons with human tissue Physica Medica 73 29–42. DOI: 10.1016/j.ejmp.2020.04.001

[9] Montgomery L, Lund C M, Landry A and Kildea J 2021 Towards the characterization of neutron carcinogenesis through direct action simulations of clustered DNA damage Physics in Medicine & Biology 66(20) 205011. DOI: 10.1088/1361- 6560/ac2998

[10] Manalad J, Montgomery L, Kildea J. A study of indirect action's impact on simulated neutron-induced DNA damage. Phys Med Biol. 2023 Mar 31;68(7). DOI: 10.1088/1361-6560/acc237

[11] Ottolenghi A, Baiocco G, Smyth V and Trott K 2015 The ANDANTE project: a multidisciplinary approach to neutron RBE Radiation protection dosimetry 166(1-4) 311–315. DOI: 10.1093/rpd/ncv158

[12] Baiocco G et al 2016 The origin of neutron biological effectiveness as a function of energy Scientific Reports 6 34033. DOI: 10.1038/srep34033

[13] Perl J, Shin J, Schumann J, Faddegon B and Paganetti H 2012 TOPAS: An innovative proton Monte Carlo platform for research and clinical applications Medical Physics 39(11) 6818– 6837. DOI: 10.1118/1.4758060

[14] Agostinelli et al, Geant4 – a simulation toolkit, Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment, Volume 506, Issue 3, 2003, Pages 250-303. DOI: 10.1016/S0168-9002(03)01368-8

[15] Cullen DE, Hubbell JH, Kissel L, EPDL97: the evaluated photo data library 97 version. Technical Report. Lawrence Livermore National Lab., CA.

[16] Perkins ST, Cullen DE, Seltzer SM, Tables and graphs of electron-interaction cross sections from 10 eV to 100 GeV derived from the LLNL evaluated electron data library (EEDL), Z= 1–100. UCRL-50400 31; 1991a, 21–24.

[17] Perkins ST, Cullen DE, Chen MH, Rathkopf J, Scofield J, Hubbell JH, Tables and graphs of atomic subshell and relaxation data derived from the LLNL Evaluated Atomic Data Library (EADL), Z= 1–100. Technical Report. Lawrence Livermore National Lab., CA; 1991b.

[18] Rene Brun, Fons Rademakers, ROOT — An object oriented data analysis framework, Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment, Volume 389, Issues 1–2, 1997, Pages 81-86, ISSN 0168-9002. DOI: 10.1016/S0168-9002(97)00048-X.

[19] Schuemann J, McNamara A L, Ramos-Mendez J, Perl J, Held K D, Paganetti H, Incerti S and Faddegon B 2019 TOPAS-nBio: An Extension to the TOPAS Simulation Toolkit for Cellular and Sub-cellular Radiobiology Radiation Research 191(2) 125–138. DOI: 10.1667/rr15226.1

[20] IAEA Phase-space database for external beam radiotherapy, IAEA NAPC Nuclear Data Section.

Recent progress in gMicroMC package for microscopic Monte Carlo simulations

Youfang Lai[1], Yuting Peng[1], Meiying Xing[3], Yujie Chi[2], and Xun Jia[1]

[1]Department of Radiation Oncology and Molecular Radiation Sciences, Johns Hopkins University, Baltimore, MD 21287, USA
[2]Department of Physics, University of Texas at Arlington, Arlington, TX 76019, USA
[3]Department of Radiation Oncology, University of Texas Southwestern Medical Center, Dallas, TX 75287, USA

Abstract Fundamentally understanding biological responses to ionizing radiation is of crucial importance. Mechanistic modeling by microscopic Monte Carlo (MC) simulation of the radiation physics and chemistry process triggered by radiation, as well as the process to generate initial DNA damages, plays an important role. However, such simulation is often very time consuming, limited the scope of simulation affordable. Our group has been developing open-source high performance microscopic MC simulation package – gMicroMC based on Graphical Processing Units (GPU). This paper summarizes our recent progress in recent developments in modeling the radiation transport process and DNA damage calculations for various DNA structures. To demonstrate the utility of this tool, we also present two example applications, namely radical production in electron FLASH irradiation, and relative biological effectiveness (RBE) of heavy ions calculated by considering DNA damage of various complexities.

1 Introduction

DNA damages and lesions are the primary lethal factors for radiation induced cell death. Of these, DNA double strand breaks (DSBs) play a key role. Recent developments from experiments by electrophresis and foci detection reveal the importance and complexity on the relationship between DNA damages and cell death. Besides the number of damage sites, the size and complexity also matter. Therefore, it is fundamental to understand the distribution of DNA damages after irradiation, especially considering that the damages are modulated by various factors of irradiation, e.g. dose, dose rate, linear energy transfer (LET) etc.

Monte Carlo (MC) simulation has been widely used in medical physics field due to its faithful description of radiation physics process and environmental conditions. The extension of conventional macroscopic MC simulation, e.g. for dose calculation purpose, into microscopic simulation, e.g. at micrometer or nanometer scale, including the particle transportation and chemical reaction, becomes an essential tool to model the DNA damage immediately (e.g. not accounting damage repairs) induced by ionizing radiation at sub-cellular levels. There is a continuous effort in developing such packages [1–3]. Along with the tool development for simulations, the modeling of DNA structure becomes more and more delicate to dig out more subtle yet basic perspectives into radiation damages. For instance, the change from a simple cylinder sensitive volume to detailed multi-scale DNA structure in a cell nucleus enables the modeling of the effect of histone proteins or a transformation from base damage to sugar damage.

A major hurdle faced by microscopic MC simulation with fine DNA structure modeling is low efficiency. Factors accounting for this challenge include, but are not limited to, the complex radiation physics and chemistry process to model, the many-body simulation problem in the radiation-chemical stage, the long chemical stage over multiple orders of magnitude in time, and the complex DNA geometry. Due to limitation in computation, suboptimal modeling setups have been employed, impeding resulting accuracy. For years, our group has been striving to develop high-performance MC packages with the use of Graphical Processing Units (GPU). We designed and optimized proper data structure and parallelism algorithms that fit the logic of GPU. Specifically, we developed a method to voxelize the space and sort particles based on their voxel index [4], thereby reducing the burden in tracking mutual interactions in the many-body radiation chemistry simulation. We built a DNA model for lymphocyte cell nucleus and the framework to transport electrons and decay of excited water molecules, which gave rise to the first version of gMicroMC [5]. Since then, this package has been maintained by our group and publicly shared to the research community *. We later added the function to explicitly consider oxygen molecules [6] and the support for proton and heavy ions in a concurrent transport framework of radicals and DNA structure[7]. The application of gMicroMC has enabled studies such as understanding oxygen depletion in FLASH condition [8] and modeling of relative biological effect (RBE) of carbon ions [9].

In this manuscript, we present the advancements in the gMicroMC package, highlighting our recent developments. We also provide illustrative examples demonstrating its utility in modeling DNA damages, thereby enhancing radiobiology modeling. We hope that this paper will introduce this versatile tool to the broader research community.

2 Materials and Methods

2.1 gMicroMC simulation process

The simulation was divided into four stages. In the physical stage, gMicroMC transports primary particles and secondary

*https://github.com/utaresearch/gMicroMC

electrons , recording positions and types of ionized and excited water molecules. In the physico-chemical stage, it computes the de-excitation process of these water molecules through predefined channels and probabilities, generating the initial distribution of radicals. After that, the simulation in the chemical stage diffuses these radicals and checks mutual reactions among them, as well as with DNA structure and other molecules, e.g. oxygen, in a step-by-step manner. Finally, searching along the DNA chain is performed for DNA damages of various complexities based on the DNA damages created in the first three stages. The advantage of gMicroMC is the high computational efficiency realized via GPU parallelization and GPU-friendly parallel simulation algorithms in all the four stages.

2.1.1 Transport simulation of radiation physics and chemistry process

Physical stages The physics processes modeled in gMicroMC were summarized in Table 1. We focused on the transport of charged particles. For photon irradiation, we initiated the simulation from the generated secondary electrons. The medium was considered to be water. For heavy ions, all cross sections with water were scaled based on those of proton's, $\frac{d\sigma_{ion}}{dW}(E_k) = Z^2 \times \frac{d\sigma_{proton}}{dW}(E_k/A)$, where W is the energy transfer, E_k, Z and A are the kinetic energy, charge and atomic mass, respectively. For more details of the modeling, users can refer to our previous publications [5, 7]. One thing worth pointing out is that the package was not designed for transporting particles over several centimeters. Hence, we ignored some processes that have marginal effect on transportation in micrometer scale, for instance the elastic process of proton. Due to the heavy mass of proton, the deflection angle is tiny for the length scale of interest. We validated the simulation of the physical stage by computing physics quantities, e.g. LET and secondary electron spectrum, and comparing with published data.

Physico-chemical and chemical stages After the physical stage, the excited and ionized water molecules generated a series of radicals based on a defined list of paths. In the chemical stage, the basic simulation for radicals' diffusion and chemical reactions was conducted in a step-by-step manner. For each time step, the radicals diffused freely while the reaction was determined by the distances between radicals

Table 1: Physics process between different particles and water in gMicroMC.

Process	Electron	Proton	Heavy ion
Ionization	Yes	Yes	Yes
Excitation	Yes	Yes	Yes
Elastic	Yes	No	No
Dissociative	Yes	No	No
Charge effect	No	Yes	Yes

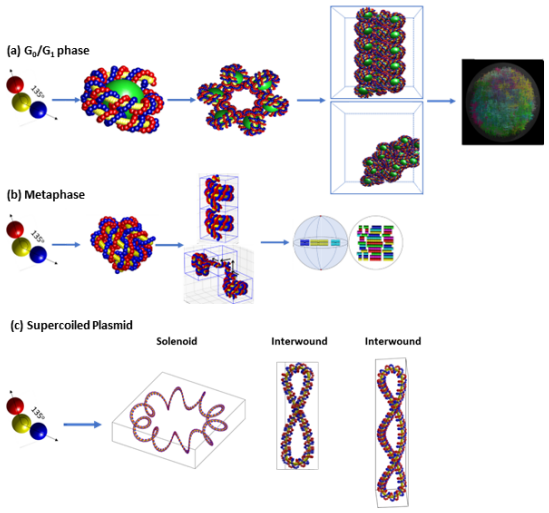


Figure 1: Illustration of different DNA models.(a) Lymphocyte cell nucleus model at G0/G1 phase, (b) meta phase, (c) supercoil plasmid model. Both the solenoid type with writhing number 10 and interwound type with writhing number 1 and 2 are displayed.

[4]. Due to the advantages of GPU's processing power, we afforded to include the presence of DNA in the chemical stage, where reactions between radicals and DNA were modeled explicitly. This increased the simulation realism, for instance the considerations of scavenge effect of DNA structure.

2.1.2 DNA models

gMicroMC currently includes two DNA models of lymphocyte cell nucleus. One is for G0/G1 phase and the other one is for meta-phase. The model for the G0/G1 phase considered the multi-scale nature of DNA packing in a cell nucleus of a sphere with 11 micrometer diameter (Fig. 1(a)). It partitioned the cell nucleus space into voxels of 55 nanometer, and filled each voxel with one of the two types of building blocks, namely the one connecting the opposite faces and the other one connecting adjacent faces. The connection path among voxels were designed by a thermal relaxation process of 46 chromosome structures. For each of the two building blocks, it was constructed by considering the hierarchical DNA structure containing chromatin fiber unit, nucleosome and base pairs. Under this multi-scale structure, 6.2 GBp DNA were packed into the the cell nucleus model.

The meta-phase DNA structure (Fig. 1(b)) used the same idea of multi-level structure. Yet, they differed in the definition of building blocks and the way to arrange them. The dimension of repeating unit was 11 nanometers and it constituted rod-shaped chromatid segment of 730 nm in diameter and 460 nm in height. The arrangement of these chromatid segments followed the coarse-grained polymer physics to match experimental Hi–C data.

Besides the two cell nucleus DNA models, we also provided a supercoiled plasmid DNA model (Fig. 1(c)). Although it is simple, it has the unique potential to allow benchmarking simulation results with experimental results, since experiments

can be performed using plasmid DNAs in a controllable fashion. We implemented both the solenoid and interwound type of folding. The elemental structure was the base pairs with the distance between adjacent base pair 0.34 nm. With different writhing number, different structure could be generated. Fig. 1(c) showed the supercoided DNA models with writhing number of 10, 1 and 2. The number of base pairs were 2050, 246, 416, respectively.

2.1.3 DNA damage calculation

The transport of the radiation physics and chemistry process generated a number of DNA damage sites along the DNA chain, directly by the charged particles' energy deposition, or indirectly by radicals' reactions with DNA sites. After damages sites were generated, we analyzed the geometrical locations of these damages to identify Single Strand Breaks, DSB and damages of higher complexities.

2.2 Example use cases

With the comprehensive developments, we presented in this section two examples to demonstrate the utility of gMicroMC.

The first example was the calculation of radical production in electron FLASH therapy realized using an electron beam of a medical linear accelerator. Specifically, we simulated the presence of electron tracks to a volume of interest within a RF pulse with a duration of 1 μs for instantaneous dose rates of 10^6 and 10^8 Gy/s. We simulated the entire radiation physics and chemistry process with oxygen in the environment explicitly considered, hence allowing us to study the dependence of radical yield on oxygen concentration.

The second example calculation was RBE modeling in heavy ion irradiation. This example included DNA damage simulation to demonstrate the utility of this function. The core component was to define a variable to measure the radiation effect and to be linked with biological endpoint of cell survival fraction (SF). As mentioned previously, both the size and complexity of DSB contributes to the biological effects. We hence defined the number of lethal DSBs (LDSBs) as a weighted sum of DSBs of various complexities $N_{\text{lethal}} = \sum_i \omega_i N_i$ where ω_i was the weight and N_i the count of ith DSB type. The weights were determined via fitting to experimental data. With this definition, we computed DNA damages of various complexities under the reference Co60 beam and various heavy ions including helium, carbon, and neon. Under the assumption of the existence of a universal survival model $SF(N_{\text{lethal}})$ over different types of irradiation, using known experimental SF data $SF_j(D_j)$, where D_j is the dose of jth type of irradiation, and the computed data $N_i = k_{i,j} D_j$, where $k_{i,j}$ is ratio between the number of the ith DSB complexity parameter and dose, we were able to derive N_{lethal} expression, and hence computing $SF(N_{\text{lethal}})$. With this function, RBE was obtained by computing the quotient of dose of x-ray and heavy ions to reach the given level of N_{lethal} corresponding to the endpoint $SF = 0.1$.

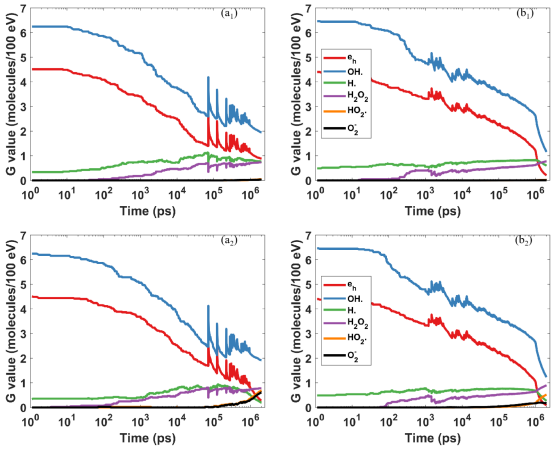


Figure 2: Comparison of the chemical yields under different dose rates of FLASH radiation with (a) 10^6 Gy/s and (b) 10^8 Gy/s. Top and bottom rows represent 0.1% (hypoxia) and 3% (physoxia).

3 Results

3.1 Computational efficiency

gMicroMC running on one GPU card is generally about 100 times faster than the CPU-based package in various simulation tasks we performed. For a test on transportation with one 750 keV electron executed on one Nvidia Titan Xp GPU (1.58GHz) and single core of Intel i7-6850K CPU (3.6GHz), the acceleration factor was 32 and 539 for physics stage and chemical stage, respectively. gMicroMC completes DNA damage pattern search in a few seconds.

3.2 Radical production in electron FLASH therapy

We presented the comparison of radical yields of under different dose rates in Fig. 2. The spikes were caused by injection of electrons within the pulse. As the instantaneous dose rate increased to a high level of 10^8 Gy/s, radical production process, as well as oxygen consumption, were significantly altered due to the spatial and temporal overlap of radicals and the mutual reactions among them.

3.3 RBE modeling in heavy ion irradiation

Fig. 3 presents the fitted N_{lethal} as a function of dose in Co60 beam and helium beam. Using this N_{lethal} quantity as a function of dose for different radiation types, we obtained the SF function of helium, carbon and neon beams with different energies and hence LETs, as presented in the top row of Fig. 4. This further allowed to compute RBE for different LETs shown in the bottom row.

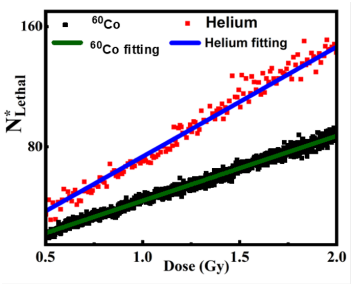


Figure 3: N_{lethal} as a function of dose for helium ions and ^{60}Co.

4 Discussion

The simulation of radical production in FLASH radiation demonstrated the power of our tool, as well as the insights regarding radical yields modulated by dose rate and oxygen environment. A more relevant question is DNA damages in these conditions, which are currently under exploration in our group. While DNA damages are generally related to radical numbers, the presence of DNA structures, e.g. histone protein, may lead to scavenging effects, complicating the dependence of DNA damage number and complexity on dose rate. More importantly, using the simple supercoil DNA, the simulation results can be compared with experimental results, providing insights to the understanding of FLASH effects.

The calculation of RBE via N_{lethal} offered a framework to model SF based on biologically relevant quantities. It hence bypass the need to empirically connect SF model parameters, e.g. α and β in the linear-quadratic model, with physics quantities such as dose, dose rate, and LET. The dependence of N_{lethal} on these quantities is naturally accounted by the microscopic MC simulation. The fundamental assumption in modeling SF and RBE is that DSBs of various complexities governs the cell survival behaviors. From the cell perspective, the biological process in response to DNA damages, e.g. the repair process, should be related to DSBs of various complexities, but it is expected that cell does not know, and react differently to, the type of radiation causeing the damages. Hence, even the process from initial DNA damages to the endpoint of cell survival is complex, we expect that there is

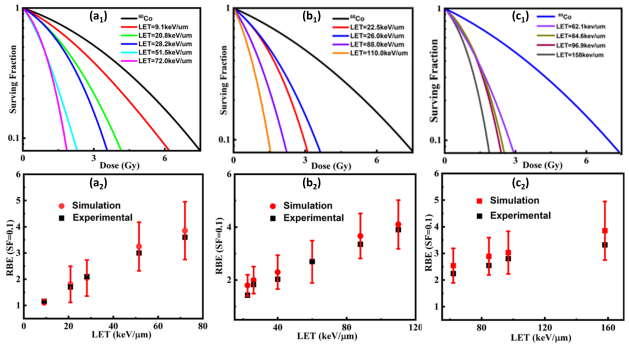


Figure 4: Simulated SF curves (top row) and RBE values (bottom row) for (a) helium, (b) carbon and (c) neon ions for V79 cells.

a universal function linking DNA damages to the endpoint. The term "universal" is in the sense that it is applicable to various radiation types. It is this assumption that allowed us to calculate SF using the same framework for various types of heavy ions and x-ray, and hence computing RBE. Yet it is possible that the universal function from DNA damages to cell survival is not only dependent on the single variable of N_{lethal}, but rather on each types of DSB N_i. In this case, a more complex function form is needed. However, due to scarcity of available data, we would run into the risk of overfitting. Hence, we chose to consider a single variable approach in this study and hope a linear combination of DSB of different complexities can capture the major behavior of the model.

5 Conclusion

This paper summarized the recent development of gMicroMC and presented two example applications. The versatile functions of microscopic MC simulation and radio-biological modeling of DNA structure and damages could benefit the research community in providing mechanistic prospective to understand the radiation damages to DNA.

References

[1] H. Nikjoo, S. Uehara, W. E. Wilson, et al. “Track structure in radiation biology: theory and applications”. *International journal of radiation biology* 73.4 (1998), pp. 355–364.

[2] S. Incerti, G. Baldacchino, M. Bernal, et al. “The geant4-dna project”. *International Journal of Modeling, Simulation, and Scientific Computing* 1.02 (2010), pp. 157–178.

[3] J. Schuemann, A. McNamara, J. Ramos-Méndez, et al. “TOPAS-nBio: an extension to the TOPAS simulation toolkit for cellular and sub-cellular radiobiology”. *Radiation research* 191.2 (2019), pp. 125–138.

[4] Z. Tian, S. B. Jiang, and X. Jia. “Accelerated Monte Carlo simulation on the chemical stage in water radiolysis using GPU”. *Physics in Medicine & Biology* 62.8 (2017), p. 3081.

[5] M.-Y. Tsai, Z. Tian, N. Qin, et al. “A new open-source GPU-based microscopic Monte Carlo simulation tool for the calculations of DNA damages caused by ionizing radiation—Part I: Core algorithm and validation”. *Medical physics* 47.4 (2020), pp. 1958–1970.

[6] Y. Lai, Y. Chi, and X. Jia. “Mechanistic modelling of oxygen enhancement ratio of radiation via Monte Carlo simulation-based DNA damage calculation”. *Physics in Medicine & Biology* 67.17 (2022), p. 175009.

[7] Y. Lai, X. Jia, and Y. Chi. “Recent developments on gMicroMC: transport simulations of proton and heavy ions and concurrent transport of radicals and DNA”. *International journal of molecular sciences* 22.12 (2021), p. 6615.

[8] Y. Lai, X. Jia, and Y. Chi. “Modeling the effect of oxygen on the chemical stage of water radiolysis using GPU-based microscopic Monte Carlo simulations, with an application in FLASH radiotherapy”. *Physics in Medicine & Biology* 66.2 (2021), p. 025004.

[9] H. Feng, W. Li, Y. Zhang, et al. “Mechanistic modelling of relative biological effectiveness of carbon ion beams and comparison withexperiments”. *Physics in Medicine and Biology* ().

Evaluation of a New Radiosurgery System (infinityARC) Using Geant4

J. Ngontie Ngounou[1,2], Y. Lemaréchal[1,2], and P. Després[1,2]

[1]Département de physique, de génie physique et d'optique, et Centre de recherche sur le cancer, Université Laval, Québec, QC, Canada
[2]Service de radio-oncologie et Axe Oncologie du CRCHU de Québec, CHU de Québec – Université Laval, Québec, QC, Canada

Abstract

In this study, Geant4 simulations were used to evaluate the infinityArc (iArc) radiosurgery system currently under development. This new system is equipped with a robotic arm that allows for the precise and safe movement of the equipment's head around the patient while delivering the beam. The radiation from 22 ^{60}Co sources passes through a collimation system towards the isocenter. To meet FDA approval standards, a comprehensive study protocol utilizing Geant4 simulations has been established, adhering to IEC 60601-2-11:2013 and NCRP 102 guidelines which are defining performance criteria to be achieved. Monte Carlo simulations were used to calculate the dose distribution in a spherical water phantom and to assess leakage radiation around the equipment's head. The dose profiles and isodose curves will be compared with physical measurements once the prototype is completed. Preliminary results suggest the efficacy of the shielding geometry. Nevertheless, further simulations and in-depth investigations are required to further characterize the system.

1 Introduction

The iArc is a novel radiosurgery system currently under development by a company founded by a group of experts, including medical physicists, engineers, and entrepreneurs dedicated to addressing global health inequalities. This new system will be used to treat functional brain disorders such as trigeminal neuralgia, arteriovenous malformations, vascular malformations, as well as benign and malignant intracranial and extracranial tumors.

The device consists of 22 ^{60}Co sources arranged in a cylindrical head. Gamma ray beams converge through a collimator system to the isocenter, where the target volume should be positioned. There are two collimation systems: the primary collimator with a fixed aperture and the secondary collimator with a variable aperture. The system is designed to rotate the treatment head around the patient at a constant distance from the isocenter.

In order to obtain FDA approval, a study protocol using Geant4 as a simulation tool has been established. The protocol, based on the IEC 60601-2-11:2013 standard and NCRP 102 report, outlines the steps to be followed and the criteria that the new system must meet to secure approval. The IEC 60601-2-11:2013 standard sets safety and essential performance requirements for gamma beam radiotherapy equipment, including multi-source stereotactic radiotherapy [1]. Manufacturers must comply with these requirements in the design and construction of such equipment. The NCRP 102 report provides recommendations regarding the design, performance, and optimal use of medical imaging and therapy systems utilizing medical X-rays, electron beams, and gamma rays for energies up to 50 MeV [2].

2 Materials and Methods

2.1 Monte Carlo simulation

The head of the iArc system was modeled with Geant4, a C++ based toolkit that employs the Monte Carlo method [3]. Geant4 is designed for high-energy physics and is also applicable to low-energy ranges, allowing its utilization in the medical field [4]. In this study, Geant4 version 11.1.0 (December 2022) was used. The system was dosimetrically characterized by assessing the dose distribution (dose profile and isodose) in a spherical water phantom. The dose profiles and isodose curves will be compared later with physical measurements, once the prototype is functional. In addition to dose profiles in a water phantom, the radiation leakage around the system head was evaluated.

2.1 IEC 60601-2-11 and NCRP 102 report

This study is based on a protocol that outlines the steps to follow and establishes the performance criteria to be achieved. The criteria are defined based on the standard 60601-2-11 and the NCRP 102 report.

These standards specify that the ratio of the dose at the isocenter to the dose at the surface of a 16 cm water phantom centered on the radiation isocenter should be greater than 30.

The assessment of leakage radiation at the system's head requires conducting several tests. In this paper, we will present two of the four tests to be performed:

- **IEC 60601-2-11 Compliance: Beam-on Stray Radiation (Clause 201.10.2.5.2.2)**: The shielding should sufficiently reduce stray radiation, ensuring that the dose rate at any point 1 m from the center of ^{60}Co sources is $\leq$ 0.5% of the maximum absorbed dose rate on the radiation beam axis at a 1 m distance.
- **IEC 60601-2-11 Compliance: Beam-Off Isocenter Dose Rate Assessment:** The beam-off dose rate should be less than 2% of the beam-on dose rate for the smallest secondary collimator size.

2.2 Dose distribution

In Geant4, three fundamental steps were considered: creating the physical geometry, defining the possible physical events, and finally execute the simulation process.

Each source in the system is enclosed within a steel bushing assembly. The configuration of the cobalt source is the same as the one used by Romano et al. (2007) [5]. In our study, the initial positions of photons were uniformly sampled within a cylinder with a diameter of 2.54 mm and a length of 20 mm. To achieve a reasonable computation time, the direction of photons was sampled within a cylinder with an angular distribution $\Delta\theta$ of 10°. Romano et al. (2007) indicated in their study that restricting the photon propagation does not have a significant effect on the final results.

The radiation emitted by each ^{60}Co source is collimated by a cylindrical primary and secondary collimator made of tungsten. The secondary collimator is available in three different configurations that provide three different dimensions of the beam diameter at the isocenter (4, 8 and 14 mm).

At 480 mm from the center of the sources is the spherical water phantom with a diameter of 160 mm, simulating the patient's head. To simulate a treatment of the patient's head, the system must rotate by 30 degrees around the x-axis in a clockwise (Figure 1) direction. Subsequently, we conduct a simulation for each angle, totaling 360 simulations to replicate the rotational movements performed during a surgical intervention. The summation of the entire set of static simulations yields the final dose distributions.

A cube with a side length of 105 mm was placed at the center to calculate the total absorbed dose at the end of the simulation. It is divided into 105 x 105 x 105 cubic voxels of 1 mm^3, where the deposited dose is recorded.

The standard electromagnetic processes were employed as the physics list in Geant4. A detailed description of physical processes can be found in Geant4 [6]. The particle production cut, expressed by Geant4 as a range of particles in the material, was set to 0.1 mm for the entire system.

A total of 1.584×10^{13} primary particles were generated for different configurations of the secondary collimator. The simulation was partitioned into 360 independent tasks, each with distinct seeds, on a Linux cluster to shorten the overall simulation duration.

2.3 Leakage radiation

The method employed for this part of the study is similar to the one described earlier, with a few exceptions. Table 1 outlines the specificities of this section.

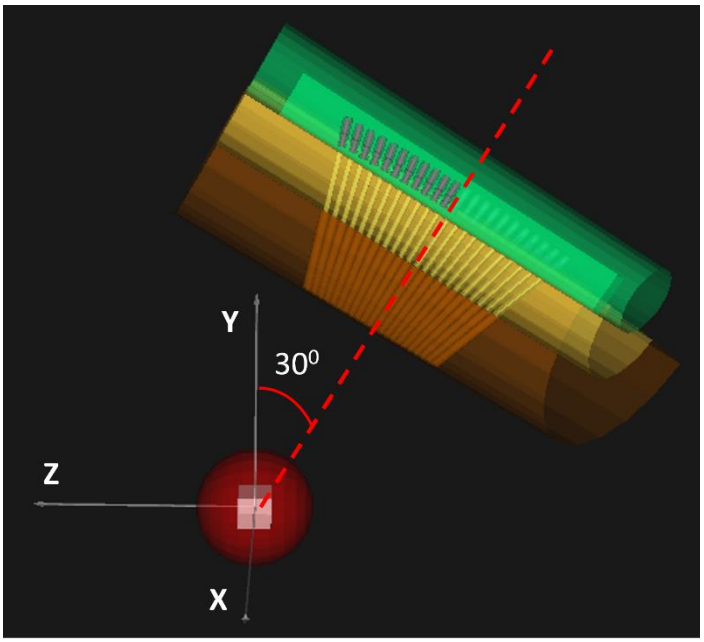


Figure 1: The iArc radiosugery system treatment head. The treatment head is in the (y,z) plane. The spherical water phantom is shown in red. The cubic detector is located at the center of the phantom.

Table 1:Detector configurations

	Particularities
Beam-Off Isocenter Dose Rate Assessment	Primary particles exhibit full isotropic angular distribution.
Beam-on Stray Radiation (Clause 201.10.2.5.2.2)	A central detector and 14 strategically positioned detectors within the water phantom. Primary particles exhibit full isotropic angular distribution.

3 Results

The results presented in this section are preliminary findings from the simulations conducted. Despite their preliminary nature, these results hold significant importance as they offer an initial understanding and assessment of the studied system.

3.1 Dose profile and isodose

For all collimator configurations, we obtained relative dose profiles along the x, y, and z axes, as well as isodose curves in the sagittal, coronal, and axial planes. We run 1 simulation per angle, for a total of 360 simulations per

collimator size, and then merged the results of these runs. The y-axis dose profile is identical to the x-axis as expected from the symmetry of the system. Figures 2-5 show the relative dose profiles and isodose curve.

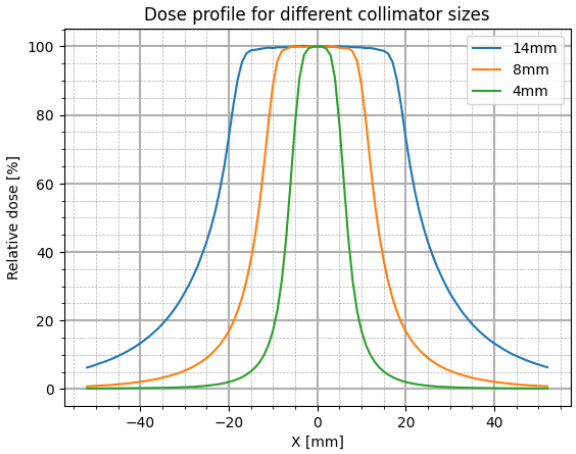


Figure 2: Dose profile along x-axis

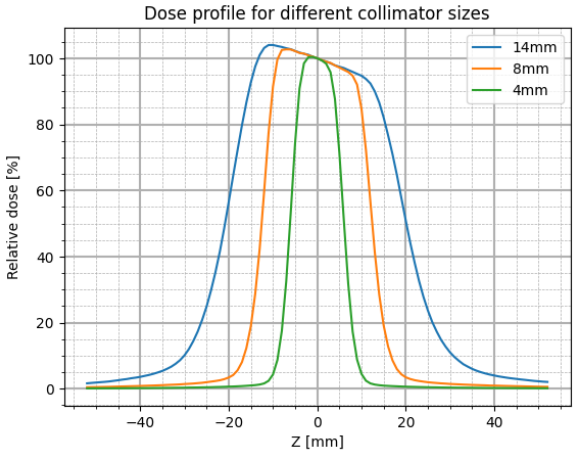


Figure 3: Dose profile along z-axis

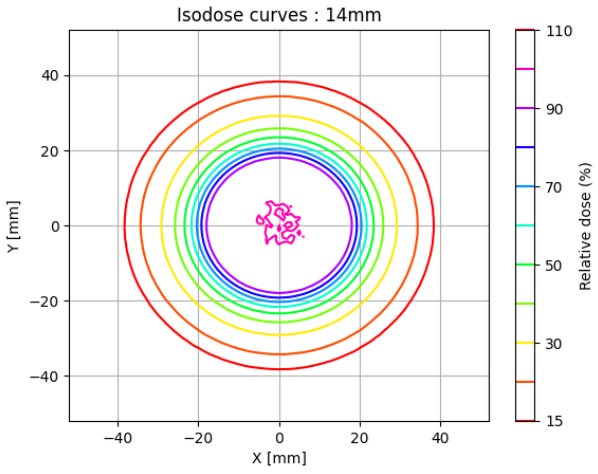


Figure 4: Isodose curves for the 14 mm collimator. Sagittal plane of cubic detector.

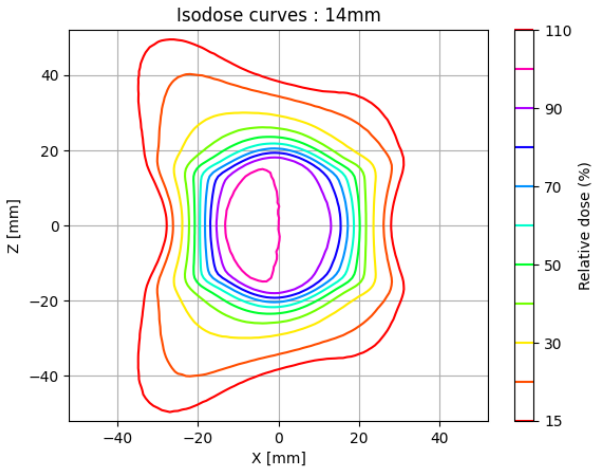


Figure 5: Isodose curves for 14 mm collimator. Axial plane of cubic detector.

3.2 Dose ratio

For dose ratio calculation, 22 detectors were positioned on the phantom surface, with an additional one at the center. Each detector is 1 x 1x 1 mm^3 in size. The ratio is obtained by dividing the total dose at the phantom center across 360 simulations by the surface dose. Table 2 shows the minimum and maximum dose ratio values for each collimator configuration.

Table 2: Dose ratios for different collimator sizes

Collimator	Ratio [min, max]
14 mm	[553,1011]
8 mm	[903,1478]
4 mm	[1742,2314]

3.3 Leakage radiation

3.3.1 Beam-off isocenter dose rate assessment

This test exclusively requires the use of the smallest secondary collimator (4 mm), as it represents the lowest dose rate when the system is in beam-off mode. In total, 360 simulations were conducted, corresponding to a one-degree rotation per simulation. Preliminary results indicate that no particles reach the isocenter, suggesting a zero- dose rate at the isocenter when the system is in beam-off mode.

3.3.2 Beam-on Stray Radiation (Clause 201.10.2.5.2.2)

The scope of this test is to evaluate the leakage radiation emitted when the iArc radiosurgery system is in the beam-on conditions measured at 1 m from the center of the ^{60}Co

source distribution. For each of the three collimator sizes, the dose resulting from leakage radiation, as measured by the 14 detectors, is compared with the dose measured by the detector positioned at the center of the phantom.

The initial results indicate that the 4 mm collimator has a single detector having a dose ratio exceeding 0.5 %. For the 8 mm collimator, four detectors exceeded this threshold, and for the 14 mm collimator, thirteen detectors surpassed it.

4 Discussion

4.1 Dose profile and isodose curves

The dose profiles and isodose curves indicate that the dose distribution along the X and Y axes is symmetrical with respect to the axis. However, along the Z axis, symmetry is not observed. This asymmetry is attributed to the 30-degree rotation of the iArc, causing the beam arriving on the Z axis to be oblique. The profiles and isodoses will be compared later to physical measurements.

4.2 Dose ratio

The results indicate that the ratio of dose at the center compared to the dose at the surface increases with the reduction of the collimator diameter. This is logical as a smaller diameter result in less scattering contributing to the surface dose. Initial simulations show that the dose ratio significantly exceeds 30, consistent with the protocol expectations.

4.3 Leakage radiation

4.3.1 Beam-off isocenter dose rate assessment

The preliminary data showing that no particles reach the isocenter when the system is in beam-off mode. These initial results highlight the potential for zero dose at the isocenter. However, it is imperative to emphasize that these conclusions are provisional, and a thorough analysis will only be possible after the completion of the entire planned set of simulations.

4.3.1 Beam-on Stray Radiation (Clause 201.10.2.5.2.2)

The leakage radiation test results offer insights into the iArc system's performance under beam-on conditions, with a criterion of maintaining a dose ratio below 0.5% for all detectors. Notably, the 4 mm collimator demonstrates promising outcomes, surpassing the threshold with a single detector. However, deviations in the 8 mm and 14 mm collimators, where four and thirteen detectors exceeded the limit, respectively, indicate challenges in achieving the desired dose ratio with larger collimator sizes. This highlights the need for further optimization and investigation to enhance system performance.

5 Conclusion

The preliminary results of the study on the iArc radiosurgery system are promising despite their preliminary nature. The dose profiles and isodose curves demonstrate a symmetrical distribution along the X and Y axes, while an asymmetry is observed along the Z axis due to the 30-degree rotation of the iArc. Dose ratios increase with the reduction of the collimator diameter, suggesting less scatter. Leakage radiation tests, particularly with the 4 mm collimator, reveal promising results, but potential challenges are highlighted with larger collimators. These findings underscore the need for optimization and in-depth investigations to enhance the overall performance of the iArc system. The data provides a solid foundation for future studies.

References

[1] https://webstore.iec.ch/publication/2618

[2] https://ncrponline.org/publications/reports/ncrp-reports-102/

[3] S. Agostinelli *et al.*, “Geant4 – a simulation toolkit”, Nucl. Instr. Meth.in Phys. Res. A **506**, 250-303 (2003). DOI:10.1016/S0168-9002(03)01368-8

[4] G. A. P. Cirrone *et al.*, “The GEANT4 toolkit capability in the hadrontherapy field: simulation of a transport beam line”, Nucl. Phys. B– Proc. Suppl. **150** (2006) 54. DOI:10.1016/j.nuclphysbps.2005.04.061

[5] F. Romano, M. G. Sabini, G. Cuttone, G. Russo, V. Mongelli and R. Foroni, "Geant4-based Monte Carlo Simulation of the Leksell Gamma Knife®," 2007 IEEE Nuclear Science Symposium Conference Record, Honolulu, HI, USA, 2007, pp. 2581-2586, DOI: 10.1109/NSSMIC.2007.4436677

[6] https://geant4-userdoc.web.cern.ch/UsersGuides/PhysicsReferenceManual/BackupVersions/V11.0/html/index.html

Oxygen Effect on Tumor Response to Irradiation: Insights from GATE Simulation Studies

Zakaria Ait Elcadi[1] and Othmane Bouhali[1]

[1]Electrical and Computer Engineering, Texas A&M University at Qatar, Doha, 23874, Qatar.

Abstract The Oxygen effect significantly influences tumor cell response to irradiation, particularly in cases of tumor hypoxia. This study, using the GATE simulation code, evaluates dose distribution in a spherical tumor within the right lung under varying oxygen concentrations. The Oxygen Enhancement Ratio (OER) is employed to quantify the impact of oxygen pressure on radiation response. Two phases of investigation involve simulations of an XCAT phantom with a 1 cm diameter spherical tumor, applying oxygen levels from 5 to 80 mmHg. Results demonstrate a correlation between deposited energy, OER values, and oxygen pressure. The study reveals that low oxygen pressure regions require higher doses for comparable effects, emphasizing the need for tailored treatments based on tumor oxygenation. This research sets the groundwork for future studies on physiological factors influencing hypoxic tumors and their spatial relationship to varying tissue densities.

1 Introduction

The Oxygen effect plays a pivotal role in how tumor cells respond to irradiation. Tumor hypoxia has been acknowledged as a limiting factor in treatment response for many years. Diverse approaches to enhance the rate of tumor control have been experimented with, aiming to amplify the radical cascade [1].

Cells with an oxygen partial pressure (pO_2) below 8-10 mmHg are generally categorized as hypoxic [2] and exhibit greater radio-resistance compared to cells with normal oxygen levels (pO_2 up to approximately 100 mmHg [2]). This phenomenon can be quantified through the Oxygen Enhancement Ratio (OER), representing the ratio between the dose at a specific oxygen pressure and the dose at normal oxygen pressure that produces an equivalent biological effect ICRU Report 30: Quantitative Concepts and Dosimetry in Radiobiology. [3]

In this study, the dose distribution was assessed in a spherical tumor positioned at the center of the right lung. This tumor was generated using an XCAT phantom of an elderly woman, and various oxygen concentrations were considered. The evaluation was conducted using the GATE simulation code (V9.3) [4], employing the Monte Carlo method.

2 Materials and Methods

The investigation into the impact of oxygen on dose distribution was conducted using a female XCAT phantom, featuring a spherical tumor with a 1 cm diameter positioned at the center of the right lung (figure 1). This selection was influenced by the effect of low-density material on a spherical tumor with a density comparable to that of soft tissue. Various levels of oxygenation were applied to the spherical tumor to assess its response under different oxygen concentrations. The study utilized previously validated phase specifications of the Varian Clinac IX for a photon energy of 6MV [5] . The field size was fixed at 1.2x1.2 cm², employing removable jaws for a spherical tumor with a 1 cm radius, ensuring the tumor's central placement within the irradiation field. Oxygenation levels applied to the tumor ranged from 5 to 80 mmHg. The XCAT phantom was positioned at a source-surface distance of 100 cm. Results were analyzed based on deposited energy and OER values relative to O_2 pressure in mmHg, following a specific equation [6].

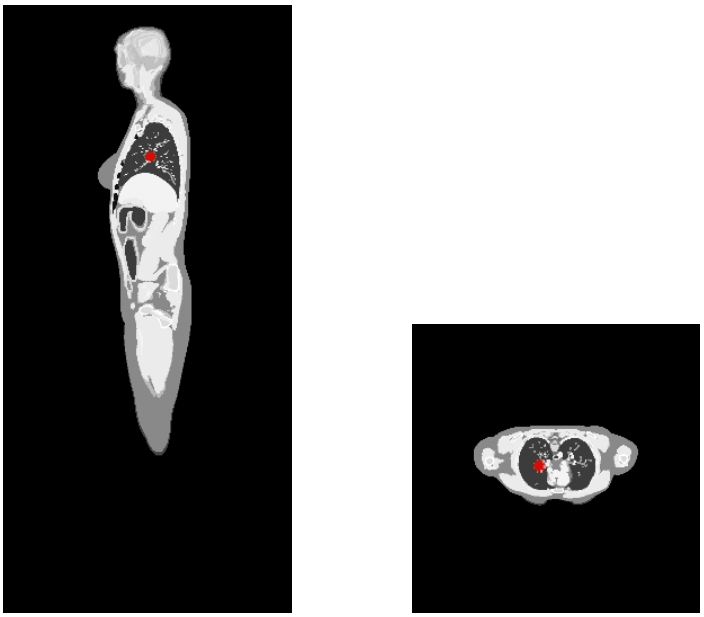


Figure 1: Coronal and transversal views of the female XCAT phantom show a spherical tumor centered in the right lung with a diameter of 1 cm.

3 Results

Nine simulations were conducted for the selected pressures in the initial phase of our study (figure 2). Figure (3) illustrates the deposited energy based on the O_2 pressure in the spherical tumor, while figure (4) displays OER values across O_2 pressures from 5 to 80 mmHg. The statistical error was a minimum of 2% for each simulation.

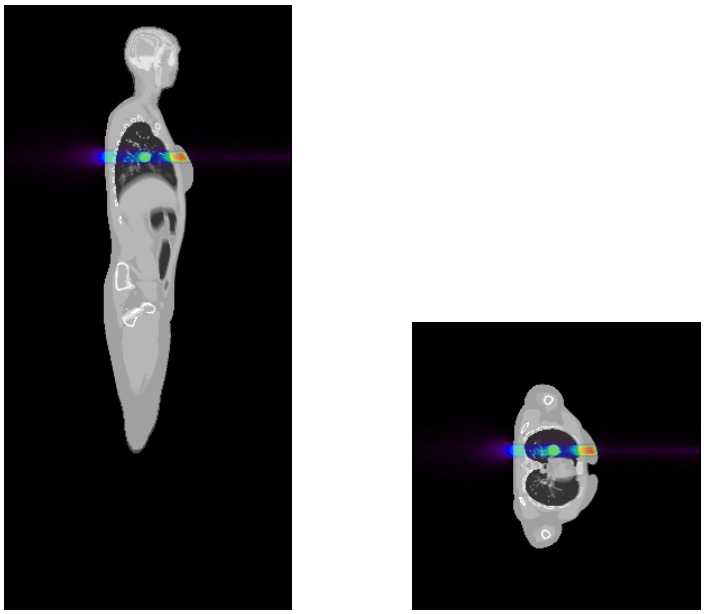


Figure 2: Coronal and transversal views of the female XCAT phantom, simulated in GATE and targeting the spherical tumor in the right lung, exhibit a 1 cm diameter with a 1x1 cm² field size.

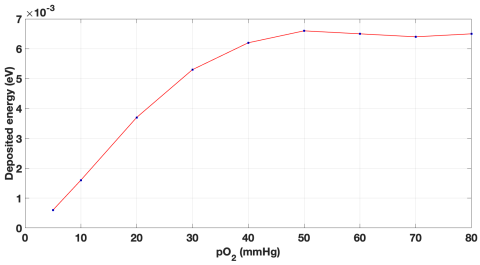


Figure 3: Energy deposition for various O_2 pressures applied to the spherical tumor.

4 Discussion

The figure 3 indicate that the deposited energy increases as the O_2 pressure grows, reaching a plateau at 60 mmHg. Beyond this point, the ratio remains constant (>60 mmHg), as illustrated in the figure. The OER values relative to O_2 pressure exhibit a linear growth up to 60 mmHg, followed by stabilization (figure 4).

In summary, the figures lead to the conclusion that regions with low oxygen pressure necessitate characterization by low deposited energy, requiring a higher dose to achieve effects comparable to regions with normal oxygenation. This consideration extends to the same biological effects to be achieved and the specific environment of the tumor, influencing the choice of a region with low density for this study.

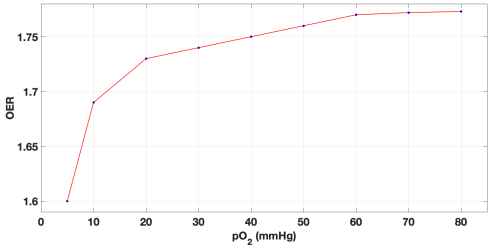


Figure 4: OER values are associated with each O_2 pressure applied to the spherical tumor.

5 Conclusion

The aim of this study was to explore the importance and necessity of oxygen (O_2) pressure in radiotherapeutic treatments. It was noted that the selected dosage for these treatments may differ, not dependent on the tumor and its location, but rather on the oxygen pressure within the tumor itself. This leads to the subsequent step of initiating a new study on physiological factors and biological effects that must be considered for tumors exhibiting hypoxia. Additionally, the study will examine the location of these tumors in relation to regions with varying densities, along with evaluating the dose received by neighboring organs.

References

[1] R. D. olveira and R. Alves. “Agentes antineoplásicos biorredutiveis: uma nova alternativa para o tratamento de tumores sólidos.” *Quim. Nova* 25 (2002), pp. 976–984.

[2] A. Carreau, B. E. Hafny-Rahbi, A. Matejuk, et al. “Why is the partial oxygen pressure of human tissues a crucial parameter? Small molecules and hypoxia.” *J Cell Mol Med* 15 (2011), pp. 1239–1253.

[3] I. R. 30. *Quantitative Concepts and Dosimetry in Radiobiology*. Tech. rep. International Commission on Radiation Units and Measurements, 2023.

[4] opengatecollaboration. *http://www.opengatecollaboration.org/*. Tech. rep. opengatecollaboration, 2023.

[5] Z. A. Elcadi, A. Bannan, R. E. Baydaoui, et al. “FEASIBILITY OF EXTERNAL RADIOTHERAPY DOSEESTIMATION IN HOMOGENOUS PHANTOM USINGMONTE CARLO MODELING”. *JATIT* 98 (2020), pp. 1151–1162.

[6] J. Kirkpatrick, L. Cárdenas-Navia, and M. Dewhirst. “Predicting the effect of temporal variations in po2 on tumor radiosensitivity”. *International Journal of Radiation Oncology*Biology*Physics* 59 (2004), pp. 822–833.

Monte-Carlo modeling of intratumoral radionuclide distribution in TAT: comparison to experiment

Victor Levrague[1], Mario Alcocer-Ávila[2], Etienne Testa[2], Michaël Beuve[2], and Rachel Delorme[1]

[1]Univ. Grenoble Alpes, CNRS, Grenoble INP, LPSC-IN2P3, 38000 Grenoble, France
[2]IP2I, Université Claude Bernard Lyon 1, CNRS/IN2P3, 4 rue Enrico Fermi, 69622 Villeurbanne Cedex, France

Abstract In the context of predicting biological effects in Targeted Alpha Therapy, nano- and micro-dosimetry are essential because of the very heterogeneous dose deposition at cell level.
The NanOx [1] biophysical model has been developed and adapted to predict cell survivals to low energy ion irradiations. Combining CPOP to create complex spheroidal multicellular geometries and Geant4 to perform Monte-Carlo simulations, cell survivals can be calculated through NanOx algorithm for a given irradiation condition. This study's objectives are to investigate the impact of intratumoral radionuclide heterogeneities on mean cell survivals with our biophysical modeling chain, and to compare the results with the *in vitro* experiment of Neti et al. [2] and results obtained via the MIRDCell [3] tool.
The impact of a lognormal law on the initial activity per labeled cell, with varying shape factor, was evaluated, as well as the percentage of radiolabeled cells.
At 1% cell labeling, both MIRDCell and our model overestimate with several orders of magnitude cell survivals from the experiment. For 10% and 100% labeling, our simulation pipeline gave predictions with up to 0.89 R^2 for lognormal distributions.
This study highlights the consistency of our model's predictions with those of MIRDcell and the importance of considering realistic radionuclide distributions in calculations. However it also highlights the difficulty in reproducing as faithfully as possible experiments with limited controlled parameters.

1 Introduction

Targeted Alpha Therapy (TAT) is an emerging therapy, to treat localized cancers. It benefits from the short range of alpha particles emitted by specific alpha-emitter radionuclides, that are fixed to cancer sites with a biochemical vectors. Its recent rise is due to the improvements made in tumor targeting [4]. Being able to provide accurate dosimetry and predict part of the treatment efficacy would help clinicians in assessing the required injected activity for a patient. However, the short range of alpha particles (<100 µm, equivalent to a few cells) induces heterogeneity of dose deposition at the micrometric scale. A macroscopic dose would not be enough to predict cell survivals at the tumor scale.
Biophysical models, in the litterature, developed to calculate cell survivals to ion irradiation from nano and/or microdosimetry are validated for energies of hadrontherapy ($\sim$ 50-200 MeV/n). Experimental data are scarce to parameterize and validate the biophysical model's predictions at the few MeV of TAT. To that aim, we are currently adapting the NanOx biophysical model [1] to such low energies. Formalism and its hypotheses were previously described, and prediction results were compared to those of the previous formalism used for hadrontherapy applications to ensure autoconsistency and extension of the model's capabilities (article in submission process). In another study, we have explored the impact of intracellular distribution of radionuclides in TAT on several indexes like the nucleus physical dose or the Tumor Control Probability (article also in submission process). These studies were performed with a uniform intratumoral radionuclide distribution. This does not represent the variety of experimental cases. Moreover, there is a need for the biophysical model to be strengthened by experimental validations. The present study is dedicated to the comparison with the Neti et al. [2] experiment, that was chosen because their experimental parameters are given and controlled enough to be reproducible by simulation. The main objective of this study is to explore the impact of intratumoral distribution in TAT outcome predictions, and to compare the results with measures from the literature. Results will also be compared with the MIRDcell analytical-MC tool [3].

2 Materials and Methods

2.1 Multicellular irradiations

The multicellular irradiations were performed with CPOP [5]. CPOP is a layer on top of the Monte Carlo (MC) code Geant4, that allows the generation of complex multicellular geometries in which the particle tracking occurs. Due to its ability to allow flexible cell size and membrane deformation, compacted geometries can be used to reproduce as much as possible the different compaction level observed in *in vitro* or *in vivo* experiments.
Different degrees of intratumoral heterogeneity were considered in this study. Firstly, the percentage of radiolabeled cells was varied as 100%, 10% or 1%, meaning that only those cells contain alpha-emitter radionuclides. Secondly, the distribution of the initial activity per labeled cell (i.e. the number of alpha emitters) was investigated from a uniform to a lognormal law. Studies like Pasternack *et al.* [6] showed the importance of taking into account this type of distribution, to evaluate cell survivals. A lognormal law can be described with the mean activity per cell, and the shape parameter σ.

2.2 Cell survival calculations

Cell survival was chosen as the biological index in this study, due to the current knowledge on it and the model's abilities. The NanOx biophysical model was used to compute the cell survivals. This model has been developed for hadrontherapy applications, validated by comparison against experimental data and benchmarked to other biophysical models [1]. NanOx (like the other biophysical models) considers that only damage in a cell nucleus can induce cell death. A modified version of the NanOx formalism was used in this paper, to take into account the energy loss of ions in cell nuclei. This formalism relies on two main approximations, namely the "narrow track" and "charge equilibrium" approximations. The first one means that the secondary electrons emitted by ions are concentrated along the ion's trajectory at distances lower than 100 nm. Their tracking is therefore not required in the MC simulations. With the "charge equilibrium" approximation, all the energy transferred by the ion in the cell nucleus is deposited in it.
In MC simulations, any time an ion crosses a cell nucleus with an entrance energy E_i and an exit energy E_f, this can be converted into damage to the cell. These damage can be either lethal n_L, at the nanometer scale, or sublethal n_{SL} at the micrometric scale. Cell survival S is calculated as follows:

$$S = \exp(-n_L - n_{SL}) \quad (1)$$

2.3 Comparison to experimental data

In the experiment performed by Neti *et al.* [2], multicellular clusters of chinese hamster V79 lung fibroblast cells were created, by centrifugation. A defined percentage of cells were labeled with Po-210, a radionuclide that mainly emits a 5.3 MeV alpha particle. Clusters were made of millions of cells. For computing time and memory purposes, the cluster size has been reduced to 100 µm spheroid radius. According to the data given in the paper, the compaction level was $\sim 47\%$, and $\sim 22\%$ of radionuclides were internalized in nucleus, the rest remaining in the cytoplasm.
During the experiment, cells were irradiated in the centrifugation tube. The choice was made in this study to use a spheroidal geometry, to model a part of the cluster in the tube.
Three timelines could be defined: the activity uptake (0.5 h), the cluster maintenance (72 h) and the colony formation (96 h). Cells were maintained at a temperature that allowed them to stay in a cluster arrangement up to 72 h, without declining their ability to clone themselves [2]. After 72 h, cells were washed. As a result, radionuclides decays were supposed to be mainly happening during the cluster maintenance.
The parameters described in this paragraph were used for the following simulations. For 100, 10 and 1% cell labeling, the maximum activity per labeled cell were 0.1, 1 and 24 mBq. The default shape parameter of 2.0 for the lognormal distribution was used to follow the recommendations of [3], and different values will be tested to investigate different distributions. The cell survivals were also compared with results obtained via the MIRDcell tool [3], with the same parameters. To compare our work to the results published in Neti *et al.*, a R^2 test for goodness of fit was performed, on logarithmic cell survivals.

2.4 Studied data and parameters

The different data in this study are the measures from Neti et al., the calculations performed with MIRDCell and the ones from the CPOP - Geant4 - NanOx model (CGN). Simulation results will be compared between themselves, and against the experimental data. The studied parameters will be: the percentage of cell labeling, the type of activity distribution in cells (between uniform and lognormal), and the shape factor of a lognormal distribution. Results will be presented as a function of the mean injected activity per cell, calculated for a time of 94h. When the cluster activity is calculated, simulated activities are normalized to scale from the 2480 cells modeled to the 4.10^6 cells in the experiment.

3 Results

3.1 Impact of the distributions of initial activity per cell

Figure 1 shows the mean cell survival as a function of the mean initial activity per cell, with 10% cells radiolabeled. Are presented the experimental data from Neti *et al.*, and the prediction of our CGN model with an uniform and lognormal distribution of activity per cell. The measured mean cell survival is a fit given by the authors for their data. The shape factor (σ) of the lognormal law varied between 0.5 and 2.0. For a 2.0 lognormal shape factor, the CGN model overestimates cell survivals by up to 4 times higher than the experimental data. The lognormal distribution with σ equal to 1 stay within 2.5 times the cell survivals from the uniform distribution, $\sim$ 10 times lower than the lognormal with σ = 2.0.
For the 0.5, 1 and 2.0 shape factor, the R^2, compared to the experimental data, are respectively equal to 0.55, 0.89 and 0.23. It is equal to 0.44 for the uniform distribution.

Figure 2 represents the mean cell survival as a function of the mean activity per cell, with 100% of cells labeled with radionuclides. With MIRDcell and our CGN model, results with both uniform and lognormal radionuclide distribution are presented. Results are presented with a 2σ error, for the rest of the study.
In this logarithmic representation, the experimental mean cell survival in the Neti experiment follows a linear law with a slope of ~ -0.993, and starts to "saturate" with increasing cell activity.

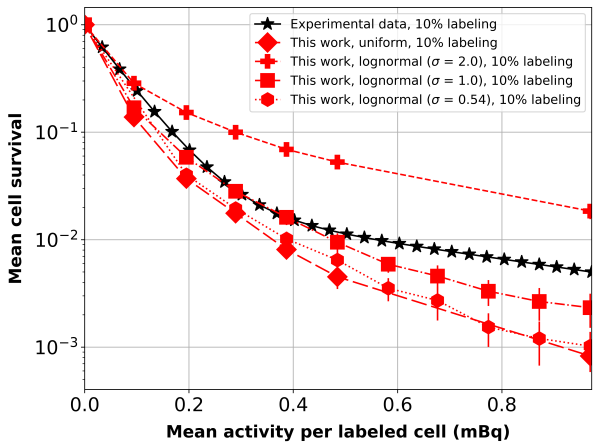


Figure 1: Mean cell survival calculated as a function of the mean activity in each cell, with 10% cells radiolabeled. Experimental data from Neti et al. [2] are in stars (solid line). Data calculated with the CGN model distributed via a lognormal law with a shape factor of 2.0, 1.0 and 0.54 are, respectively in plus signs (dashed), squares (dashdotted), hexagons (dotted). The simulations with a uniform distribution are in diamonds, with long dashes.

For lognormal distributions, MIRDcell mean survivals are 2 to 6 times higher than survivals from the CGN model.
For both models (MIRDcell and CGN), applying a lognormal distribution to the number of radionuclides in cells, increases the cell survival. For example, for the CGN results, differences can go from 10-fold differences at 0.015 mBq per cell to more than 100-fold above 0.030 mBq per cell.
For the CGN results, the R^2, compared to measured data, are 0.28 (lognormal distribution), and -47.6 (uniform distribution). For MIRDcell, R^2 are 0.78 (lognormal distribution), and -17.8 (uniform distribution)

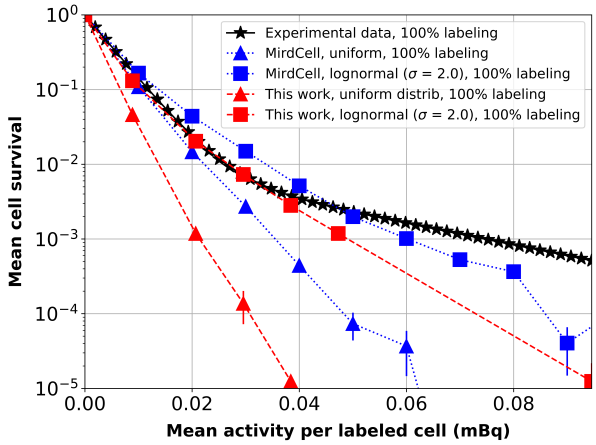


Figure 2: Mean cell survival calculated as a function of the mean activity in each cell, with all cells labeled with radionuclides. Experimental data from Neti *et al.* [2] are in stars (solid line). Simulation data with a uniform distribution are in triangle, those with a lognormal one are in squares. MIRDcell data are plotted in dotted lines, and CGN in dashed lines.

3.2 Impact of the percentage of cell labeling

Figure 3 shows the mean cell survival as a function as the total cluster activity in kBq.

For the same total cluster activity, cell survival differences range from $\sim$ 2 at 0.04 kBq to 10-folds at 0.39 kBq. The lower the % of radiolabeling, the higher the mean cell survival. In the experimental data, survivals at 1 and 10% cell labeling are similar. For MIRDcell and CGN, differences can go up 20-fold.
Looking at the 1% cell radiolabeling, the R^2, compared to experimental data, with MIRDcell are -4.61 (lognormal distribution) and -3.63 (uniform distibution). With NanOx, they are -4.47 (lognormal distribution) and -3.51 (uniform distribution).
The minimum of mean cell survival obtained with our model is 0.183 ± 0.007. For the Neti experiment, the minimum is equal to $1.7 * 10^{-4}$, for the same activity per labeled cell.

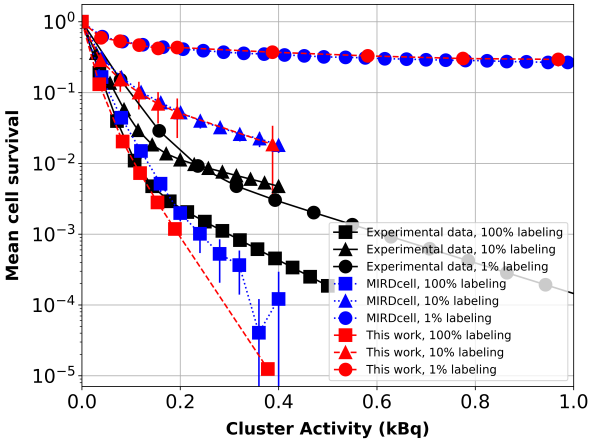


Figure 3: Mean cell survivals as a function of the total cluster activity (kBq). The measures from Neti et al. (solid lines), and calculations performed with the MIRDcell (dotted lines) and CGN (dashed lines) are presented. All calculations were done with a lognormal distribution on the number of radionuclides per labeled cell, with a shape factor of 2.0. Neti et al. data are a fit given by the authors on their experimental data. Data at 100, 10 and 1% cell labeling are, respectively, in squares, triangles and circles.

4 Discussion

This study confirms that applying a lognormal distribution seems to have a major impact in cell survivals assessment, as previously shown in the literature. A lognormal distribution could be creating some areas with lower radionuclide concentration. Even though, alpha particles can cross $\sim$ 40 µm, it is not enough to irradiate entirely a cluster of 100 µm radius. Under the activities observed, there is never a complete saturation because the emptier areas will eventually get irradiated by fixed radionuclides in this area or nearby, if a majority of cells are radiolabeled.

The shape parameter of a lognormal law can be defined as potential variations allowed for the distribution. When it tends towards zero, the distribution becomes closer to an uniform one and cell survivals get lower, as observed in Figure 1 and confirmed by experiments in the literature [6]. To model *in vitro* experiments, it appears mandatory to have measured the shape factor beforehand to set up the simulations with a lognormal distribution on the number of radionuclide per labeled cell. Considering only an uniform distribution of radionuclides could explain an overkilling in cell predictions.
The lower cell survivals when a lognormal law is applied can be associated to an overkill effect. Adding more radionuclides to a cell zone that is already mostly dead will not have an important effect on cell killing. The untouched cells will keep increasing the mean cell survival.

Results between MIRDcell and our model have similar orders of magnitude. The differences can be explained with the NanOx being closer to the physical reality. In particular, the (α, β) coefficients for cell survival calculations are constant in MIRDcell, while varying as a function of the energy in NanOx. Furthermore, the NanOx model was built to be flexible and easily enhanced.

At 100% cell labeling, according to the R^2 goodness of fit test, the simulations with a lognormal distribution at $\sigma = 2.0$ are fitting better the experimental data. The negative R^2 from the uniform distribution means that the prediction is worse than using the average horizontal value. Except for MIRDcell under 0.05 mBq per cell, models tend to overestimate cell death. Especially, the saturation of cell survivals in the Neti et al. experiment is not well reproduced, and the explanation is not straightforward. Several hypotheses are under investigation like cellular protection could be happening, as well as reparation effects.

To explain why cell survivals at 1% radiobeling are off the charts compared to Neti et al. experiments, regardless of the model used, multiple paths are guessed. Firstly, as pointed in the reference article [2], bystander could be a non-considered effect that increases cell death. Adding bystander damage to the model is an option to understand if this path is enough to explain the differences. Secondly, experimental uncertainties not taken into account in this paper could induce some deviations. For example, radionuclides could diffuse in the cluster and alpha particles emitted could hit more cells than in our simulations.
The saturation in our simulations could be explained by a geometric effect, where all cells that could be hit by helium ions are already dead, and increasing the dose would only kill those already dead cells. Moreover, the washes made to remove radionuclides from cells could not be optimal and cells could be still irradiated during the clonogenic assays, which this study doesn't take into account.
The Geant4 simulations were performed only in spheroid configurations. But, in the Neti et al. experiment, the geometry changes after 72 h, with cells still being irradiated. The impact of this geometry change has yet to be evaluated.
The effect of the percentage of radiolabeling heterogeneity, major in our simulations, is slim between the 10 and 1% Neti et al. experiments, although significantly different at 100% labeling. A comprehensive study of the impact of this type of heterogeneity on cell survivals or other physical and biological indexes will be done in the future.
This study underlines the difficulty of working with experimental data of low energy ion irradiation. Even with the multitude of controlled parameters in the Neti et al. experiment, many uncertainties restrain the perfect reproduction of it. However, the flexibility of our simulation pipeline allows the user to adapt to many complex irradiation configurations.

5 Conclusion

This ongoing study is a first validation step of the CGN pipeline for TAT. To model and predict properly *in vitro* experiments, one has to consider the intratumoral distribution of radionuclides in cells and, at best, to have it first measured depending on the vector. On top of that, knowing the ability for the targeting vector to fix to cells, i.e. the percentage of radiolabeling is mandatory to avoid errors up to several orders of magnitude.
The results given by our pipeline, using an intratumoral distribution of radionuclides is promising, although gaps have to be overcome, like the lack of saturation in our cell survivals for high radiolabeling and the severe underkilling at low radiolabeling.

References

[1] C. Monini, G. Alphonse, C. Rodriguez-Lafrasse, et al. "Comparison of biophysical models with experimental data for three cell lines in response to irradiation with monoenergetic ions". *Physics and Imaging in Radiation Oncology* 12 (Oct. 2019), pp. 17–21.

[2] P. V. S. V. Neti and R. W. Howell. "Biological response to nonuniform distributions of 210Po in multicellular clusters". *Radiation research* 168.3 (2007), pp. 332–340.

[3] S. Katugampola, J. Wang, A. Rosen, et al. "MIRD Pamphlet No. 27: MIRDcell V3, a revised software tool for multicellular dosimetry and bioeffect modeling". *Journal of Nuclear Medicine* 63.9 (2022), pp. 1441–1449.

[4] A.-M. Tsimberidou. "Targeted therapy in cancer". *Cancer chemotherapy and pharmacology* 76.6 (2015). Publisher: Springer, pp. 1113–1132.

[5] L. Maigne, A. Delsol, G. Fois, et al. "CPOP: An open source C++ cell POPulation modeler for radiation biology applications". en. *Physica Medica* 89 (Sept. 2021), pp. 41–50.

[6] J. B. Pasternack, J. D. Domogauer, A. Khullar, et al. "The advantage of antibody cocktails for targeted alpha therapy depends on specific activity". *Journal of Nuclear Medicine* 55.12 (2014). Publisher: Soc Nuclear Med, pp. 2012–2019.

DINo: Deep learning Intelligence for Nuclear reactions

Lévana Gesson[12], Greg Henning[1], Uli Weber[2], Marie Vanstalle[1]

[1]Université de Strasbourg, CNRS, IPHC UMR 7178, F-67000 Strasbourg, France
[2]Biophysiks, GSI-Fair, Darmstadt, Germany

Abstract In radiation therapy, precise dose calculations are crucial for treatment plan improvement, relying on advanced algorithms based on Monte Carlo simulations. Persistent discrepancies between experimental data and hadronic models in particle therapy highlight the need for improved accuracy. This study introduces DINo (Deep learning Intelligence for Nuclear reactions), a deep learning algorithm designed to explore solutions for these challenges. DINo uses partial proton nuclear reactions cross-sections to extrapolate total cross-sections, mitigating limitations in existing models. Focused on protons, with an extensive representation in the TENDL nuclear data library, DINo, as a dense neural network, learns from TENDL's 2019 model values, effectively predicting total cross-sections for unexplored initial energies. Preliminary results tend to indicate DINo's superior predictive capabilities compared to the TENDL model, promising enhanced precision in nuclear reaction data and refined dose calculations for particle therapy.

1 Introduction

Dose calculations in radiation therapy heavily rely on advanced algorithms incorporating both physical and biological processes, based on data provided by Monte Carlo simulations. However, discrepancies persist between experimental data and hadronic models, particularly in nuclear reactions in particle therapy, leading to potential inaccuracies in dose calculations. Therefore, the implementation of measured data into existing codes is challenging but essential [6]. This abstract proposes the development of a deep learning algorithm called DINo (Deep learning Intelligence for Nuclear reactions) that proposes to leverage partial proton nuclear reactions cross-sections to extrapolate total cross-sections data, mitigating the data limitations within current models.

2 Materials and Methods

The TENDL nuclear data library [1], based on TALYS theoretical code, includes the state of the art in nuclear reaction models, taking into account all major reaction mechanisms encountered in light particle-induced nuclear reactions. The library encompasses a great amount of proton data, leading to the choice of protons as the focal point for this study. Figure 1 illustrates the cross-sections for incident protons on a ^{12}C target, given by the TENDL library, focusing on the comparison between the TENDL model results with experimental data from EXFOR [7] for the p(^{12}C,X)^{11}C reaction alongside with the total cross-section.

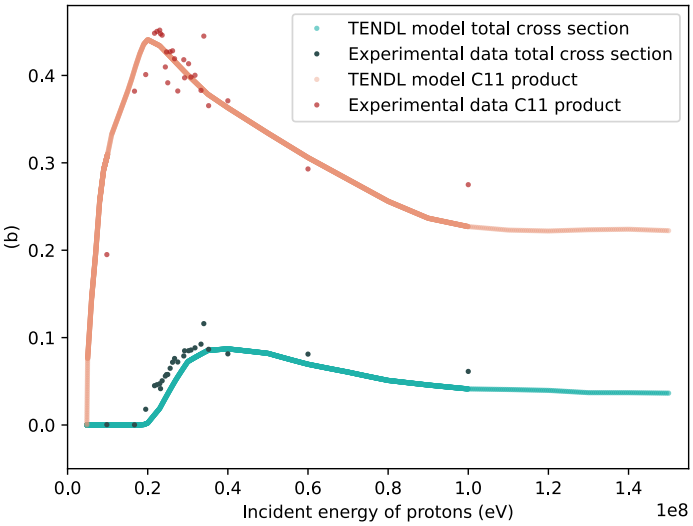


Figure 1: ^{11}C product cross-sections and total cross-section for incident protons on a ^{12}C target – TENDL model results compared to experimental data (data error in the order of 2% not visible on the plot).

The $\chi^2 = \sum \frac{(predicted-expected)^2}{expected}$ value for the comparison between TENDL model and experimental data for total cross-section is 4.5±1.0.

Introduced as a dense neural network (DNN), the DINo algorithm learn from the TENDL 2019 model values, focusing on charge changing cross-section of ^{11}C as a function of proton projectile initial energy, alongside with their corresponding total cross-section. DINo aims at predicting total cross-sections for initial energies not covered by current experimental data. On figure 3 is shown a simplified scheme of the algorithm global structure.

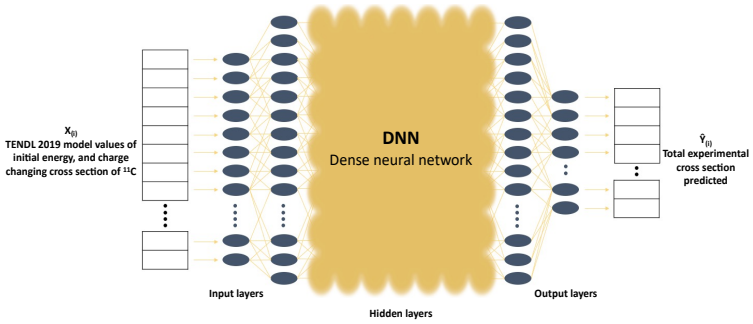


Figure 3: DINo simplified structure

During the algorithm learning phase, 70% of the TENDL model values of ^{11}C charge-changing cross-sections, alongside with their corresponding initial energy, serve as input data. Simultaneously, the total cross-section values, linked to the respective initial energies, are designated as output. To evaluate the algorithm's proficiency, 30% of the TENDL model values are reserved for testing at each epoch. Following 60 epochs, the algorithm is ready for application to experimental data.
During the algorithm testing phase on data, it receives experimental values for ^{11}C charge-changing cross-sections with associated initial energy, enabling it to predict the corresponding experimental values for total cross-sections at those specific initial energies.

3 Results

Figure 4 presents predicted experimental total cross-sections values by DINo alongside TENDL model values and actual experimental data.

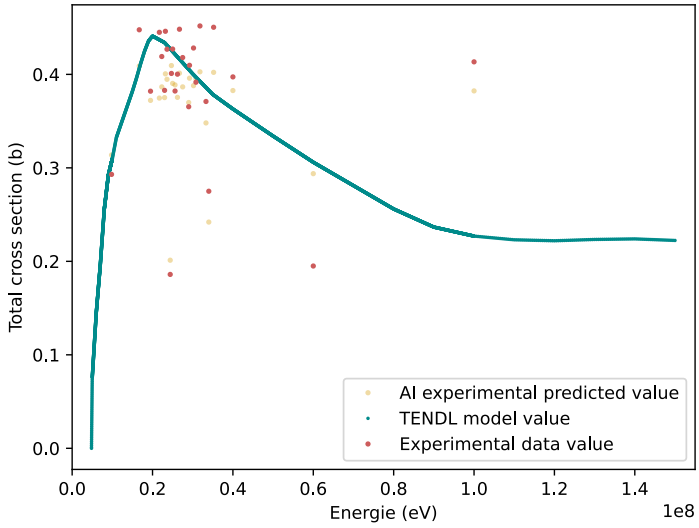


Figure 4: Total cross-section for incident protons on a ^{12}C target as a function of incident energy: TENDL model results and DINo prediction compared to experimental data (data error in the order of 2% not visible on the plot).

The comparison in Figure 5 highlights the efficacy of DINo predictions in relation to both TENDL model results and experimental data. The χ^2 values demonstrate a superior compatibility between the experimental data and the DINo algorithm predictions compared to the TENDL model.

Indeed, the χ^2 between the experimental data and the TENDL model is, like said in previous section, 4.5 ± 1.0 when the χ^2 between the experimental data and the DINo algorithm prediction is 2.9 ± 0.6.

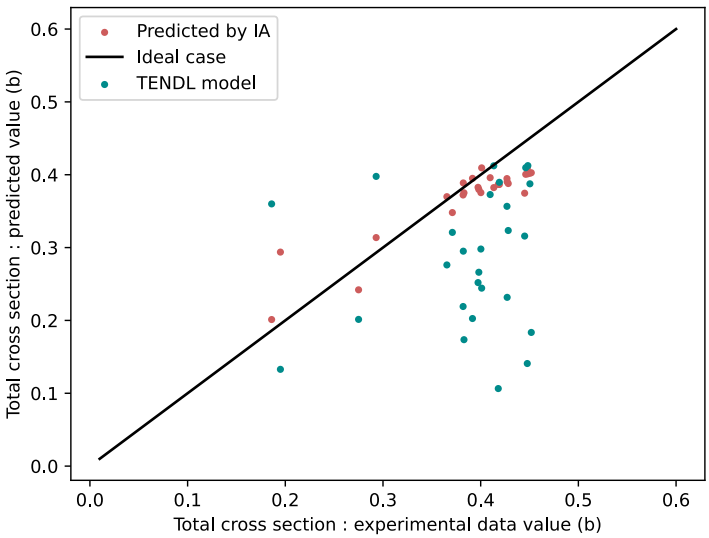


Figure 5: ^{11}C product cross-sections and total cross-section for incident protons on a ^{12}C target: TENDL model results and DINo prediction compared to experimental data (data error in the order of 2% not visible on the plot).

4 Discussion

In this study, the deep learning algorithm DINo exhibits enhanced predictive capabilities for total cross-sections compared to the TENDL model. Even in the worst case scenario, DINo's prediction align with TENDL's performance. Indeed, the χ^2 of 3.5 is reached as the best TENDL's performance and the worst DINo's performance. In optimal situations, DINo outperforms the TENDL model by a factor 2.4.

5 Conclusion

These findings underscore the potential of the deep learning approach in improving the accuracy of predicting nuclear reaction data, in order to to enhance the precision of dose calculations in particle therapy.

To establish the efficacy and reliability of the proposed algorithm in predicting missing experimental data, further investigations on additional charge-changing cross-sections and different target materials are needed. This study lays the foundation for future research aimed at expanding the scope and applicability of the deep learning algorithm in the realm of nuclear reactions.

References

[1] A.J. Koning, D. Rochman, J.-Ch. Sublet, N. Dzysiuk, M. Fleming, S. van der Marck,
TENDL: Complete Nuclear Data Library for Innovative Nuclear Science and Technology,
Nuclear Data Sheets, Volume 155, 2019, Pages 1-55,
ISSN 0090-3752,
https://doi.org/10.1016/j.nds.2019.01.002.

[2] Van Rossum, G. & Drake, F.L., 2009. Python
3 Reference Manual, Scotts Valley, CA: CreateSpace.
[3] TensorFlow Developers. (2023). TensorFlow (v2.15.0). Zenodo. https://doi.org/10.5281/zenodo.10126399

[4] 2022-2023 FIDLE Formation – MIAI, UGA Grenoble, EFELIA project, CNRS
https://gricad-gitlab.univ-grenoble-alpes.fr/talks/fidle/-/wikis/home

[5] "How to produce accurate inelastic cross sections from an indirect measurement method?", by Kerveno, Maëlle and Henning, Greg and Borcea, Catalin and Dessagne, Philippe and Dupuis, Marc and Hilaire, Stéphane and Negret, Alexandru and Nyman, Markus and Olacel, Adina and Party, Eliot and Plompen, Arjan in EPJ N - Nuclear Sciences & Technologies 4, (2018). 10.1051/epjn/2018020
https://hal.archives-ouvertes.fr/hal-02109918

[6] Norbury JW, Battistoni G, Besuglow J, et al. Are Further Cross Section Measurements Necessary for Space Radiation Protection or Ion Therapy Applications? Helium Projectiles. Front. Phys. 2020;8:565954.

[7] N. Otuka, E. Dupont, V. Semkova, B. Pritychenko, A.I. Blokhin, M. Aikawa, S. Babykina, M. Bossant, G. Chen, S. Dunaeva, R.A. Forrest, T. Fukahori, N. Furutachi, S. Ganesan, Z. Ge, O.O. Gritzay, M. Herman, S. Hlavač, K. Katō, B. Lalremruata, Y.O. Lee, A. Makinaga, K. Matsumoto, M. Mikhaylyukova, G. Pikulina, V.G. Pronyaev, A. Saxena, O. Schwerer, S.P. Simakov, N. Soppera, R. Suzuki, S. Takács, X. Tao, S. Taova, F. Tárkányi, V.V. Varlamov, J. Wang, S.C. Yang, V. Zerkin, Y. Zhuang,
Towards a More Complete and Accurate Experimental Nuclear Reaction Data Library (EXFOR): International Collaboration Between Nuclear Reaction Data Centres (NRDC), Nuclear Data Sheets, Volume 120, 2014, Pages 272-276, ISSN 0090-3752, https://doi.org/10.1016/j.nds.2014.07.065.

Development of the low dose radiotherapy machine for the osteoarthritis: optimization of design based on the Monte Carlo simulation

Hyojun Park[1], Jung-In Kim[1], Sang Hoon Heo[2], Song Yi Heo[2], Minjae Choi[1], Ju Yeol Shin[1], and Chang Heon Choi[1,*]

[1]Department of Radiation Oncology, Seoul National University Hospital, Seoul, Republic of Korea
[2] Papricalab Co., Ltd, Seoul, Republic of Korea

Abstract This study aims to optimize the design of the low dose radiotherapy machine for the osteoarthritis to develop the dedicated machine. Monte Carlo (MC) simulation was employed in the optimization by using the GEometry ANd Tracking 4 (Geant4). Firstly, the appropriate source for the osteoarthritis treatment was defined by comparing physical characteristics of different radionuclides currently used in radiotherapy. Next, geometry of the collimator, housing and source were optimized. Geometry of the collimator and housing was optimized in terms of its thickness by evaluating transmission rate of photons from the source according to different tungsten thickness. The source geometry was optimized in terms of its length and diameter as it had cylindrical shape. Finally, the dose distribution produced by the dedicated machine was evaluated according to its rotation steps. The results is expected to contribute development of compact and simple treatment machine for the osteoarthritis.

1 Introduction

Osteoarthritis is one of the most common disease in those over 65 in Korea with incidence rate of approximately 37.8% [1]. Typically, both surgical and conservative treatment have been employed for the osteoarthritis. However, these treatment methods have been reported to have limitations and side effects, for example, cardiovascular damage and increment of gastro-intestinal hemorrhage due to usage of nonsteroidal anti-inflammatory drugs [2, 3].

Low dose radiotherapy (LDRT) was reported as one of the effective methods to cure the osteoarthritis with 63-90% reduction of its symptom [4]. The LDRT have been employed for the osteoarthritis in Europe since early 1900s and over 85% of the therapeutic facilities carry out the LDRT in countries such as German [5]. Meanwhile, the LDRT has never been employed to cure the osteoarthritis in Korea even considered as a treatment option for the osteoarthritis.

Currently, the dedicated RT machine is not available in hospital. The LDRT for the osteoarthritis is carried out by using medical linear accelerators (LINACs), which is inefficient because the LINAC mainly targets cancer patients and its cost is very expensive to buy new machine just for treating the osteoarthritis. Therefore, the LDRT machine for the osteoarthritis with appropriate structure and reasonable price is required.

The aim of this study was to define the optimal design of the dedicated RT machine to develop the clinically available machine. Monte Carlo (MC) simulation was carried out to decide the optimal design of the RT machine. Subsequently, dose distributions in cylindrical water phantom, which mimicked knee joint, were assessed according to different beam delivery technique.

2 Materials and Methods

The design optimization was carried out by using the GEometry ANd Tracking 4 (Geant4) tool-kit, which was written in C++ language with high flexibility and powerful calculation of particle transportation. The optimization was carried out for the radiation source, collimator, and housing of the treatment machine. Figure 1 shows illustration of the dedicated RT machine. The machine delivers therapeutic beams from different beam directions by rotating the radiation source around patient's leg located at rotation axis of the RT machine.

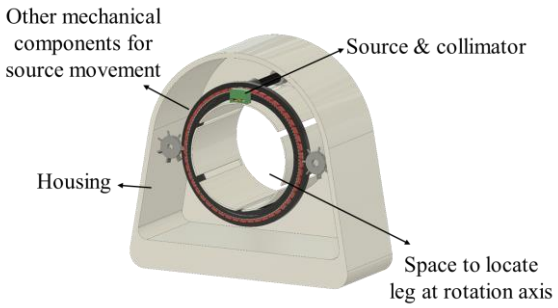


Figure 1. Illustration of the LDRT machine for the osteoarthritis

The RT machine was designed to use radionuclide as the radiation source. The optimization started from decision of which type of radionuclide would be appropriate for treatment of the osteoarthritis. Evaluation was carried out on physical characteristics of radionuclides being commonly used in current radiotherapy, such as gamma knife, brachytherapy, and etc. The energy of the gamma ray decides the available depth of the treatment. The half-life refers replacement period of the source. The specific activity is related to available dose rate of the treatment machine. The source was defined with consideration of the requirements for the LDRT of the osteoarthritis.

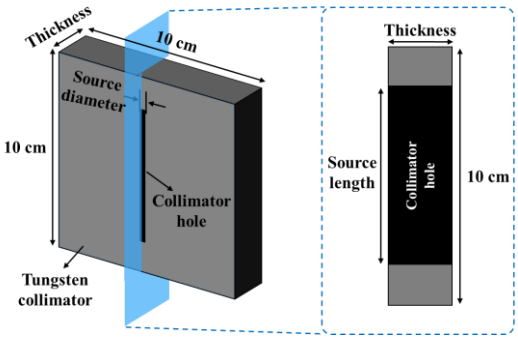


Figure 2. Design of the tungsten collimator

The collimator and housing consisted of tungsten (ρ = 19.3 g/cm^3). The collimator was designed to have parallel hole, which minimize the beam divergence due to isotropic emission of the gamma ray from the radionuclide (Figure 2). It was important because sequential beam delivery is required to cover the treatment site. The fall-off region of each field should be steep as possible to avoid dose overlap at field junction, especially at surface of patient. Size of the collimator hole was defined as that of the source.

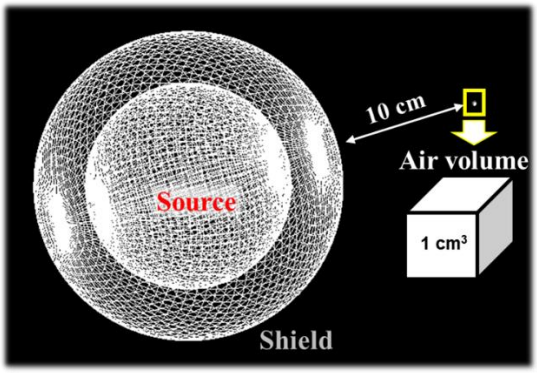


Figure 3. Monte Carlo simulation to calculate gamma transmission according to tungsten thickness

Thickness of the collimator and housing was optimized to stop at least 95% of the gamma ray from the source. The MC simulation was employed to define the optimal thickness of the collimator (Figure 3). Transmission of the gamma ray was calculated according to different thickness of tungsten, which varied from 1 to 4 cm with 0.5 cm steps. Total of 8×10^7 photons were generated at each simulation. Fluence of the gamma ray was assessed in the air volume that located at 10 cm from the tungsten surface. The results were compared to that calculated mathematically based on the equation below:

$$I = I_0 e^{-\mu t}$$

Shape of the source was determined as cylinder. To decide the optimal source geometry, dose distributions in the knee phantom were simulated according to source diameters and lengths. Diameter of the source was varied as 1, 5, 10, and 20 mm. Nominates for source length were 10, 27, and 50 mm where 27 mm was the length of the radiation source embedded in the MR/^{60}Co machine manufactured by ViewRay®.

Figure 4 shows the simulation setup for calculating dose distribution in the knee phantom. The source was attached to the collimator while the beam angle was fixed at 0. The distance between the source and center of the phantom was 27.5 cm, which refers the source-to-surface of the phantom of 15 cm. The number of generated photon in each simulation was 8×10^7. The knee phantom was modelled to simulate dose distribution in patients given the LDRT for the osteoarthritis (Figure 4). The phantom was filled with water. It had cylindrical shape with diameter and height of 25 and 20 cm, respectively. The diameter of the phantom was decided by considering Korean standard of knee width (16 cm) with its variation, and compensation to produce uniform dose distribution in the entire treatment site. The length was decided as it describes stretched leg.

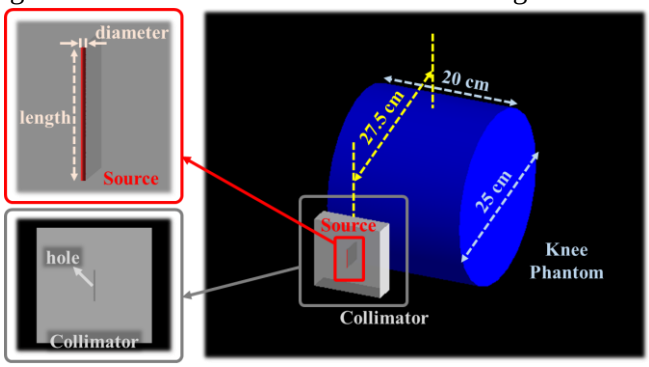


Figure 4. Monte Carlo simulation for calculating dose distribution in the knee phantom

Finally, the dose distributions were calculated according to different intervals of beam angles consisting a single arc. The interval was 60, 45, 40, 30, 20 and 10 degrees. Simulation condition was identical to that represented in Figure 4 except that the optimized source rotated around the phantom.

3 Results

Table 1 shows physical characteristics of radionuclides used in the radiation treatment. Among the radionuclides, ^{125}I and ^{103}Pd are not appropriate for the osteoarthritis treatment since the photons cannot reach the treatment site at center of leg joint. In addition, ^{103}Pd has too short half-life that requires frequent source replacement, which can make the performance maintenance of the RT machine be challenging. Meanwhile, ^{137}Cs has much lower specific activity than other radionuclides, which can lead to lower treatment efficiency due to low beam output. Therefore, ^{60}Co was defined as the radiation source of the RT machine with better stability of activity for long period and availability to dose delivery to deeper region than ^{192}Ir. Despite that ^{192}Ir has higher specific activity than ^{60}Co, ^{60}Co has sufficient specific activity to be used in LDRT and even larger specific activity requires thicker radiation shielding due to unnecessarily high dose rate.

Table 1. Physical characteristics of radionuclides used in the radiation treatment

Source	Energy (keV)	Half life	Specific activity (Ci/g)
^{137}Cs	661.6	30.05 yr	86.49
^{60}Co	1173, 1331	5.2 yr	1,130
^{192}Ir	310 - 380	74 d	9,220

	(average)		
^{125}I	28 (average)	60 d	17,000
^{103}Pd	21 (average)	16.99 d	> 67,600

Figure 5 shows the transmission of the ^{60}Co gamma ray according to different tungsten thickness with comparison between the MC simulation and mathematical calculation. Both results decreased exponentially as tungsten thickness increased. The difference between the simulation and mathematical calculation increased with larger tungsten thickness due to the contribution of the scatter, which was excluded in the mathematical calculation. In the results, tungsten with thickness larger than 3 cm could stop 95% of the gamma ray. This implies that thickness of the collimator and housing needs to be larger than 3 cm.

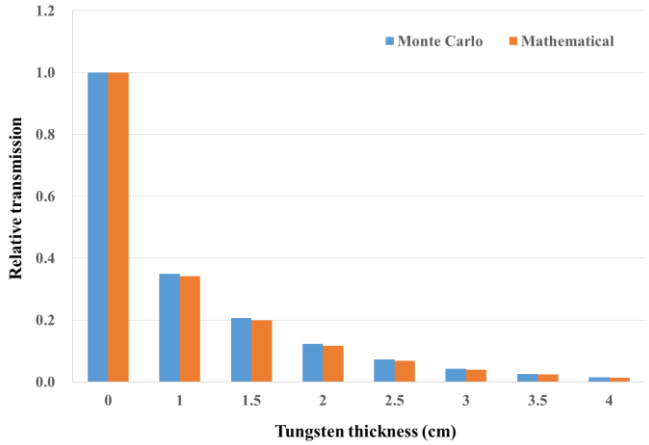


Figure 5. Transmission rate of ^{60}Co gamma ray according to tungsten thicknesses with comparison between the Monte Carlo simulation and mathematical calculation

The dose distributions according to different source diameter and length were shown in Figure 6. The diameter was fixed at 1 mm in the investigation of effect of the source length on dose distribution. The dose distribution was obtained on the transverse plane at center of the phantom. The source produced broader dose distribution with length of 27 and 50 mm than 10 mm. Since larger source has higher activity that requires more bulky shielding, the source length was decided as 27 mm.

In the investigation of dose changes because of different source diameter, the source length was fixed at 27 mm. The dose distribution was obtained on the coronal plane at center of the phantom. Larger source diameter produced broader dose distribution with increased penumbra. This can lead to enhancement of hot or cold spots at the field junction when the beam scans patient's leg along the rotation axis. The beam divergence along superior-to-inferior direction of the phantom should be minimized. Therefore, the diameter of the source was defined as 1 mm.

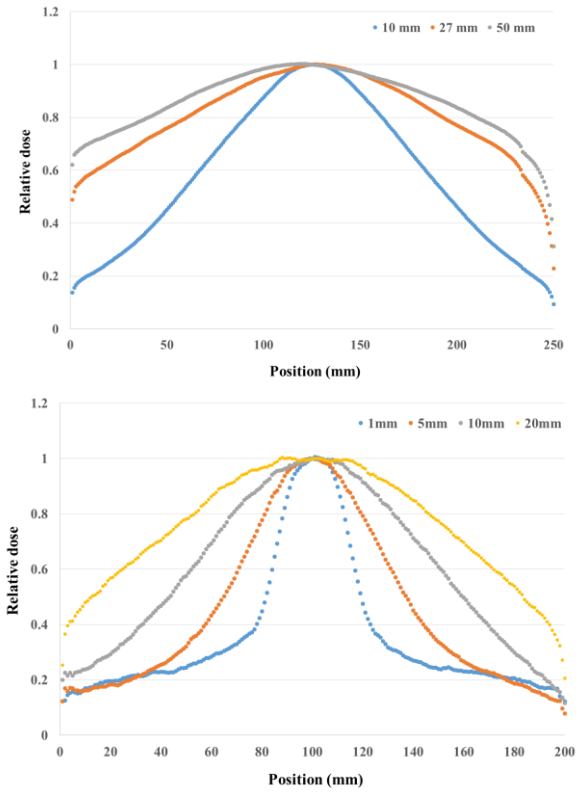


Figure 6. Dose distributions according to different source length (upper) and diameter (lower). Note that, the dose distributions with different source length and diameter were obtained on the transverse and coronal plane of the phantom, respectively.

As shown in Figure 7 and 8, the dose distributions were compared among the beam delivery techniques with different beam angle interval consisting a single arc. The lateral dose of 60-degree interval was higher than the other when the profile was obtained along vertical direction. Meanwhile, it was the lowest in the horizontal profile. This was due to sparse beam delivery compared to the other beam delivery conditions. However, all of the results showed that the region with diameter of 20 cm at the center was given the dose with the variation less than 10%. In addition, the dose distribution was almost identical when the beam angle interval was larger than 45 degrees.

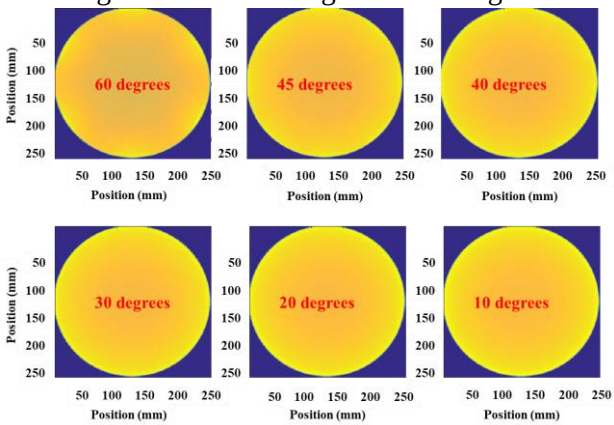


Figure 7. Dose distributions according to different beam angle intervals

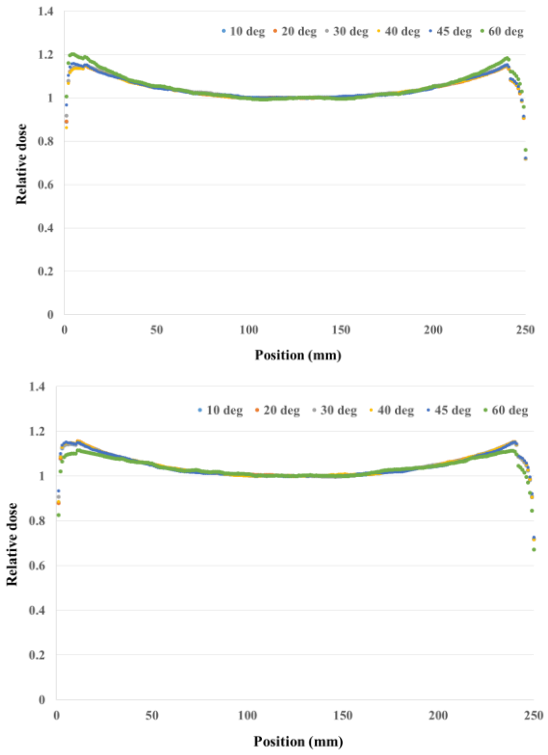


Figure 8. Vertical (upper) and horizontal (lower) profiles of the dose distribution in Figure 7

4 Discussion

In this study, investigation for the optimal design of the LDRT machine for the osteoarthritis was carried out. The machine consisted of the radiation source, the collimator, and the housing. The machine showed uniform dose distribution in broad region, which can sufficiently cover patient's leg.

As the radiation source, radionuclide was selected rather than the x-ray generator. The machine was designed to be compact and simple. The x-ray generator requires additional space for producing accelerated electrons, which can make the machine more bulky than using radionuclides even required x-ray energy is hundreds keV. In addition, radionuclide as the radiation source does not require mechanical maintenance. In the case of radionuclide with long half-life, such as ^{60}Co in this study, the dose output can be stable for more than year.

The dose distributions according to the source lengths showed acceptable tendencies. The difference was smaller between 27 and 50 mm than 10 and 27 mm. It was due to the cylindrical shape of the phantom. Depth of water was insufficient for build-up of the gamma rays reaches border of the phantom. The dose distribution was broader with larger source diameter. However, the width of the dose distribution was much larger than that of the source. Even parallel-hole collimator produced diverged beam due to the clearance between the collimator and patient. Since the dose output can be decreased as the collimator is located closer to patient, further investigation on the optimal distance between the collimator and patient is required.

The dose distribution produced by the arc beam delivery showed possibility of the dedicated RT machine to be used in the clinic. Even the beam was discretely delivered at different directions but produced uniform dose distribution with the variation less than 10%. Since the radiation treatment of the osteoarthritis requires uniform dose in large treatment site, the results proved availability of the RT machine. In addition, the results also showed that uniform dose can be delivered by continuous rotation of the radiation source with fixed speed. Further, relation between the rotation speed and dose per a single arc should be investigated to apply the dedicated machine in the clinic.

Current study focused on design of the dedicated machine based on the MC simulation. Study on application of the machine to actual treatment plan is required based on both the simulation and measurement. Algorithm for establishing treatment plan of the patients undergoing the osteoarthritis is also necessary. For clinical application of the dedicated RT machine, investigations on safety-related managements, such as evaluation of leakage dose, are also required. Therefore, further studies will be carried out to develop the LDRT machine for the osteoarthritis and to apply the machine in the actual treatment.

5 Conclusion

In this study, the design of the dedicated RT machine for the osteoarthritis was defined and its availability on clinic was evaluated. It is expected that the machine can be applied to the actual treatment by manufacturing its hardware and developing its treatment planning algorithm. Further, the machine can contribute improved treatment quality of the osteoarthritis.

References

[1] Kim, Jung-Suk, Chun-Ja Kim, and Elizabeth A. Schlenk. "Psychometric properties of the Korean version of the patient knowledge questionnaire-osteoarthritis." Clinical Nursing Research 31.1 (2022), pp. 69-79.

[2] Bhala, N., et al. "Vascular and upper gastrointestinal effects of non-steroidal anti-inflammatory drugs: meta-analyses of individual participant data from randomised trials." Lancet (London, England) 382.9894 (2013), pp. 769-779.

[3] Masclee, Gwen MC, et al. "Risk of upper gastrointestinal bleeding from different drug combinations." Gastroenterology 147.4 (2014), pp. 784-792.

[4] Dove, Austin PH, et al. "The use of low-dose radiation therapy in osteoarthritis: a review." International Journal of Radiation Oncology* Biology* Physics 114.2 (2022), pp. 203-220.

[5] Leer, Jan Willem Hendrik, Paul Van Houtte, and Jaap Davelaar. "Indications and treatment schedules for irradiation of benign diseases: a survey." Radiotherapy and oncology 48.3 (1998), pp. 249-257..

Prompt-gamma timing with a track-length estimator in proton therapy

Jean M Létang[1], Oreste Allegrini[2], and Étienne Testa[2]

[1]Univ. Lyon, INSA-Lyon, Université Lyon 1, UJM-Saint Étienne, Centre Léon Bérard, CREATIS CNRS UMR 5220 Inserm U1294, 69373 Lyon, France
[2]Univ. Lyon, Université Lyon 1, IP2I CNRS/IN2P3 UMR5822, 69622 Villeurbanne, France

Abstract We present a variance reduction technique (VRT) for Geant4/GATE Monte Carlo (MC) simulations, tailored to the simulation of prompt-gamma (PG) radiation caused by proton inelastic processes in the patient. This technique is based on a track-length estimator to generate 3D sources of PG energy and emission-time yields. Results show minor discrepancies with respect to analog MC. A gain in efficiency of about 50 is reported for the detection of PG.

1 Introduction

Compared with standard radiotherapy, monitoring the dose delivered in proton therapy is much more problematic because no primary particles leave the patient. On-line monitoring of secondary radiation such as prompt-gammas (PG) emitted after nuclear inelastic processes is one of the techniques being developed to ensure that the dose delivered complies with the treatment plan. The design and optimization of the detector setup require multiple Monte Carlo (MC) simulations, and the use of variance reduction techniques is highly desirable. An estimator based on track length was originally proposed to build 3D sources of PG energy yields [1, 2]. It has recently been adapted to compute 3D sources of PG emission-time yields, in addition to PG energy yields. We present a benchmark of this module named vpgTLE-tt against analog MC in the Geant4/GATE simulation framework.

2 Materials and Methods

2.1 Description of the vpgTLE-tt module

The vpgTLE-tt module consists of three stages as depicted in Figure 1. Stage 0 is run off-line to build the database of PG energy yields (indexed by proton energy E_p) for a set of elements characterized by their atomic number Z. Stage 1 is then set according to the scene to be studied. The outputs of Stage 1 are two 3D voxelized sources (assumed to be uncorrelated) of PG energy $\boldsymbol{E}_{pg}$ and emission-time $\boldsymbol{t}_{pg}$ yields: for each step of the proton track in a given voxel, the PG energy yield is read (from Stage 0 corresponding to the voxel material and the proton energy) and scored in an histogram with a weight corresponding to the track length. Likewise, the PG emission time is randomly drawn between the pre- and post-step points and scored in an histogram with a weight corresponding to the track length. In Stage 2, the voxelized yield sources $\boldsymbol{E}_{pg}$ and $\boldsymbol{t}_{pg}$ are then read and a number of PG events is run according to the total PG yield (integral of $\boldsymbol{E}_{pg}$) and the requested number of proton N_p.

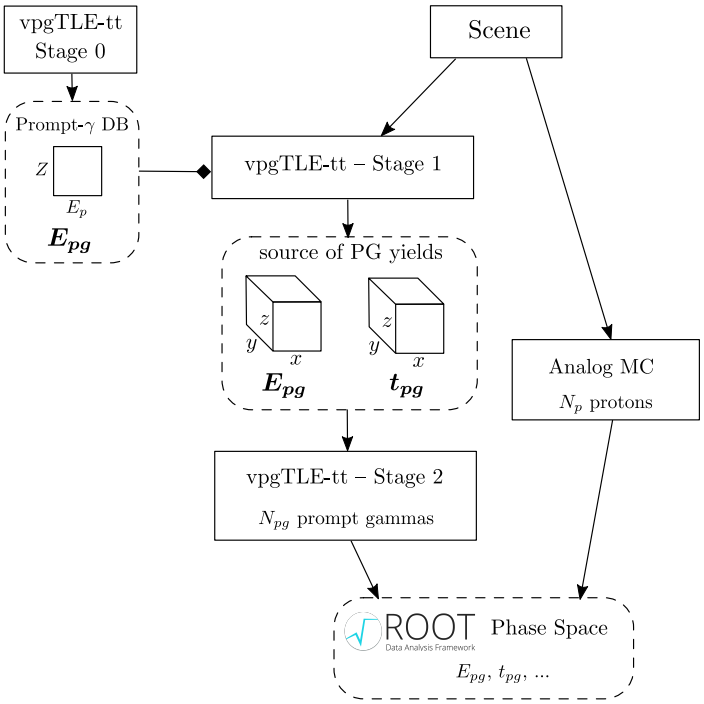


Figure 1: Diagram of the vpgTLE-tt module (see description in section 2.1).

2.2 Test case

A 3D Geant4/GATE view from the clinical Radiation-Therapy plan (RT-plan) is illustrated Figure 2. Only the 7 spots of the distal layer are considered, the proton energy is 133.08 MeV. The Geant4 physics list QGSP_BIC_HP_EMY has been used with 1 mm cuts for all particles. A statistics of 10^8 proton events has been used, which gives a statistics of 1.2×10^7 emitted prompt-gamma events. The ROOT scoring volume is a sphere shell (55 cm radius) around the patient head. The creator processes of the detected gammas in the analog MC simulation are first investigated to estimate the relative importance of the gammas coming from proton inelastic processes. The 2D gamma-yield distributions of the emission-time vs energy are therefore saved for the major creator processes from the ROOT output file. A benchmark study is finally conducted to assess the accuracy of the vpgTLE-tt module against analog MC. The relative error distributions correspond to the difference of the 2D gamma yield distributions divided by the respective analog maximum PG yields.

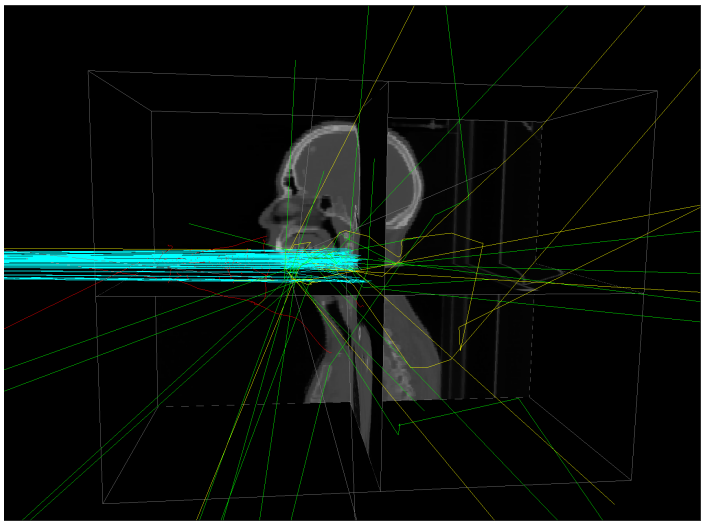


Figure 2: 3D scene view (with 100 incident protons) of the clinical RT-plan test case with 100 protons with Geant4/GATE. Protons are in cyan, gammas in green, neutron in yellow, electrons in red and positrons in blue.

3 Results

3.1 Creator processes of the detected gammas

The 2D distributions (emission-time vs energy) for several creator processes of the detected gammas in the scoring volume (sphere shell) are shown in Figure 3. Among all processes, the major six (in descending order) are: proton inelastic, annihilation, neutron inelastic, neutron capture, radioactive decay, and e-/e+ bremsstrahlung. Note that annihilation photons other than 511 keV have been either scattered in the patient (below 511 keV) or comes from positrons which are not at rest (above and below 511 keV). Table 1 lists the relative importance of all these processes.

process	counts	%
proton inelastic	1.17×10^7	53%
annihilation	7.14×10^6	32%
neutron inelastic	1.57×10^6	7.1%
neutron capture	7.94×10^5	3.6%
radioactive decay	6.96×10^5	3.2%
e-/e+ bremsstrahlung	8.96×10^4	0.4%

Table 1: Creator processes of detected gamma rays (any time). See the corresponding 2D histograms in Figure 3.

3.2 Benchmark: vpgTLE against analog MC

The 2D distribution of the detected gamma yields given by the vpgTLE-tt module is depicted at the top of Figure 4. The corresponding distribution given by the analog MC simulation is shown in the second row of Figure 4 for all creator processes (left column) and only proton inelastic (right column). The relative error of the vpgTLE-tt module with respect to each of these analog outputs are shown in the bottom row of Figure 4.

4 Discussion

The prompt-gammas (PG) resulting from proton inelastic processes represent the majority of all gammas detected (any time) by the scoring sphere shell. In the $[4,6]$ ns time window and above 1 MeV, the only non-negligible contribution comes from neutron inelastic processes, which is nevertheless an order of magnitude below proton inelastic. This confirms our decision to restrict to proton inelastic processes.
In the present study, the efficiency gain of Stage 2 compared to analog MC is about 50: the variance is the same since a comparable PG statistics is simulated, the gain just comes from the computation time gain. Huisman et al. [2] reported efficiency gain for Stage 1 of the order of 10^3 to reach the same variance at the voxel level to compute the PG source yields. The efficiency of the vpgTHE-tt method (Stages 1 and 2) should therefore only be marginally affected by the on-line simulation of protons by Stage 1.
The relative error remains within $\pm 20\%$, and even then this is for very localized places (in energy and emission time), most of the other regions have a relative error of only a few percent. The main sources of these small discrepancies come from (i) the de-excitation times of some nuclear products which are much greater than the nanosecond (and they are not modeled by vpgTLE-tt) and (ii) for some PG, energy and emission time are correlated (in the Bragg peak area).

5 Conclusion

We have presented and validated a new acceleration module based of on a track-length estimator to compute voxelized prompt-gamma yields in energy and emission-time. Further work is needed to handled gamma lines with long de-excitation time.

Acknowledgments

This work was supported by the LABEX PRIMES (ANR-11-LABX-0063). We gratefully acknowledge support from the CNRS/IN2P3 Computing Center in Lyon (France).

References

[1] W. El Kanawati, J. M. Létang, D. Dauvergne, et al. “Monte Carlo simulation of prompt γ-ray emission in proton therapy using a specific track length estimator”. *Physics in Medicine and Biology* 60.20 (2015), pp. 8067–8086. DOI: 10.1088/0031-9155/60/20/8067.

[2] B. F. B. Huisman, J. M. Létang, É. Testa, et al. “Accelerated prompt gamma estimation for clinical proton therapy simulations”. *Physics in Medicine and Biology* 61.21 (2016), pp. 7725–7743. DOI: 10.1088/0031-9155/61/21/7725.

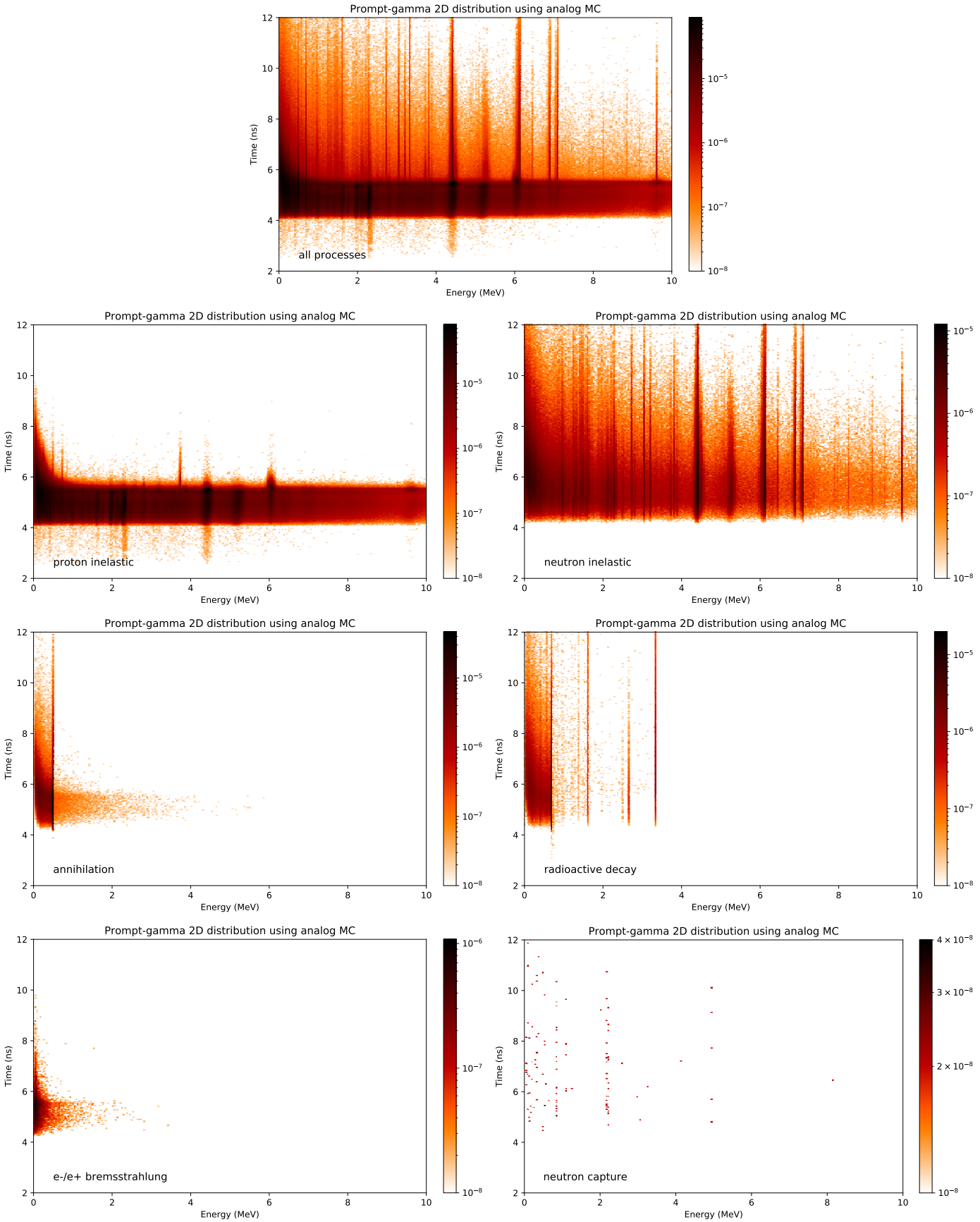


Figure 3: 2D histograms of prompt-gamma rays per creator process. Total counts of gamma rays per process are given in Table 1. Top row: all creator processes. Then in the reading direction: proton inelastic, neutron inelastic, annihilation, radioactive decay, electron/positron bemsstrahlung, and neutron capture. Caution: the colorbars (log scale) are normalized to their maximum value.

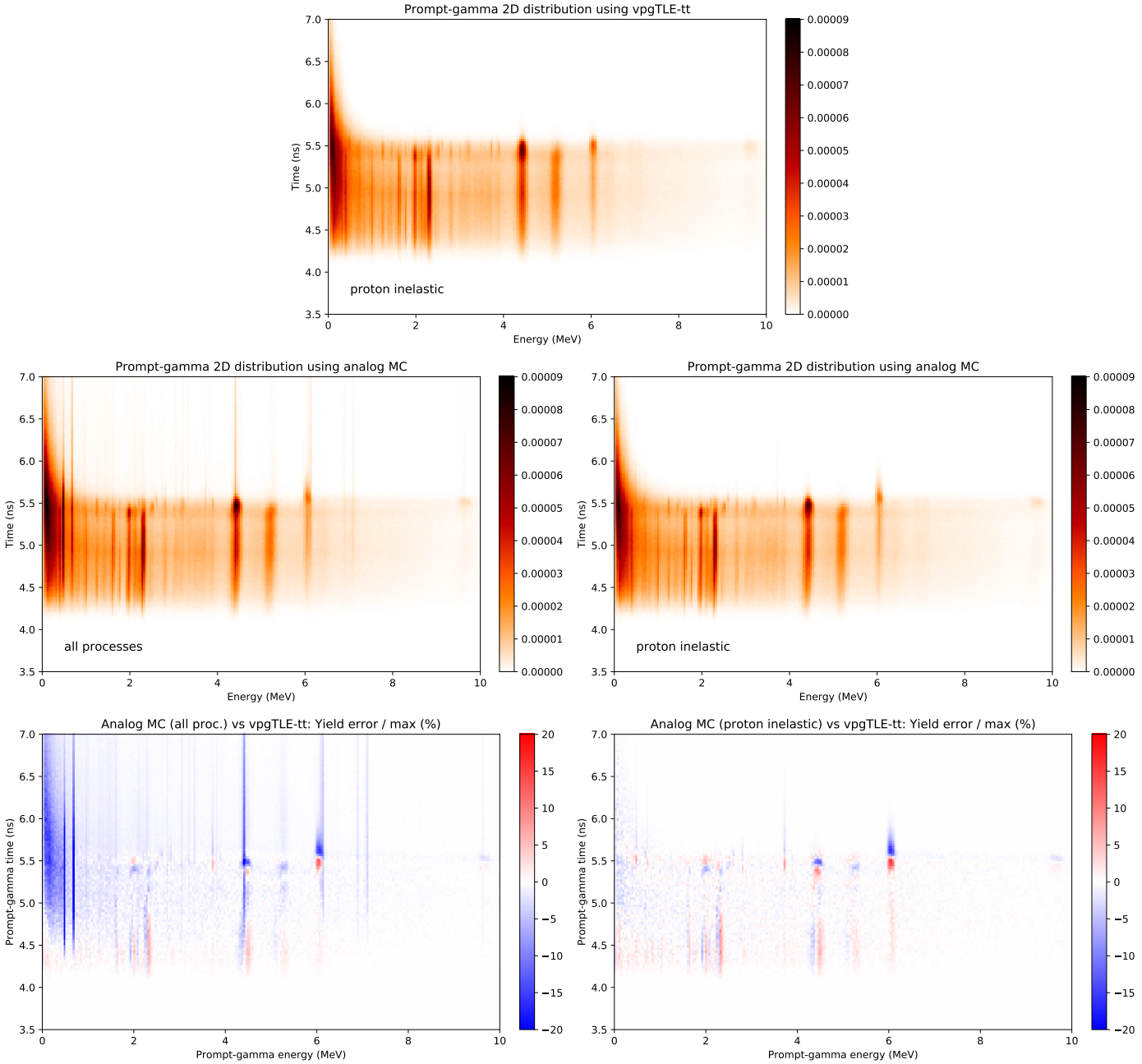


Figure 4: 2D histograms of prompt-gamma yields. Top row: vpgTLE-tt module. Middle row: analog MC with gammas from all processes (left) or from just proton inelastic (right). Bottom row: corresponding relative yield error (normalized by the bin maximum). Caution: the colorbars of the PG 2D distributions are similar (same linear scale) and the error plot with respect to all processes (bottom left) has been clipped to $\pm 20\%$ for better visual comparison.

Identification of suitable probing locations for a proton radiography-based quality control of CBCT-based synthetic-CTs

Måns Lundberg[1,2], Arturs Meijers[2], Antony Lomax[2,3], Damien C Weber[2,4,5], Kevin Souris[6], and Antje Knopf[1]

[1]Institute for Medical Engineering and Medical Informatics, School of Life Science FHNW, Muttenz, Switzerland
[2]Center for Proton Therapy, Paul Scherrer Institute, Villigen, Switzerland
[3] Department of Physics, ETH Zurich, Zurich, Switzerland
[4] Department of Radiation Oncology, University Hospital of Zürich, Zürich, Switzerland.
[5]Department of Radiation Oncology, Inselspital, Bern University Hospital, University of Bern, Bern, Switzerland
[6]Ion Beam Applications SA, Louvain-La-Neuve, Belgium

Abstract

Background

Cone-Beam CT (CBCT)-based adaptive proton therapy is nowadays technically feasible and has the potential to enhance patient outcomes. To use CBCTs for proton dose calculations, synthetic CTs (sCT) need to be generated with sufficient precision in their density information that proton dose calculations can be performed. Proton Radiography (PR) can be employed as a quality control (QC) tool for such sCTs in order to validate their accuracy with respect to prediction of proton ranges. The objective of this study is to identify PR probing locations that are robust towards any confounding factors like patient positioning errors, and therefore can be suitable as sCT QC positions.

Method

A cohort of five patients with cranial CTs was used to predict and evaluate PR probing locations. As a measure for the robustness of the probing locations, the energy bandwidth in a surrounding region was investigated. To assess this, positioning errors were applied to the probing locations, and resulting range shifts were evaluated.

Result

The total number of investigated locations per patient was ranging from 76 to 92. A correlation between the energy bandwidth and range shifts was found. This trend was observed in all five patients and was already apparent for small positioning errors of 1mm. A minimum of 19 suitable probing locations was found in each of the patients with a maximum range error of 0.6% of transverse water equivalent thickness.

Conclusion

This study showed that PR locations suitable for sCT QC can be found by determining and assessing the energy bandwidth in a surrounding area.

1 Introduction

Proton therapy (PT) has the potential to deliver highly conformal dose distributions with a low integral dose, thus sparing healthy tissue. Protons offer a low entrance dose followed by a sharp dose rise and a subsequent dose decline called a Bragg peak. The location of the steep dose gradient is sensitive to changes in the patient anatomy, such as tumor shrinkage, filling of nasal cavities, or weight loss/gain [1].

A typical treatment course of 30 fractions lasts 6 weeks, during which anatomical changes are likely to occur. As the treatment plan is usually based on a single planning CT, safety margins and robust optimization are applied to avoid loss of target coverage in the case of such changes. Resulting in a compromise of target dose conformity. Adaptive Proton Therapy (APT) monitors the dosimetric impact of anatomical change and aims to modify the treatment plan accordingly [2]. With knowledge of anatomical changes, the dose to healthy tissue can be minimized whilst maintaining conformal target dose coverage.

To monitor anatomical changes, frequent medical images of the patient are needed. Frequent medical imaging places a significant imaging dose burden on the patient.

Cone-Beam CT (CBCT) offers low-dose 3D images and is generally available in the treatment room for daily patient positioning.

Whilst suitable for positioning, CBCT's suffer from various artifacts, such as beam hardening and scattering. They therefore have a significantly lower image quality than diagnostic planning-CTs.

To enable accurate proton dose calculations based on CBCT, corrections are therefore needed [3].

For this, information from the initial planning CT and the daily CBCT can be combined to generate synthetic CTs (sCT's), with the aim of providing updated patient anatomy with the image quality of the diagnostic planning CT. Several studies have reported the ability to generate sCT's that are suitable for proton dose calculations [4–6]. Their use in a clinical workflow for adaptive proton therapy however requires an appropriate Quality Control (QC) procedure. Proton Radiography (PR) is a promising approach for this [7, 8]. PR uses proton beams of high enough energy to penetrate the patient such that the residual Bragg peak can be measured at the opposite side of the patient. From the residual range of this measured Bragg peak, the Water-Equivalent Thickness (WET) of the patient at the traversed location can be determined. By comparing PR simulations performed on the sCT with PRs measured after the patients, the HU accuracy and imaging quality of the sCT can therefore be assessed [9].

When using PR to check the accuracy of HU in the sCTs, it is vital however to minimize the influence of any confounding factors. Proton beams are very susceptible to density variations and even patient positioning errors can alter the Bragg peak location. This study aims to locate and predict PR locations that are robust to such changes and would therefore be suitable for sCT QC in a clinical environment. The assessment of suitable locations is based on a calculation of Energy Bandwidths (EB). For five patients with cranial CTs, we calculated optimal PR locations and assessed their robustness by simulating positioning errors and analyzing resulting range shifts.

2 Materials and Methods

2.1 Dataset

This study is based on diagnostic planning CT images in the cranium of five patients.

2.2 Robust PR Probing Locations

A sagittal WET image is generated from the planning CT data with a resolution of 1mm×1mm. The WET values are then translated to proton energy by PSTAR Continuous Slowing Down Approximation (CSDA) in liquid water (https://physics.nist.gov/PhysRefData/Star/Text/PSTAR.html) resulting in an energy map.

For each patient, the outline of the skull was manually drawn as the Region of Interest (ROI) on the energy map, as shown on the left side of fig. 2. Each pixel within the ROI is considered as a potential PR probing location.

As a measure for the robustness of a PR probing location, the EB within an area surrounding the probe location is calculated.

In fig. 1, a sketch of the 17mm×17mm considered area for the EB calculation is shown. The size of the area is well covering the two-sigma extent of a typical clinical proton pencil beam. For the next steps, the EB distribution is truncated between the 1st and 99th percentiles to mitigate the impact of outlier values.

The robustness of a probing location is ranked according to the EB. To avoid clusters of suitable probing locations, each location had to fulfill eq. (1)

$$(AP \geq 5\,\text{mm} \cap IS \geq 5\,\text{mm}) \cup (d \geq 12\,\text{mm}) \qquad (1)$$

where AP corresponds to the distance between two probing locations in Anterior-Posterior direction, IS in Inferior-Superior, and d is the diagonal distance. The locations of the spots in the order of robustness are written into a DICOM file that can be used as input in the simulation environment described below.

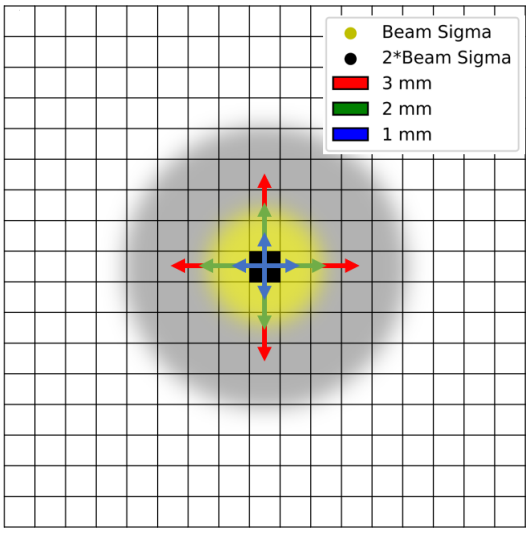


Figure 1: Illustration of the 17mm×17mm box considered for the EB calculation in relation to a typical proton beam sigma. In black, the investigated probing location is shown. The red, green, and blue arrows indicate the positioning errors which are simulated to evaluate the robustness of the probing location with respect to confounding factors.

2.3 Evaluation of PR Probing Locations

To assess the robustness of each probing location with an EB threshold below 10 MeV, the resulting range shifts when applying positioning errors were assessed. Positioning errors of ± 1, 2 and 3mm in the AP and IS directions were applied as illustrated by the blue, green, and red arrows in fig. 1. For each positioning error, an Integrated Depth Dose (IDD) curve for the residual Bragg peak was created by integrating the dose along the beam axis. The IDDs of the nominal and the positioning error simulations were fitted using a least-squares method to determine the resulting range shifts.

2.4 Simulation Environment

PR simulations were performed within the open-source program OpenREGGUI (https://openreggui.org/) with proton beams being simulated to enter from the left of the patient. The residual Bragg peaks were scored on the patient's right side, in a water structure that was included in the patient's planning CT to mimic a PR detector.

3 Results

The number of investigated PR probing locations per patient is listed in table 1, showing a cumulative count of locations with increasing EB. A maximum of 92 PR locations with an EB below 10 MeV was found in patient 1, as shown on the right of fig. 2. The distribution of PR locations appears symmetrical, with low EB locations in the center of the brain and higher EB locations towards the outer regions. Similar distributions were found for all five patients.

Table 1: Cumulative count of PR Locations Found per EB per Patient

EB[MeV]	1	2	3	4	5	6	7	8	9
Pat. 1	0	1	10	26	41	58	66	78	92
Pat. 2	1	2	5	16	21	35	50	57	77
Pat. 3	0	0	0	9	19	36	47	56	76
Pat. 4	3	14	22	32	44	54	64	74	82
Pat. 5	0	0	4	17	25	37	53	60	82

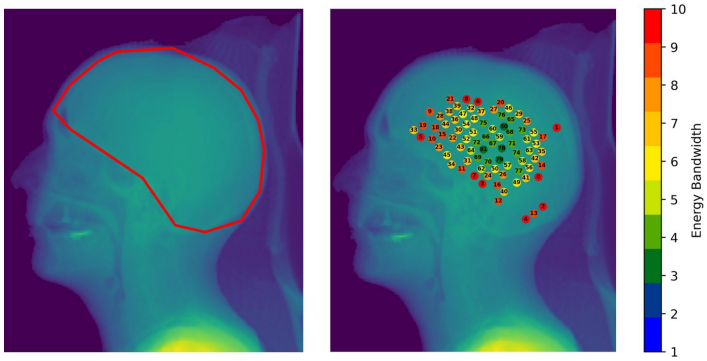


Figure 2: Left: Illustration of the ROI in which PR probing locations were investigated. Right: Set of PR probing locations with an EB below 10 MeV. Both images show Patient 1

When simulating positioning errors, an increase in resulting range shifts can be observed in relation to an increase in EBs, as shown in fig. 3.

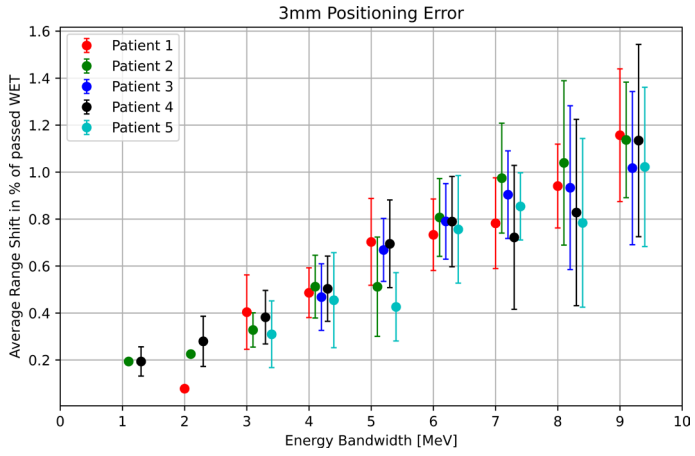


Figure 3: Average range shifts for simulated 3 mm positioning shifts in the AP and IS directions.

This correlation is clearly observable even for smaller positioning errors of 1mm, as shown in table 2, and was apparent in all five investigated patients.
For a positioning error of 3mm, the maximum range shifts in % of the average WET in an area based on two-sigma around the PR location, in relation to the EB is shown in table 3.

Table 2: Average range shift in % of passed WET for 1, 2 and 3 mm positioning errors (PE) for each EB averaged over all five patients

EB MeV	1mm PE	2mm PE	3mm PE
1	0.06	0.12	0.19
2	0.06	0.12	0.19
3	0.12	0.24	0.36
4	0.16	0.32	0.48
5	0.2	0.4	0.6
6	0.26	0.52	0.78
7	0.28	0.56	0.85
8	0.30	0.60	0.90
9	0.36	0.73	1.09

Table 3: Maximum range shifts in % of passed through WET in relation to the EB for positioning error simulations of 3mm.

EB	Patient1	Patient2	Patient3	Patient4	Patient5
1		0.36		0.48	
2	0.10	0.40		0.73	
3	1.03	0.59		0.91	0.72
4	1.21	1.08	0.96	1.15	1.11
5	1.86	0.98	1.26	1.61	1.01
6	1.58	1.60	1.58	1.64	1.68
7	1.50	2.19	1.78	1.82	1.61
8	2.09	2.16	2.04	1.95	1.69
9	2.19	2.67	2.62	2.51	2.34

4 Discussion

PR is a promising tool for the QC of CBCT based sCTs generated for adaptive proton therapy. Due to the sensitivity of protons to density variations however, it is essential to locate PR probing locations that are robust to confounding factors. In this study, we conducted an analysis of potential PR probing locations for five patients. The PR probing locations were selected based on the EB in the areas surrounding proton pencil beams.
The EB was determined by considering an area of 17mm×17mm on the energy map around a potential probing location. The size was chosen considering a typical sigma of proton pencil beams and potential positioning errors.
To evaluate the robustness of the predicted PR probing locations, and to verify that the EB can be used as a metric for robustness, positioning errors of ± 1, 2, and 3mm in the AP and IS directions were applied, and the resulting range shifts analyzed. During treatment planning for head and neck patients, positioning uncertainties up to 3mm are considered [10]. The positioning errors considered for this study therefore reflect the magnitude of uncertainty that can occur clinically.
For an EB threshold of 10MeV, at least 76 PR probing locations were identified in each of the five investigated patients.

Realistically, not more than twenty PR QC acquisitions can be done per patient due to dose and time constraints. Consequently, we did not explore PR locations with a higher EB.

Our findings showed that the EB of a probing location correlates to the size of observed range shifts when simulating positioning errors of the probing location. This trend was observed for all five patients and for all investigated magnitudes of positioning errors.

To understand the maximum size of possibly occurring range uncertainties for a specific EB, we also calculated the maximum range shifts per EB. Range uncertainties for proton dose calculation are estimated to be around 3% [11]. They depend on several factors, such as the type and model of the CT scanner, the CT calibration, and the conversion of HU to stopping powers [12]. All maximum range shifts reported in table 3 are smaller than 3% and could therefore be considered as suitable PR probing locations for sCT quality control.

The investigation for this study was limited to patients with cranial CTs. When expanding to other body regions, more inhomogeneous anatomical regions will be encountered. For instance, sCT generation might work well for soft tissue, but less so for muscle and bone.

5 Conclusion

This study showed that PR locations suitable for sCT QC can be found by determining their EB. By simulating positioning errors, we demonstrated that the EB of an area surrounding a probing location correlates with the magnitude of the resulting range shift when simulating positioning errors.

We were able to locate a minimum of 19 suitable PR probing locations for sCT QC in each of the five investigated patients with a maximum of 0.6% of average range shift.

References

[1] F. Albertini, A. Bolsi, A. J. Lomax, et al. “Sensitivity of intensity modulated proton therapy plans to changes in patient weight”. *Radiotherapy and Oncology* 86.2 (2008), pp. 187–194. DOI: https://doi.org/10.1016/j.radonc.2007.11.032.

[2] H. Paganetti, P. Botas, G. C. Sharp, et al. “Adaptive proton therapy”. *Physics in Medicine Biology* 66.22 (Nov. 2021), 22TR01. DOI: 10.1088/1361-6560/ac344f.

[3] V. Giacometti, A. R. Hounsell, and C. K. McGarry. “A review of dose calculation approaches with cone beam CT in photon and proton therapy”. *Physica medica : PM : an international journal devoted to the applications of physics to medicine and biology : official journal of the Italian Association of Biomedical Physics (AIFB)* 76 (2020), pp. 243–276. DOI: 10.1016/j.ejmp.2020.06.017.

[4] A. Thummerer, P. Zaffino, A. Meijers, et al. “Comparison of CBCT based synthetic CT methods suitable for proton dose calculations in adaptive proton therapy”. *Physics in Medicine Biology* 65.9 (Apr. 2020), p. 095002. DOI: 10.1088/1361-6560/ab7d54.

[5] A. Thummerer, C. Seller Oria, P. Zaffino, et al. “Clinical suitability of deep learning based synthetic CTs for adaptive proton therapy of lung cancer”. *Medical Physics* 48.12 (2021), pp. 7673–7684. DOI: https://doi.org/10.1002/mp.15333.

[6] C. J. O’Hara, D. Bird, B. Al-Qaisieh, et al. “Assessment of CBCT–based synthetic CT generation accuracy for adaptive radiotherapy planning”. *Journal of Applied Clinical Medical Physics* 23.11 (2022), e13737. DOI: https://doi.org/10.1002/acm2.13737.

[7] C. Seller Oria, A. Thummerer, J. Free, et al. “Range probing as a quality control tool for CBCT-based synthetic CTs: In vivo application for head and neck cancer patients”. *Medical Physics* 48.8 (2021), pp. 4498–4505. DOI: https://doi.org/10.1002/mp.15020.

[8] A. Thummerer, C. Seller Oria, P. Zaffino, et al. “Deep learning–based 4D-synthetic CTs from sparse-view CBCTs for dose calculations in adaptive proton therapy”. *Medical Physics* 49.11 (2022), pp. 6824–6839. DOI: https://doi.org/10.1002/mp.15930.

[9] C. Seller Oria, A. Thummerer, J. Free, et al. “Range probing as a quality control tool for CBCT-based synthetic CTs: In vivo application for head and neck cancer patients”. *Medical Physics* 48.8 (2021), pp. 4498–4505. DOI: https://doi.org/10.1002/mp.15020.

[10] W. Liu, S. J. Frank, X. Li, et al. “Effectiveness of robust optimization in intensity-modulated proton therapy planning for head and neck cancers”. *Medical Physics* 40.5 (2013), p. 051711. DOI: https://doi.org/10.1118/1.4801899.

[11] H. Paganetti. “Range uncertainties in proton therapy and the role of Monte Carlo simulations”. *Physics in Medicine Biology* 57.11 (May 2012), R99. DOI: 10.1088/0031-9155/57/11/R99.

[12] M. Yang, X. R. Zhu, P. C. Park, et al. “Comprehensive analysis of proton range uncertainties related to patient stopping-power-ratio estimation using the stoichiometric calibration”. *Physics in Medicine Biology* 57.13 (June 2012), p. 4095. DOI: 10.1088/0031-9155/57/13/4095.

External validation of an AI-based tool for the early detection of radiotherapy treatment planning errors-An updated multicentric study

Petros Kalendralis*[1], Samuel M.H. Luk*[2], Alan M. Kalet[3], Andre Dekker**[1], Inigo Bermejo**[1,4], Tomasz Piotrowski**[5,6,7], Adam Ryczkowski**[5,6], Joanna Kazmierska**[5,8]

1. Department of Radiation Oncology (Maastro), GROW School for Oncology and Reproduction, Maastricht University Medical Centre+, Maastricht, The Netherlands
2. Department of Radiation Oncology, The University of Vermont Medical Center, Burlington, Vermont, United States
3. Department of Radiation Oncology, University of Washington Medical Center, Seattle, United States
4. Data Science Institute (DSI), Hasselt University, Hasselt, Belgium
5. Department of Electroradiology, Poznan University of Medical Sciences, Poznan, Poland
6. Department of Medical Physics, Greater Poland Cancer Centre, Poznan, Poland
7. Department of Biomedical Physics, Adam Mickiewicz University, Poznan, Poland
8. Radiotherapy Department II, Greater Poland Cancer Centre, Poznan, Poland

*Equal contribution

**Equal senior authors

Abstract: Bayesian Networks (BNs) represent a graphical modelling approach derived from the field of artificial intelligence (AI). They have demonstrated promise in assisting with the identification of errors in radiotherapy treatment plans. This study externally validated the BN developed with data from the University of Washington (UW), University of Vermont (UVM), and MAASTRO clinic in the Netherlands, using an independent dataset from the radiotherapy department of the Greater Poland Cancer Centre (GPCC) in Poznan. We used a dataset of 378 Volumetric Modulated Arc Therapy VMAT treatment plans, consisting of 802 treatment arcs. The results showed that the BN trained on the three centres (MAASTRO, UVM and UW) outperformed (AUC:78%) of the BNs trained in single centres (UW 66% and MAASTRO 75%). Future steps include the construction of a BN applied to specific treatment modalities and tumour locations for the creation of a more robust model.

1 Introduction

Radiotherapy (RT) represents an intricate multidisciplinary process wherein diverse professionals collaborate to formulate and execute a treatment plan [1]. Every stage of the RT workflow [2] is vulnerable to errors that can adversely affect the treatment outcome, carrying severe implications. For instance, errors such as radiation overdose during treatment delivery may result in radiation toxicity, leading to inflammatory or autoimmune diseases [3]. Presently, clinical RT care heavily relies on manual quality assurance (QA) tests to identify abnormalities and potential errors in radiotherapy processes [4]. Given the rapid technological advancements and increasing complexity in RT, ensuring quality through QA is paramount for delivering high-quality healthcare in radiation oncology.

In recent years, artificial intelligence (AI) techniques have been integrated into various facets of the RT workflow [5]. Specifically, AI has played a pivotal role in automating and expediting the RT treatment planning procedure [6], delineating organs at risk (OARs)[7], and developing decision support systems for patient treatment [8]. However, AI developments in quality assurance have been somewhat constrained[9]. Some efforts have been directed towards evaluating patient-specific QA and machine QA using datasets based on (inter)national treatment protocols and safety guidelines [10]. AI holds the potential to enhance the efficiency and effectiveness of QA without necessitating an increase in human resources [11].
Bayesian networks are probabilistic graphical models that depict interactions among a network of variables, offering the ability to estimate the probability of an event based on partial information. Their capacity to handle missing datavalues, a common occurrence in RT datasets, coupled with the intuitive nature of probabilistic reasoning, positions BNs as an ideal method for decision support in radiation oncology [12]. In comparison to rules-based algorithms and checklists designed to aid treatment plan review, both in-house and commercially [13], [14], [15], [16], BNs possess the advantage of emulating human reasoning and adapting to changes in clinical practice by updating the network model with new data [9].

Luk et al. [12] introduced a Bayesian network (BN) model for early detection of potential errors in RT treatment plans. Drawing on various diagnostic, treatment planning, radiation dose prescription, and patient setup variables, Luk et al.[17] devised an intelligent alert system that can notify medical physicists of potentially erroneous or suboptimal parameters in the RT treatment plan. The BN structure was constructed using a dependency-layered ontology [12], [18], [19] and insights from RT experts at the University of Washington [UW] (Seattle, WA, USA). Later, Kalendralis et al. built upon the BN proposed by Luk et al. and trained it on data from UW, University of Vermont (UVM), and MAASTRO clinic in the Netherlands.

The external validation of models in radiotherapy (and also in every scientific field) is essential for their potential clinical implementation [20]. The main goal of this study is to externally validate the BN model developed by Kalendralis et al. [21] by using an independent/external patient dataset collected at the radiotherapy department of the Greater Poland Cancer Centre (GPCC) in Poznan, Poland.

2 Materials and Methods

We used a patient dataset of 378 Volumetric Modulated Arc Therapy (VMAT) treatment plans who were treated at GPCC between January to May 2022 with TrueBeam accelerators (Varian Medical Systems, Palo Alto, USA).Furthermore, we have included head and neck, thorax, abdomen and pelvic cancer patients who were treated according to the treatment protocols prescribed based on the diagnosis and the localisation of the tumour. The treatment plans we used were created with 6 MeV photons based on the GPCC protocols and guidelines, based on conventional and stereotactic fractionation schemes. For the external validation of the BN we used simulated errors in 3% of the total number of treatment plans. The simulated errors were categorised into three types: Linear Accelerator (LINAC), dose prescription and general radiotherapy errors (Table 1). For the statistical analysis and the calculation of the error probabilities we have used the Java based application programming interface (API) Hugin Researcher 7.4 [22]. Additionally, for each error category we instantiated the variables Anatomic tumour location, T stage, N stage and M stage and calculated the error probabilities for the remaining variables of the dataset. The overall variables we had available in our dataset can be found in table 2.

To assess the model's performance in comparison to the results reported by its authors, we generated a receiver-operating characteristic curve (ROC) and computed the area under the curve (AUC) on our dataset. The AUC serves as an indicator of the model's discriminatory capability. We conducted this analysis for the entire dataset, combining all variables, and separately for variables with simulated errors: collimator angle, table angle, gantry angle, PTV dose, and bolus. Using the R language (version 2022.02.0 +443)[23] and the "classifierplots" package, we plotted the ROC but also calculated the AUC.

Table 1: Error categories

LINAC	Collimator angle
LINAC	Gantry angle
Dose prescription	Planning tumour volume (PTV)
General radiotherapy errors	Monitor Units (MU) per CentiGray (cGy)
General radiotherapy errors	Monitor Units (MU) per degree (Gantry angle)

Table 2: Variables used

Variable categories	Variable names
Clinical variables	Clinical T stage, Clinical N stage, Clinical M stage,Anatomic tumour location, Treatment intent
LINAC	Table angle, Gantry angle, Collimator angle
Radiotherapy generic	Number of plan adaptations,Radiation type, Plan technique, Number of beams, Control points, Source to surface distance (SSD), Beam energy, Table orientation, Monitor Units/cGy, Monitor Units/degree,
Dose prescription	Dose per fraction,Total dose, Number of fractions

3 Results

As presented in figure 1, in terms of the area under the receiver operating characteristic curve (AUC), the BN trained on the three centres (MAASTRO, UVM and UW) outperformed (AUC=78%) of the BNs trained in single centres (UW-66.1% and MAASTRO 75.%).

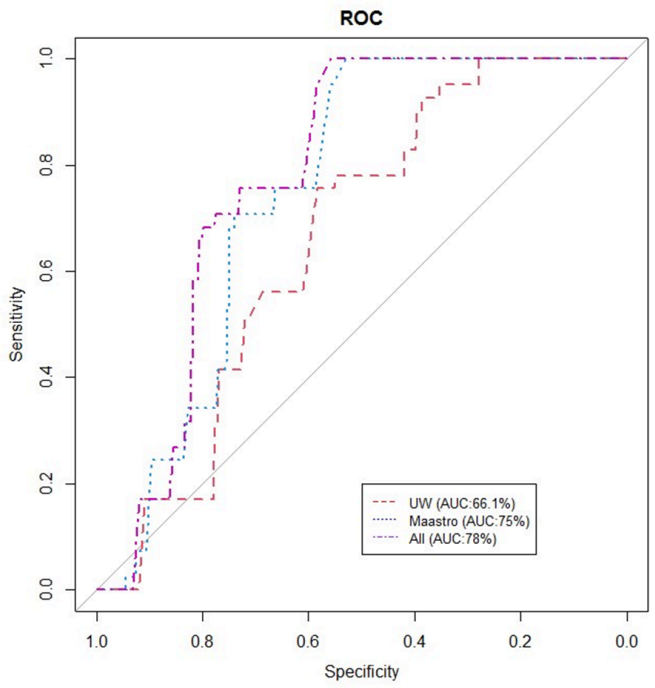


Figure 1: The receiver operating characteristic curves (ROCs) for the three Bayesian Networks (BNs) externally validated on GPCC data: the one trained on Maastro, UW and UW+UVM+Maastro.

4 Discussion

In this study we performed an external validation experiment having as a goal to test the performance of the BN developed by Kalendralis et al.[21] using an independent dataset of patients treated with different protocols and machines than the patients of the original BN dataset. The results showed that the BN is reproducible from other external institutes. Furthermore, according to the AUC results, the external validation showed that a BN trained on the data from multiple centres outperformed BNs trained on data from single centres.

This study includes several limitations. Initially, the categories of simulated errors were less than the original BN studies[17], [21]due to the absence of the "Bolus", "Wedge" and "Tolerance table" variables, which were included in the initial training of the BN[17], [21]. Secondly, the simulated error categories were selected based on the consultation of medical physicists and radiation oncologists of the centre in order to investigate the variables of interest based on their clinical practice.

Future steps will include the development of a novel error simulation method that will be applicable and reproducible by a wide range of international centres. Additionally, the development of more sophisticated algorithms that can predict the probabilities of radiotherapy treatment planning errors to happen is more than essential. For instance, deep learning (DL) algorithms have already shown their potential [24] as decision support systems in radiotherapy. Finally, the inclusion of additional treatment planning quality parameters such as imaging, Dose Volume Histogram (DVH) and dosimetric variables will have the potential to enrich the results of an AI assistance for the detection of treatment planning outliers for medical physicists.

5 Conclusion

In conclusion, we have presented an improvement of the BN performance, including an independent dataset from a European institute in order to potentially aid medical physicists and radiotherapy professionals in the treatment planning procedure.

References

[1] S. J. Gardner, J. Kim, and I. J. Chetty, "Modern Radiation Therapy Planning and Delivery," *Hematol. Oncol. Clin. North Am.*, vol. 33, no. 6, pp. 947–962, Dec. 2019, doi: 10.1016/j.hoc.2019.08.005.

[2] C. Henry, C. Kuziemsky, and J. Renaud, "Understanding Workflow in Radiation Therapy: Process Implications of Providing Outpatient Radiation Therapy Services to the Inpatient Population," *J. Med. Imaging Radiat. Sci.*, vol. 49, no. 1, p. S11, Mar. 2018, doi: 10.1016/j.jmir.2018.02.032.

[3] R. Yahyapour *et al.*, "Radiation-induced inflammation and autoimmune diseases," *Mil. Med. Res.*, vol. 5, no. 1, p. 9, Dec. 2018, doi: 10.1186/s40779-018-0156-7.

[4] T. K. Yeung, K. Bortolotto, S. Cosby, M. Hoar, and E. Lederer, "Quality assurance in radiotherapy: evaluation of errors and incidents recorded over a 10 year period," *Radiother. Oncol.*, vol. 74, no. 3, pp. 283–291, Mar. 2005, doi: 10.1016/j.radonc.2004.12.003.

[5] M. F. Chan, A. Witztum, and G. Valdes, "Integration of AI and Machine Learning in Radiotherapy QA," *Front. Artif. Intell.*, vol. 3, p. 577620, Sep. 2020, doi: 10.3389/frai.2020.577620.

[6] C. Wang, X. Zhu, J. C. Hong, and D. Zheng, "Artificial Intelligence in Radiotherapy Treatment Planning: Present and Future," *Technol. Cancer Res. Treat.*, vol. 18, p. 153303381987392, Jan. 2019, doi: 10.1177/1533033819873922.

[7] M. Chen, S. Wu, W. Zhao, Y. Zhou, Y. Zhou, and G. Wang, "Application of deep learning to auto-delineation of target volumes and organs at risk in radiotherapy," *Cancer/Radiothérapie*, vol. 26, no. 3, pp. 494–501, May 2022, doi: 10.1016/j.canrad.2021.08.020.

[8] G. Valdes *et al.*, "Clinical decision support of radiotherapy treatment planning: A data-driven machine learning strategy for patient-specific dosimetric decision making," *Radiother. Oncol.*, vol.

125, no. 3, pp. 392–397, Dec. 2017, doi: 10.1016/j.radonc.2017.10.014.
[9] S. M. H. Luk, E. C. Ford, M. H. Phillips, and A. M. Kalet, "Improving the Quality of Care in Radiation Oncology using Artificial Intelligence," *Clin. Oncol.*, vol. 34, no. 2, pp. 89–98, Feb. 2022, doi: 10.1016/j.clon.2021.11.011.
[10] T. R. McNutt, K. L. Moore, B. Wu, and J. L. Wright, "Use of Big Data for Quality Assurance in Radiation Therapy," *Semin. Radiat. Oncol.*, vol. 29, no. 4, pp. 326–332, Oct. 2019, doi: 10.1016/j.semradonc.2019.05.006.
[11] A. M. Kalet, S. M. H. Luk, and M. H. Phillips, "Radiation Therapy Quality Assurance Tasks and Tools: The Many Roles of Machine Learning," *Med. Phys.*, vol. 47, no. 5, May 2020, doi: 10.1002/mp.13445.
[12] A. M. Kalet, J. N. Doctor, J. H. Gennari, and M. H. Phillips, "Developing Bayesian networks from a dependency‐layered ontology: A proof‐of‐concept in radiation oncology," *Med. Phys.*, vol. 44, no. 8, pp. 4350–4359, Aug. 2017, doi: 10.1002/mp.12340.
[13] E. E. Furhang, J. Dolan, J. K. Sillanpaa, and L. B. Harrison, "Automating the initial physics chart-checking process," *J. Appl. Clin. Med. Phys.*, vol. 10, no. 1, pp. 129–135, Dec. 2009, doi: 10.1120/jacmp.v10i1.2855.
[14] D. Yang and K. L. Moore, "Automated radiotherapy treatment plan integrity verification: Plan checking using PINNACLE scripts," *Med. Phys.*, vol. 39, no. 3, pp. 1542–1551, Feb. 2012, doi: 10.1118/1.3683646.
[15] B. Sun *et al.*, "Initial experience with TrueBeam trajectory log files for radiation therapy delivery verification," *Pract. Radiat. Oncol.*, vol. 3, no. 4, pp. e199–e208, Oct. 2013, doi: 10.1016/j.prro.2012.11.013.
[16] C. Holdsworth *et al.*, "Computerized System for Safety Verification of External Beam Radiation Therapy Planning," *Int. J. Radiat. Oncol.*, vol. 98, no. 3, pp. 691–698, Jul. 2017, doi: 10.1016/j.ijrobp.2017.03.001.
[17] S. M. H. Luk *et al.*, "Characterization of a Bayesian network‐based radiotherapy plan verification model," *Med. Phys.*, vol. 46, no. 5, pp. 2006–2014, May 2019, doi: 10.1002/mp.13515.
[18] C. J. A. Wolfs, R. A. M. Canters, and F. Verhaegen, "Identification of treatment error types for lung cancer patients using convolutional neural networks and EPID dosimetry," *Radiother. Oncol.*, vol. 153, pp. 243–249, Dec. 2020, doi: 10.1016/j.radonc.2020.09.048.
[19] A. M. Kalet, J. H. Gennari, E. C. Ford, and M. H. Phillips, "Bayesian network models for error detection in radiotherapy plans," *Phys. Med. Biol.*, vol. 60, no. 7, pp. 2735–2749, Apr. 2015, doi: 10.1088/0031-9155/60/7/2735.
[20] C. L. Ramspek, K. J. Jager, F. W. Dekker, C. Zoccali, and M. Van Diepen, "External validation of prognostic models: what, why, how, when and where?," *Clin. Kidney J.*, vol. 14, no. 1, pp. 49–58, Feb. 2021, doi: 10.1093/ckj/sfaa188.
[21] P. Kalendralis *et al.*, "Automatic quality assurance of radiotherapy treatment plans using Bayesian networks: A multi-institutional study," *Front. Oncol.*, vol. 13, p. 1099994, Feb. 2023, doi: 10.3389/fonc.2023.1099994.
[22] S. K. Andersen, K. G. Olesen, F. V. Jensen, F. Jensen, G. Shafer, and J. Pearl (Eds.), "Hugin – A shell for building belief universes for expert systems," in *Reading in Uncertainty*, 1990, pp. 332–337.
[23] "RStudio." [Online]. Available: https://www.rstudio.com/
[24] S. Liu *et al.*, "Implementation of a knowledge‐based decision support system for treatment plan auditing through automation," *Med. Phys.*, vol. 50, no. 11, pp. 6978–6989, Nov. 2023, doi: 10.1002/mp.16472.

Development of an Automated Radiotherapy Daily QA System

Marcus Tyyger[1], Matthew Carlisle[1], Jack P.C. Baldwin[1], Richard Homer[1], Phillip Rixham[1], and Steve Weston[1]

[1]Medical Physics and Clinical Engineering, Leeds Teaching Hospitals NHS Trust, UK

Abstract

Purpose: Develop an EPID-based daily constancy check system, which does not require a phantom, and can reduce overall quality assurance (QA) burden.

Materials and Methods: Daily images were acquired across 12 linear accelerators for 6 and 10MV FFF and flattened beams. Local computing infrastructure was setup to automatically transfer data and analyse images. Results were made visible through a webserver for online and offline analysis. Image analysis was validated against standard QA measurements.

Results: Output median error 0.1% (range: -0.9, 1.7). The median symmetry and flatness errors were less than 0.1 % (range: -2.3, 2.0), and -0.6% (range: -3.1, 3.3) respectively.

Conclusion: In-house development can provide solutions for RT departments where commercial solutions are unsuitable or unavailable. The developed system can be employed as an output daily constancy check, and trend analysis of beam characteristics enables potential reduction of monthly QA tasks.

1 Introduction

Radiotherapy (RT) departments continue to see increasing patient workloads[1] and increasingly time-consuming treatment pathways. Given this, departments are incentivised to increase linear accelerator (linac) capacity where possible. One potential solution is to reduce the quality assurance (QA) burden, on a monthly and daily timescale.

Daily output constancy checks ensure linac performance for that day, and are regularly performed using external devices which require setup, training, and management. Ideally, daily QA would capture a wide range of linac performance metrics, and enable assessment of long-term performance. This may allow a reduction in monthly quality control (QC) tasks when linac performance is stable. For departments to substantially benefit from this approach, daily workloads cannot simply be increased. Therefore, daily QA checks must require minimal human and linac resources for setup, analysis, and recording of results.

Amorphous-silicon electronic portal imaging devices (EPID) have previously been shown to be suitable for linac QA[2], which led to various commercial products[3,4]. However, these can be manufacturer or linac specific, require a phantom, or only provide online analysis whilst the operator is performing the test. RT departments may be unable to purchase a tool that suits their clinical pathways. This can lead to departments spending significant resources on investigation, evaluation, and commissioning of products that do not make QA as efficient as possible.

Herein, the development of a system for daily QA which uses EPID images without additional build-up or phantoms is described. The aim is to show the potential for RT departments to develop bespoke tools to suit their clinical need, and highlight the potential of EPID-based QA to reduce the overall QA burden.

2 Materials and Methods

Daily images were acquired on twelve VersaHD linacs (Elekta AB, Stockholm, Sweden) between 18th December 2023 and 26th January 2024. Flattening filter free (FFF) and flattened beams were delivered at 6 MV and 10 MV at maximum dose rate, with 100 monitor units (~1Gy), and a 20 x 20 cm open field at isocentre. All images were acquired on an IViewGT EPID panel without any additional build-up, at gantry and collimator angle 0°.

Data Pipeline and Infrastructure

Acquired images were automatically exported as DICOM images from the IView acquisition station to a PACS server and then on to a dedicated virtual machine (VM) running a DICOM service class provider (SCP). Each field's associated xml file was automatically copied from the local database to the VM, as this was required for dosimetric analysis.

Image processing was performed on the VM using software developed in Python, with results stored in an SQL database. Users accessed results through a locally hosted webserver which directly queried the database. As processing was automatic and results were stored, this allowed online review and offline trend-analysis.

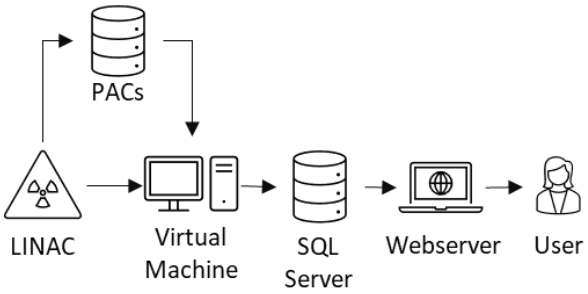


Figure 1: Developed data transfer and processing pipeline

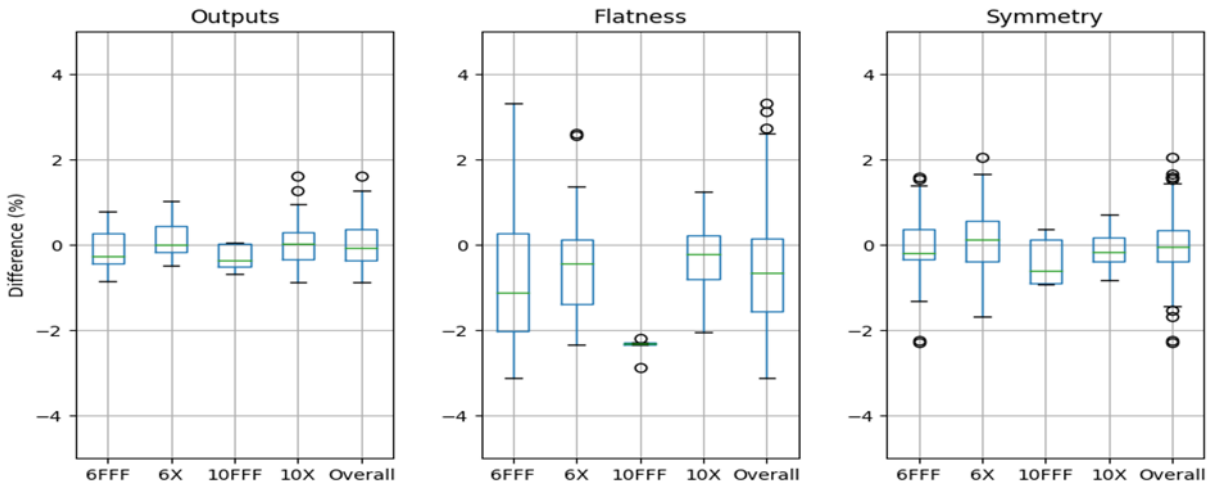


Figure 2: Validation results, "overall" is all beams combined. Outputs compared to ionisation chamber measurements. Flatness and symmetry results compared to 2D array measurements.

Image Processing and Analysis

Outputs were determined through the equation below, where P_{avg} is the average pixel value within the central 50 x 50 pixel (~ 12.5 x 12.5 mm measured at isocentre) region. F_{cal} is the beam and linac specific calibration factor. F_{pix} is the image specific dosimetric pixel factor.

$$Output = \frac{P_{avg} \cdot F_{cal}}{F_{pix}}$$

Output validation was performed by comparing ionisation chamber readings with software results of images taken within 1 day of the measurement.

Flatness and symmetry were determined through profiles through the x- and y-axes within the "field region". Initially 10 pixel wide profiles through the central beam axis were averaged and subsequently normalised. FFF normalisation was performed using the technique proposed by Fogliata et al[5] using Elekta VersaHD factors from Fogliata et al[6]. Flattened beams were normalised to 100% on the beam central axis. Field sizes were calculated by determining 50% dose points, and subsequently the field region was defined as 80% of the field size centred on the beam central axis.

Beam flatness was calculated through the ratio of the maximum (D_{max}) and minimum (D_{min}) doses. FFF unflatness was determined through the ratio of central axis dose (D_{cax}) and the dose off-axis at the edge of the field region (D_{edge}). For simplicity, "flatness" refers to both flattened beam flatness and FFF beam unflatness.

$$Flatness = \frac{D_{max}}{D_{min}} \qquad Unflatness = \frac{D_{cax}}{D_{edge}}$$

Beam symmetry was calculated by the maximum ratio of dose values at opposing points either side of the central axis, where x is the distance from the central axis.

$$Symmetry = \left(\frac{D_x}{D_{-x}}\right)_{max}$$

Flatness and symmetry results were validated by monthly QA results, measured using a 2-dimensional ionisation chamber array STARCHECK, with an additional 20 mm RW3 build up (PTW, Frieburg, Germany). Images taken within 1 day of the measurements were used for the comparison.

3 Results

Development was undertaken by a small team of clinical scientists and technologists alongside their routine workloads, and the first prototype was finished within six months. Additional work was required over the following months to correct minor infrastructure issues, such as VM access, DICOM SCP stability issues, and setting up automatic processing. Changes were also required to improve FFF beam analysis. Tolerance development and long-term validation is on-going and predicted to take approximately 6-months. This would give an overall time period from conception to clinical release of approximately 15-18 months, including contingency for potential delays.

Outputs were found to be in close agreement with ionisation chamber results. The median error was under -0.1 % across all energies and linacs, with a minimum and maximum error of -0.88 % and 1.61 % respectively. Due to issues with beam setup, five 10FFF results were analysable compared to ~30 for the other beams.

Across all energies, linacs, and profile orientations compared to array measurements symmetry errors were found to have a median value of -0.1 % (range: -2.3, 2.0).

For flatness this median value was -0.7 %, with a broader range of -4.3 to 3.3 %. Due to issues with availability, only six 10FFF images were processed.

4 Discussion

The development life-cycle of this tool has been roughly equivalent to the estimated time required to procure and commission a commercial product. Given only a small team was used, and no member was pulled from routine clinical duties this shows the potential for in-house development when no suitable commercial products are available. However, it cannot be ignored that in-house development requires staff expertise which may not be readily available in all RT departments.

Output results were within the recommend constancy action limit[7], giving confidence the developed system could be used as a constancy device. Locally this can reduce the time taken for daily QA as data can be acquired from morning run-ups rather than from a dedicated QC session. The results presented are only over approximately 30 days, meaning long-term stability needs to be validated prior to clinical deployment, although previous studies have shown EPID panels to be stable for this purpose[8,9].

Symmetry results found a minor median error compared to 2-dimensional array measurements, although, the range of errors did exceed +/- 2 %. Given the recommended monthly QA limit is also 2 %[7], this approach should not be used for individual accurate measurements. This is also true for flatness measurements which had an even broader range of errors compared to the Starcheck at approximately +/- 4 %. These results are likely in part due to the lack of charge particle equilibrium for EPID measurements of 10 MV beams. This work has demonstrated that the system can be used as a constancy device to monitor performance over time. Given stable results within local tolerances, the frequency of conventional output, flatness and symmetry measurements could be reduced. Tolerance analysis is necessary to commission this approach, and this work is currently locally on-going.

Flatness measurements showed a distinct difference between FFF and flattened beams, with FFF beams showing a larger range of errors and a bias towards underestimating flatness. It is hypothesised that this might be due to underlying issues with panel response rate; as if the panel is under recording the maximum delivered dose then it will calculate a lower FFF beam flatness. EPID panels have been shown to have differing responses due to dose rate[8]. Additionally, the results here show 10FFF having larger errors than 6FFF and it has a significantly higher delivered dose rate. If this is the underlying cause then simply reducing the delivered dose rate would remove this source of error.

5 Conclusion

The potential for in-house development where commercial solutions are not suitable has been demonstrated in the context of EPID-based daily QA. Validation against commissioned QA devices show the system performs adequately for use as an output constancy check, without the need to setup a phantom. Trend analysis of beam characteristics: beam flatness and symmetry, enables potential reduction of certain monthly QC tasks, or as a predictive tool of linac failures.

References

1. Willmann J, Poortmans P, Monti AF, et al. Development of staffing, workload and infrastructure in member departments of the European Organisation for Research and Treatment of Cancer (EORTC) radiation oncology group. *Radiother Oncol*. 2021;155:226-231. doi:10.1016/J.RADONC.2020.11.009
2. Budgell GJ, Zhang R, Mackay RI. Daily monitoring of linear accelerator beam parameters using an amorphous silicon EPID. *Phys Med Biol*. 2007;52(6):1721-1733. doi:10.1088/0031-9155/52/6/012
3. Barnes MP, Greer PB. Evaluation of the truebeam machine performance check (MPC) geometric checks for daily IGRT geometric accuracy quality assurance. Published online 2017. doi:10.1002/acm2.12064
4. Stambaugh C, Yancey J, Shukla U, Melhus C, Stambaugh N. Daily Quality Assurance Efficiency Evaluation Using SunCHECK Machine and Machine Performance Check. Published online 2023. doi:10.7759/cureus.35695
5. Fogliata A, Garcia R, Knöös T, et al. Definition of parameters for quality assurance of flattening filter free (FFF) photon beams in radiation therapy. *Med Phys*. 2012;39(10):6455-6464. doi:10.1118/1.4754799
6. Fogliata A, Fleckenstein J, Schneider F, et al. Flattening filter free beams from TrueBeam and Versa HD units: Evaluation of the parameters for quality assurance. *Med Phys*. 2016;43(1):205-212. doi:10.1118/1.4938060
7. Imran Patel, ed. *Physics Aspects of Quality Control in Radiotherapy (Report No. 81)*. 2nd ed. Institute of Physics and Engineering in Medicine; 2018.
8. Winkler P, Hefner A, Georg D. Dose-response characteristics of an amorphous silicon EPID. *Med Phys*. 2005;32(10):3095-3105. doi:10.1118/1.2040711
9. Ma Y, Wang X, Mai R, et al. An electronic portal image device (EPID)-based multiplatform rapid daily LINAC QA tool. *J Appl Clin Med Phys*. 2021;22(1):45-58. doi:10.1002/ACM2.13055

Isocenter deviation assessment and model for Halcyon Radiotherapy Systems

Linda Lankinen[1,2], Antti Kulmala[3], Jouko Lehtomäki[1], and Ari Harju[1]

[1]Varian Medical Systems, a Siemens Healthineers Company
[2]Department of Physics, University of Helsinki, Finland
[3]Clinical Research Institute HUCH Ltd., Helsinki, Finland

Abstract The radiation isocenter position on a Halcyon radiotherapy system was measured and the deviation from the intersection of the rotation axes, the ideal isocenter, implemented in Monte Carlo (MC) model. With a test VMAT plan, the isocenter deviation was assessed by analyzing the difference between the film measurement and MC simulated dose distribution. The implementation of the deviation improved the agreement between the MC model and the measured dose distribution especially in the steep dose gradient regions. While the treatment system delivery met all isocenter related specifications, the improvements demonstrated by the modeling of the isocenter displacements could lead to more accurate dose predictions for treatments where steep dose gradients are needed around small targets.

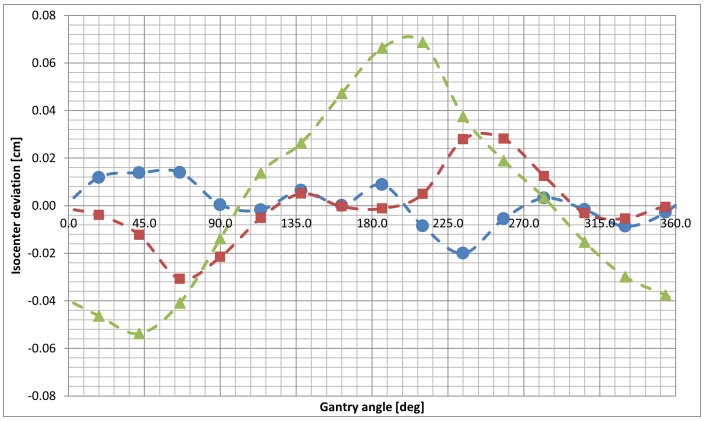


Figure 1: The radiation isocenter deviation of a Halcyon treatment unit (S4, Cancer Center, Helsinki, Finland) as a function of a gantry angle in IEC61217 coordinate system. The variations along the X-axis (blue dots), Y-axis (green triangle) and Z-axis (red square) are shown.

1 Introduction

The isocenter is the primary reference location of external radiotherapy beam. In ideal conditions, the mechanical isocenter of linear accelerator is defined as the point of intersection of gantry, collimator, and treatment table rotation axes. However, the gravity pull on the beam centerline and the mechanical uncertainties of the gantry cause systematic deviations of the isocenter position as a function of the gantry angle [1,2]. The geometric position of the radiation isocenter during rotation is usually assumed to be inside a virtual spherical volume. The AAPM Task Group Report 142 (2009) recommends that up to $\pm$ 1 mm deviation is acceptable for stereotactic treatments [3]. Discrepancies in the isocenter position may lead to loss of target coverage or increased dose delivery to the healthy tissues as the exact position of the target is relative to the mechanical isocenter, and by considering the isocenter movement, the accuracy of stereotactic treatments could be improved [1].

The goal of this study was to measure the radiation isocenter position on a Halcyon radiotherapy system (Varian Medical Systems, a Siemens Healthineers Company). The results also represent Ethos Radiotherapy Systems as the platform shares the same beam generation and collimation components [4]. The deviation of the isocenter position was replicated with Monte Carlo (MC) model. On one hand, the MC model can act as a computational copy of an actual machine and provide a reference without uncertainties due to the mechanical variations, and on the other hand, it can inherit the known mechanical variations and mimic the behavior of the actual machine.

2 Materials and Methods

The deviation of the isocenter position on the Halcyon radiotherapy system ("S4") installed in the Helsinki Cancer Center was measured. Lateral (= XIEC61217) and vertical (= ZIEC61217) variation was determined using a star-shot procedure with fifteen equally spaced gantry angles [1]. For the longitudinal (= YIEC61217) variation, a special film measurements procedure was designed. The projections of the star-shot beams were measured on XY- and YZ-planes with beam projections from the reference angles. Then, the alignment of the beams was compared to the alignment of the reference geometry and the offset was determined. The measurements were repeated to reduce the random uncertainty.
The measured isocenter deviation data was implemented to the MC model by adjusting the source position and treatment head rotation accordingly. Each direction was interpolated and processed separately. The MC model has been validated to match to the vendor provided pre-configured beam model [5]. The machine-specific mechanical leaf gap (MLG) of -0.65 mm in nominal positions was used for both proximal and distal MLC stacks.
To analyze the impact and quantify the agreement between the machine delivery and predicted dose distributions, a test Volumetric Modulated Arc Therapy (VMAT) plan was produced. The test VMAT plan, from a manuscript under review [6], had two optimized VMAT arcs (Photon Optimizer

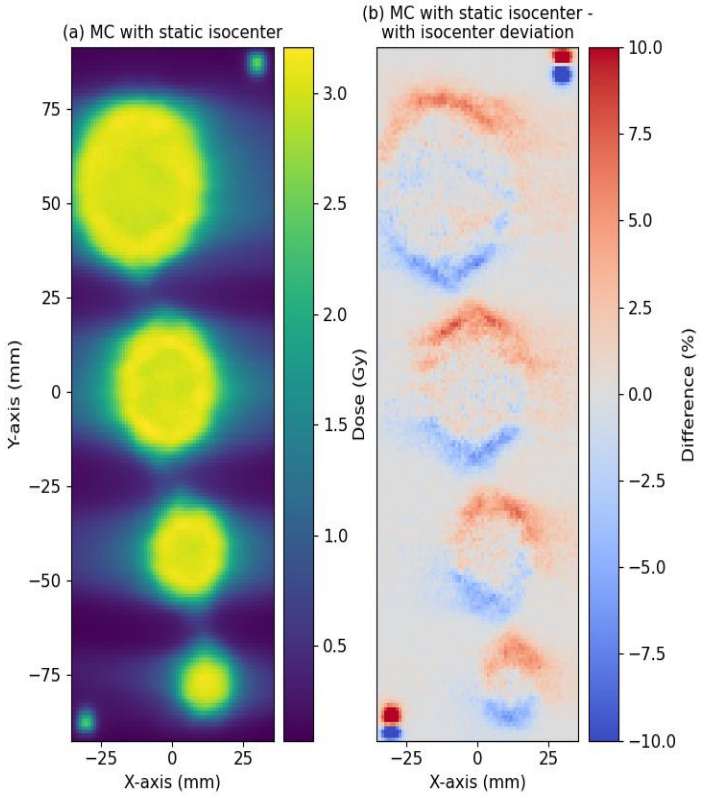


Figure 2: The MC simulated dose plane with static isocenter on the left and the relative dose plane difference between MC simulated dose planes with static isocenter and with isocenter deviation. Both graphs share the same Y-axis.

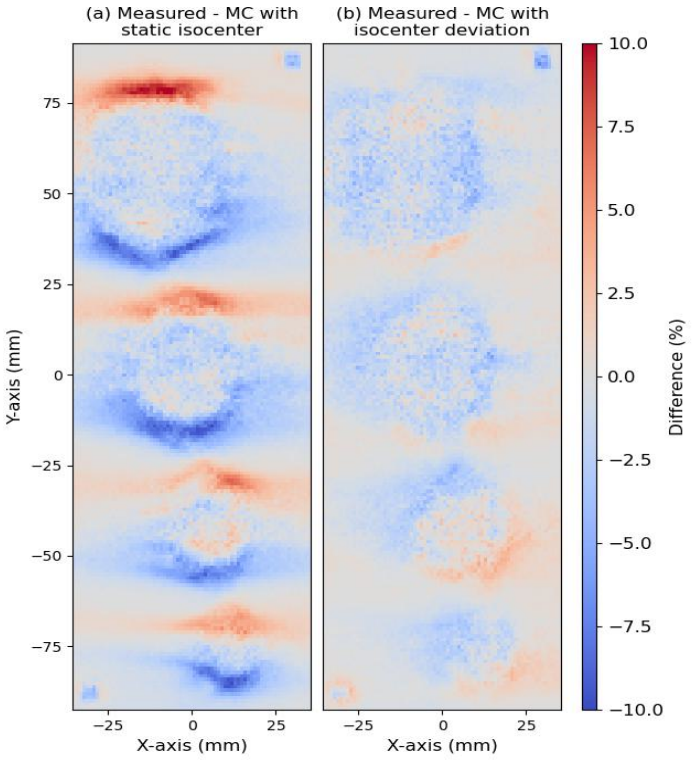


Figure 3: The relative dose plane difference between the measured and MC with static isocenter on the left and with isocenter deviation on the right. Both graphs share the same Y-axis.

18.0.0[7]) with the collimator angles 30° and 330°, and one static alignment field with zero-degree gantry, which was used for confirm the film position. The plan had four spherical targets with the diameters of (4.0, 3.0, 2.0, 1.0) cm and the static alignment field had square 5 mm x 5 mm openings. The measured dose distributions were obtained by using GafChromic™ EBT3 radiochromic film (Ashland Speciality Ingredients G.P., Bridgewater, NJ), the Vidar DosimetryPro Advantage (red) densitometer (Vidar Systems Corporation, Herndon, VA), and the RW3 water-equivalent plastic phantom (PTW, Freiburg, Germany). The phantom size was 30 cm × 30 cm × 10 cm and source to surface distance (SSD) was 95 cm. The machine type had fixed 800 MU/min dose rate.

3 Results

The minimum distance between the center of the beam and the ideal treatment isocenter on a Halcyon radiotherapy system was measured with different gantry rotations. Figure 1 shows the isocenter deviation for each direction in IEC61217 coordinate system (X, Y, Z) as a function of a gantry angle. The range of the variation along X-direction was ± 0.02 cm (blue dots), ± 0.07 cm in Y-direction (green triangles), and ± 0.04 cm in Z-direction (red squares). The largest value of 0.069 cm was measured along Y-direction at 210°. The uncertainty in determining the isocenter deviation in each direction separately was found to be <= 0.01 cm and the gantry angle uncertainty was ± 0.15°.

The measured isocenter deviation was implemented in the MC model and the results in comparison to static isocenter analyzed. For a test VMAT plan with the two fields and the collimator angles of 30° and 330°, the MC simulated dose plane is presented in Figure 2 (a). The dose plane has the four sphere targets, and on the two corners, 5 mm x 5 mm openings from the alignment field with zero-degree gantry. The relative dose difference between MC simulated dose with static isocenter and with isocenter deviation is shown in Figure 2 (b). The most profound differences are shown with the alignment field openings. With the VMAT fields, the effect is smaller, but shift-like differences up to 8 % are observed on the Y-edges of the targets.

The differences between the measured and MC dose planes are shown in Figure 3. The dose plane difference between the measured and MC with static isocenter is presented in Figure 3 (a). A systematic difference is observed on the Y-edges of the sphere targets similar to Figure 2. The largest dose differences reach up to 10 %. The positioning is correct as the alignment fields openings show no systematic differences.

An improved agreement can be observed between the measured and MC with the implementation of the isocenter deviation in Figure 3 (b). Systematic differences are no longer present and the maximum difference is below 4.1%. By taking profiles passing through the spherical targets, the same

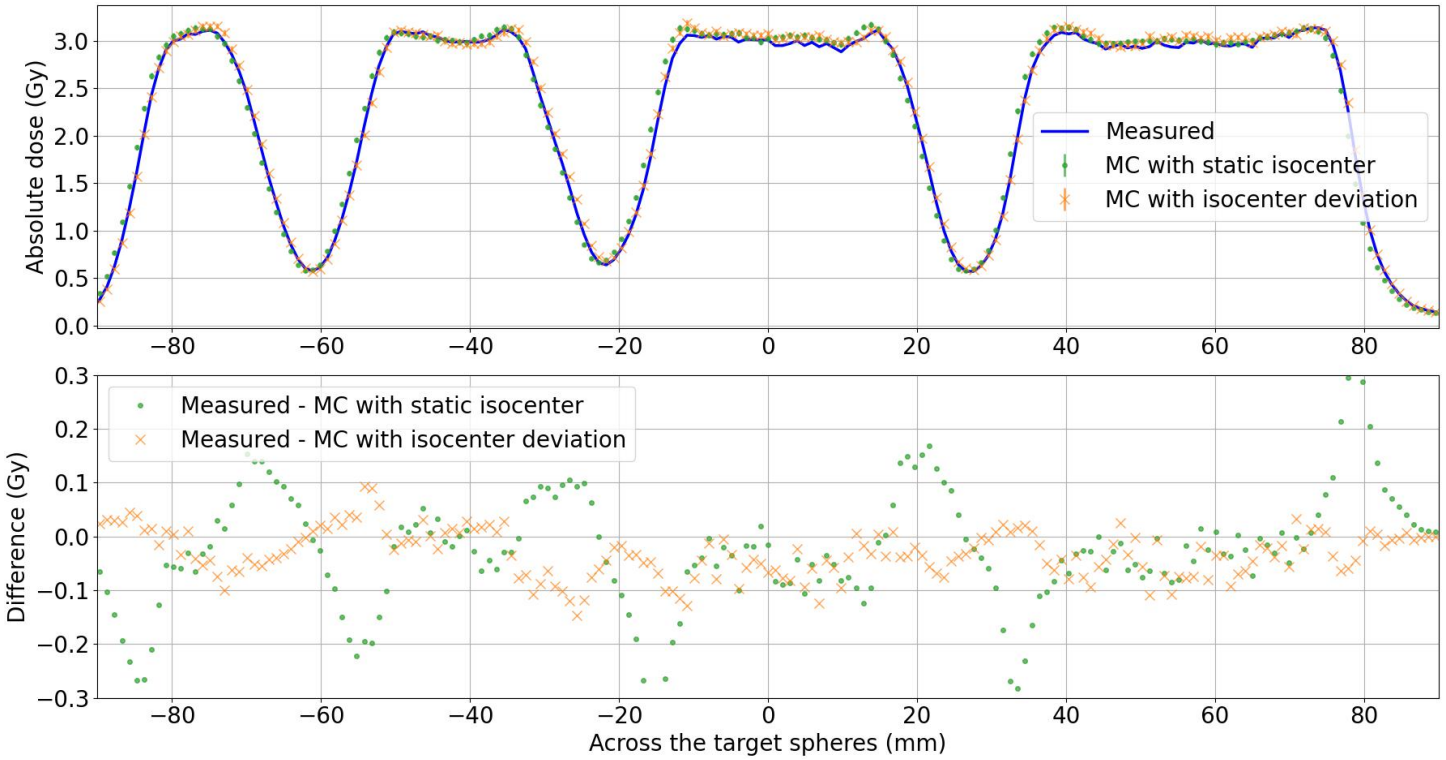


Figure 4: The dose profiles across the target spheres and the dose difference between the measured and MC.

observations can be confirmed in Figure 4. The dose profiles of the measured (blue line), MC with static isocenter (green dots), and MC with isocenter deviation (orange crosses) are presented on the top of the graph and beneath the local dose difference between the measured and MC dose profiles. The systematic shift between the measured and MC with static isocenter creates local deviations up to $\pm$ 0.3 Gy in the dose gradient regions. Between the measured and MC with isocenter deviation, the difference is smaller and the local deviations are only up to $\pm$ 0.15 Gy.

4 Discussion

The radiation isocenter deviation as a function of a gantry rotation was measured in all three dimensions on a Halcyon radiotherapy system. The deviation was less than 0.1 cm. The largest range of offset was observed with the longitudinal Y-axis up to $\pm$ 0.07 cm. The vertical Z-axis range was less than 0.04 cm and the lateral X-axis range was less than 0.02 cm. However, all directions are coupled by the collimator and the gantry angle. The effect of the isocenter deviation on the VMAT test plan was analyzed by taking into account the difference between the center of the beam and the ideal treatment isocenter in the MC model as a function of the gantry rotation. This study quantified the radiation isocenter deviation and its impact on the delivered dose for a test VMAT plan with spherical targets and high dose gradients. Although, the coincidence of the radiation beam to the mechanical axis was well below 1 mm, the effect can be modeled, leading to improved accuracy in the predicted dose delivery. As the focus of this study was on the isocenter, the uncertainties from the MLC position and the gravity effect on the position during the gantry rotation was ignored. It should be kept in mind that with a clinical case, target delineation, patient setup error, and target movement would be larger sources of uncertainties.

5 Conclusion

Based on the measured data, a gantry angle-specific isocenter deviation was applied to a Monte Carlo model validated for the Halcyon and Ethos radiotherapy systems. The consideration of the isocenter variation with the gantry angle improved the argeement between the measured and simulated dose in the gradient regions of the test VMAT plan. Further considerations such as the impact on clinical plan quality of more sophisticated radiation isocenter modeling will be addressed in future studies.

References

[1] A. Gónzalez, I. Castro, and J. A. Martínez, A procedure to determine the radiation isocenter size in a linear accelerator, Medical Physics **31**, 1489–1493 (2004).

[2] W. Du, S. Gao, X.-C. Wang, and R. J. Kudchadker, Quantifying the gantry sag on linear accelerators and introducing an MLC-based compensation strategy, Medical Physics **39 4**, 2156–62 (2012).

[3] E. E. Klein, J. Hanley, J. Bayouth, F.-F. Yin, W. Simon,

S. Dresser, C. Serago, F. Aguirre, L. Ma, B. Arjomandy, C. Liu, C. Sandin, and T. Holmes, Task Group 142 report: Quality assurance of medical acceleratorsa, Medical Physics **36**, 4197–4212 (2009).

[4] Varian Medical Systems Inc., Halcyon and Ethos Radiotherapy System Instructions for Use, 2021.

[5] L. Laakkonen, J. Lehtomäki, A. Cahill, M. Constantin, A. Kulmala, and A. Harju, Monte Carlo modeling of Halcyon and Ethos radiotherapy beam using CAD geometry: validation and IAEA-compliant phase space, Physics in Medicine and Biology **68**, 044002 (2023).

[6] L. Lankinen, A. Kulmala, J. Lehtomäki, and A. Harju, The delivery assessment for small targets on Halcyon radiotherapy system - measured and calculated dose comparison, Journal of Applied Clinical Medical Physics (Under Review).

[7] Varian Medical Systems, Inc., Eclipse Photon and Electron Algorithms Reference Guide, 2022.

Patient-Specific Quality Assurance in Radiotherapy: Prediction with Deep Hybrid Learning for VMAT plans on Rapid Arc and IMRT plans on Halcyon machines

Noémie N Moreau[1,2], Christine Boutry[1], Laurine Bonnor[1], Cyril Jaudet[1], Nadia Falzone[3], Alain Batalla[1], Laetitia Lechippey[1], Cindy Bertaut[4] and Aurélien Corroyer-Dulmont[1,2]

[1]Medical Physics Department, CLCC François Baclesse, 14000 Caen, France
[2]Normandie University, UNICAEN, CNRS, ISTCT, CYCERON, 14000 Caen, France
[3]GenesisCare Theranostics, Building 1 & 11, The Mill, 41-43 Bourke Road, Alexandria, NSW 2015, Australia
[4]Medical Physics Department, Cherbourg Hospital, 50100 Cherbourg, France

Abstract: Arc therapy in radiotherapy (RT) allows for better conformation of the dose deposition but requires patient-specific quality assurance (QA). The aim of this study was to develop a predictive model of Delta4-QA outcomes based on RT-plan complexity indices to reduce QA workload.

Methods: From 1632 RT-VMAT-plans, six complexity indices were extracted. Machine learning (ML) and deep hybrid learning (DHL, for complex tumor locations, ie breast, pelvis and head and neck) models were developed to predict QA-plan compliance.

Results: ML model achieved 100% specificity and 98.8% sensitivity for non-complex RT-plans (with brain and thorax tumor locations). For more complex RT-plans, sensitivity drops to 87%, innovative method using DHL achieved 100% sensitivity and 97.78% specificity.

Conclusion: QA results were predicted with high accuracy. A similar approach was realized to predict intensity-modulated RT QA compliance for breast tumor treatment with Halcyon machine. These results will also be presented during the conference.

1 Introduction

Over the past decade, numerous technological strides have advanced radiotherapy techniques, resulting in enhanced methodologies. One notable technique among them is volumetric intensity modulated arc therapy (VMAT) [1]. VMAT operates by dynamically adjusting the beam fluence as the accelerator arm rotates (or arcs), facilitating the delivery of a tumoricidal dose while preserving the integrity of healthy tissue. During the rotation of the accelerator arm, the multi-leaf collimator (MLC) modulates photon fluence. This complexity adds an additional layer of complexity to arc therapy treatment plans compared to conventional treatments [2]. As the complexity of treatment plans increase, a significant challenge to accommodate the patient anatomy, dosimetric constraints, optimization algorithms and the capabilities of the linear accelerator (linac) to guarantee that the administered treatment aligns with the initially planned parameters on the treatment planning system (TPS). In this context, it is imperative to conduct patient-specific quality assurance (QA) for each treatment plan before the actual patient treatment takes place. Using dedicated phantoms, the QA process assesses whether the planned dose in the TPS aligns with the dose administered to the patient on the accelerator. This evaluation uses two metrics, mean gamma and gamma index [3], to determine the compliance of the treatment plan. It is important to emphasize that QA is a time-consuming process, leading to a substantial period of unavailability for the machines dedicated to patient treatment. Furthermore, only five percent of the treatment plans at out center are non-compliant, with a gamma index < 0.95 and a tolerance of 3%/3mm.

Various alternatives to QA have been explored and discussed in diverse literature reviews [4]–[8]. Among the approaches investigated for predicting QA using complexity indexes derived from treatment plans, one method involves employing a threshold approach for these complexity indexes [9]. The second approach involves a more complex method relying on prediction models like machine learning regression and classification. As outlined in the recent review [10], various studies have explored the application of artificial intelligence (AI) in patient QA for radiotherapy. Nevertheless, apart from two studies [11], [12], most AI models were developed using a limited number of radiotherapy (RT) plans (between 250 and 600). This limited dataset size hinders the assessment of model robustness across a larger set of plans. A study [13], an auto encoder deep learning-based model was trained ton 426 RT plans and validated on 150 RT plans. The results indicated improved performance, achieving 100% sensitivity and 83% specificity for technical validation and 100% sensitivity and 72% specificity for clinical validation.

These results were obtained with a high degree of accuracy, but still with a high degree of error, especially for RT plans with a high arc count, high flow variability, high accelerator arm speed, leaves speed, and the opening of multi-leaf collimators (MLC). Furthermore, the absence of publicly available solutions derived from these models hampers their applicability in clinical practice. Consequently, the demand persists for more complex AI models tailored to complex RT plans, along with user-friendly solutions accessible for widespread clinical use.

Subsequently, more complex and innovative AI models have been developed. These include the combination of machine learning and deep learning approaches known as deep hybrid learning (DHL). Automating the analysis of radiographic images to diagnose various pathologies is one example of its application [14]. No study has used the DHL model to predict QA outcomes for multiple RT plans to our knowledge.

The aim of this study was to develop clinically applicable AI models. These models are more complex than those proposed in the literature. The models predict QA outcomes based on complexity indices of treatment plans. Machine learning classification models for simple to moderately complex RT VMAT plans were developed in the first step. More complex and innovative DHL models for more complex RT VMAT plans were developed in the second step. Finally, a freely accessible online platform was developed for the use of these models in clinical practice.

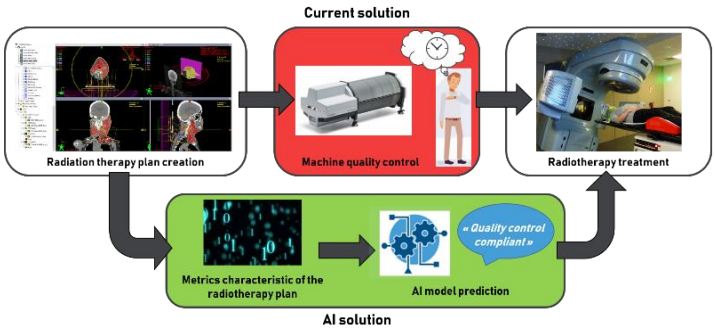


Figure 1: Can AI predict radiotherapy plan compliance?

2 Materials and Methods

2.1 Patient Cohorts

For this prospective study one thousand, six hundred and thirty-two patients with cancer referred to our oncology center between June 2019 and March 2022 were included. MR-004 a French national institution (INDS) that sets guidelines for the conduct of health research was used for this study. A few locations (digestive, rachis, other, member and skin tumor location) have a small number of cases, which is too small to build efficient models for these sites. Therefore, we only combine these cases in the "all" model and we did not develop specific models for these locations.

2.2 Radiotherapy Plan

Radiation therapy was delivered using volumetric modulated arc therapy (VMAT). A RapidArc (Varian®, Le Plessis-Robinson, France) machine was used. Eclipse® TPS was used to calculate radiotherapy plans (RT-plans). The complexity indexes were obtained from RT-plan using Raystation® TPS and an in-house python code available at https://github.com/AurelienCD/DeepHybridLearning_RadiotherapyQA_Depository_ManuscriptID_Diagnostics.
Only six complexity indexes were used to achieve a simple solution applicable to routine clinical settings: Leaf Sequence Variability (LSV), Aperture Area Variability (AAV), Modulation Complexity Score (MCS), Leaf Travel (LT), combination of LT and MCS (LTMCS) and Small Aperture Score (SAS10) [5]. Complexity indexes were derived from the treatment plan and not individually from each arc.

2.3 Patient Specific Quality Assurance

The Delta4® phantom manufactured by Scandidos® was employed for patient-specific QA. To verify the QA conformance, the local gamma index must surpass 95% with a 3%/3mm criterion, and the gamma mean should be below a predefined threshold. By identifying gamma mean values that are significantly ($p<0.05$), a specific gamma mean threshold was determined for each tumor location.
For AI construction models, conformance and non-conformance QA was assigned to 1 and 0, respectively.

2.4 Statistical Analysis

The ML and DHL performance models was assessed by examining the accuracy scores from Sklearn and TensorFlow on the test dataset, along with the Area Under the Curve (AUC) values derived from the ROC curves, specificity, sensitivity and the number of false positives. All statistical analyses were conducted in Python, utilizing the statistical data visualization and Seaborn library [15], [16].

3 Results

3.1. Model predictions for all tumor location

The initial strategy involved creating machine learning models capable of predicting QC conformity regardless of the tumor specific location. Various ML models were employed for this purpose. The large dataset of RT plans contributed to robust training and validation scores across the diverse ML models used. Random Forest. Classifier exhibited the highest training and validation scores at 100% and 89.9 respectively.

The performance of the Random Forests ML algorithm is depicted through a confusion matrix with a sensitivity of 98.6% and a specificity of 86.47%. Notably, only 1.10% of false positive (FP) values were recorded. Additionally, the ROC analysis conducted on only 18 RT-plans using the Random Forests ML algorithm resulted in an AUC value of 0.97.

While achieving positive outcomes, the application of general ML models to specific tumor locations did not provide acceptable results for clinical implementation. For instance, utilizing the Random Forests Classifier model for brain tumor radiation therapy plans results in a 4.49% false positive rate and a specificity of 30%. Consequently, it became necessary to develop ML models to each tumor location to enhance overall model performance.

3.2 Machine Learning Models: Prediction for Brain and Thorax Tumor Location

ML models to specific tumor locations outperformed the model developed for all tumor locations. In the case of brain tumor RT plans, a perfect prediction was achieved with zero false positives, leading to an AUC of one. The accuracy and specificity were both 100%. Similar outcomes were observed when employing the specific ML model for thorax tumor RT plans. Nevertheless, for RT plans with greater complexity, such as those for the pelvis, breast and head and neck (H&N) tumors, standard ML models did not exhibit sufficient accuracy. For example, ML models for breast tumor RT plans demonstrated a specificity of only 87%. Consequently, more AI models, using deep hybrid learning (DHL), were specifically developed for pelvis, breast and H&N.

3.3 Deep Hybrid Learning: Prediction for Pelvis, Breast and H&N Tumor Location

The DHL models presented highly accurate results. For instance, the DHL model designed for breast cancer significantly decreased the number of FP from 9 with ML to 0. The enhancement elevated the specificity to 97.7% compared to the 87% achieved with the ML only model.

Semblable outcomes were observed with DHL models designed for H&N tumor locations (FN from 5.88% with the standard ML model to 1.47% with the DHL model).

3.4 Clinical Practice Solution

The requirement of this study was to create a simple approach that could be implemented in clinical practice. Consequently, an application programming interface (API) was developed which can easily be used in a clinical environment. The API is accessible at: https://aureliencd-radiotherapy-quality-api-cq-patient-predictor-2clxve.streamlitapp.com/. The performance of the API has been exclusively optimized on out dataset and need to be validation before external application. The average time (in minutes) required for computing the complexity indexes and predicting conformance using the platform was five minutes per plan. When focusing just on plans with a non-conformance predicting, the time saved per week amounts to approximately 135 minutes compared to the typical 150 minutes required for patient-specific QA. This time efficiency potentially allows for the treatment of seven additional patients.

4 Discussion

Patient-specific QC is mandatory in radiotherapy, but it can be time-consuming and affect the availability of RT machines to treat patients [17]. To address this issue, we propose to use prediction models that use RT plan complexity indices as input to predict QC outcomes while ensuring treatment safety [2].

VMAT is a more complex technique compared to intensity-modulated radiation therapy (IMRT). Therefore, the development of a predictive solution for QA is of high importance. To the best of our knowledge, our study is the first to use the innovative deep hybrid learning approach for the prediction of QA on a patient-by-patient basis. Referring to the literature, Kimura and colleagues developed a deep learning solution for QC of prostate plans and obtained a specificity value of 0.986 [18], which is consistent with our study (specificity value for pelvic tumor location=1). However, no studies have been published for complex VMAT dosimetry plans (breast or H&N).

To speed up the process of patient quality assurance in RT, a significant number of studies have developed AI models. However, the number of studies with clinical implementation of the models has been very small. In one study, a clinical implementation was performed, but the results obtained for the AUC were at best 0.869 [19]. Our first goal in this study was to develop an AI solution for QC in VMAT. This solution should be easy to use for everyone and should be implemented in clinical routine. For this reason, we used only six complexity indices in the modeling, which can be easily obtained during the dosimetry step. At the same time, we packaged the solution in an API for easy use by everyone.

Prediction of QA for IMRT breast tumor treatment using Halcyon machine

We also develop a model for new RT machines such as Halcyon can deliver a maximum dose rate of 600 monitor units per minute, allowing an increase in the number of patients treated per day. However, patient-specific QA is still required to ensure treatment safety, which dramatically decreases machine availability for patients. Innovative AI algorithms could predict the QA results immediately after dosimetry planning, based on complexity metrics to be used have not been definitively determined. The aim of this study was to develop an AI solution capable of firstly determining the complexity indexes to be obtained and secondly predicting patient-specific QA in routine clinical setting. Three hundred and eighteen beams from 56 patients with breast cancer were used in this study. Seven complexity indexes including Modulation Complexity Score (MCS), Small Aperture Score (SAS10), Beam Area (BA), Beam Irregularity (BI), Beam Modulation (BM), Gantry and Collimator angles were used as input to the AI model. ML and DL models using artificial neural network with TensorFlow and Keras were set up to predict the classification problem with conforming and non-conforming QC.

Statistical analysis showed that MCS, BI, gantry and collimator angle are not significantly correlated with QC compliance. Therefore, ML and DL models were trained using SAS10, BA and BM complexity indices. ML models did not show satisfactory performance with an AUC of 0.87 and specificity and sensitivity of 0.89 and 0.84. However, the optimized DL model showed better performance with an AUC of 0.96 and specificity and sensitivity of 0.98 and 0.92. Furthermore, DL model have been able to predict QA conformity in five cases of dose difference/distance to agreement: 2.5%/2.5mm, 3%/3mm, 3%/2mm, 2%/3mm, 2%/2.5mm. As for the first study with VMAT, API have been developed for whom which are interested to use the DL model for IMRT Halcyon QA prediction: https://aureliencd-halcyonqualitycontrolapi-ba47xul4ukaxjrwn4vxp5s.streamlit.app

5 Conclusion

Patient-specific QA requires time on the treatment machine, during which time it is impossible to treat patients. Therefore, the objective of this study was to propose an alternative solution for patient-specific QA that would make treatment machines more available to patients. To predict patient-specific QA compliance for treatments, we developed AI models based on complexity indices. In a busy RT department, the AI solution was used to automate patient-specific assurance. The results were conclusive for the VMAT in Rapid Arc machine as well as for IMRT in Halcyon machine. To implement this study in clinical routine, APIs were developed.

References

[1] K. Otto, "Volumetric modulated arc therapy: IMRT in a single gantry arc.," *Med Phys*, vol. 35, no. 1, pp. 310–7, Jan. 2008, doi: 10.1118/1.2818738.

[2] M. Antoine *et al.*, "Use of metrics to quantify IMRT and VMAT treatment plan complexity: A systematic review and perspectives.," *Phys Med*, vol. 64, pp. 98–108, Aug. 2019, doi: 10.1016/j.ejmp.2019.05.024.

[3] D. A. Low, W. B. Harms, S. Mutic, and J. A. Purdy, "A technique for the quantitative evaluation of dose distributions.," *Med Phys*, vol. 25, no. 5, pp. 656–61, May 1998, doi: 10.1118/1.598248.

[4] Y. Ge and Q. J. Wu, "Knowledge-based planning for intensity-modulated radiation therapy: A review of data-driven approaches.," *Med Phys*, vol. 46, no. 6, pp. 2760–2775, Jun. 2019, doi: 10.1002/mp.13526.

[5] S. Chiavassa, I. Bessieres, M. Edouard, M. Mathot, and A. Moignier, "Complexity metrics for IMRT and VMAT plans: a review of current literature and applications.," *Br J Radiol*, vol. 92, no. 1102, p. 20190270, Oct. 2019, doi: 10.1259/bjr.20190270.

[6] C. Shen *et al.*, "Clinical experience on patient-specific quality assurance for CBCT-based online adaptive treatment plan.," *J Appl Clin Med Phys*, vol. 24, no. 4, p. e13918, Apr. 2023, doi: 10.1002/acm2.13918.

[7] R. R. Savjani *et al.*, "A Framework for Sharing Radiation Dose Distribution Maps in the Electronic Medical Record for Improving Multidisciplinary Patient Management.," *Radiol Imaging Cancer*, vol. 3, no. 2, p. e200075, Mar. 2021, doi: 10.1148/rycan.2021200075.

[8] B. Sadowski, K. Milewska, and J. Ginter, "Machine Learning Based Prediction of Gamma Passing Rate for VMAT Radiotherapy Plans.," *J Pers Med*, vol. 12, no. 12, Dec. 2022, doi: 10.3390/jpm12122071.

[9] Z. L. Jazouli, S. Muraro, and D. Julian, "Patient-QA prediction: a new approach of complexity indexes," *Physica Medica*, vol. 92, p. S87, Dec. 2021, doi: 10.1016/S1120-1797(22)00184-3.

[10] M. F. Chan, A. Witztum, and G. Valdes, "Integration of AI and Machine Learning in Radiotherapy QA.," *Front Artif Intell*, vol. 3, p. 577620, 2020, doi: 10.3389/frai.2020.577620.

[11] D. A. Granville, J. G. Sutherland, J. G. Belec, and D. J. La Russa, "Predicting VMAT patient-specific QA results using a support vector classifier trained on treatment plan characteristics and linac QC metrics.," *Phys Med Biol*, vol. 64, no. 9, p. 095017, Apr. 2019, doi: 10.1088/1361-6560/ab142e.

[12] H. Hirashima *et al.*, "Improvement of prediction and classification performance for gamma passing rate by using plan complexity and dosiomics features.," *Radiother Oncol*, vol. 153, pp. 250–257, Dec. 2020, doi: 10.1016/j.radonc.2020.07.031.

[13] L. Wang *et al.*, "Multi-task autoencoder based classification-regression model for patient-specific VMAT QA.," *Phys Med Biol*, vol. 65, no. 23, p. 235023, Nov. 2020, doi: 10.1088/1361-6560/abb31c.

[14] W. Zouch *et al.*, "Detection of COVID-19 from CT and Chest X-ray Images Using Deep Learning Models.," *Ann Biomed Eng*, vol. 50, no. 7, pp. 825–835, Jul. 2022, doi: 10.1007/s10439-022-02958-5.

[15] J. D. Hunter, "Matplotlib: A 2D Graphics Environment," *Comput Sci Eng*, vol. 9, no. 3, pp. 90–95, 2007, doi: 10.1109/MCSE.2007.55.

[16] M. Waskom, "seaborn: statistical data visualization," *J Open Source Softw*, vol. 6, no. 60, p. 3021, Apr. 2021, doi: 10.21105/joss.03021.

[17] N. Agazaryan, T. D. Solberg, and J. J. DeMarco, "Patient specific quality assurance for the delivery of intensity modulated radiotherapy," *J Appl Clin Med Phys*, vol. 4, no. 1, pp. 40–50, Dec. 2003, doi: 10.1120/jacmp.v4i1.2540.

[18] Y. Kimura, N. Kadoya, S. Tomori, Y. Oku, and K. Jingu, "Error detection using a convolutional neural network with dose difference maps in patient-specific quality assurance for volumetric modulated arc therapy.," *Phys Med*, vol. 73, pp. 57–64, May 2020, doi: 10.1016/j.ejmp.2020.03.022.

[19] R. Yang *et al.*, "Commissioning and clinical implementation of an Autoencoder based Classification-Regression model for VMAT patient-specific QA in a multi-institution scenario.," *Radiother Oncol*, vol. 161, pp. 230–240, Aug. 2021, doi: 10.1016/j.radonc.2021.06.024.

.

Analysis of a decade's worth of oncology information system data from a large US healthcare system

Christian Velten[1,2], Ping Yan[1,2], Madhur K. Garg[1], and Wolfgang A. Tomé[1,2]

[1]Department of Radiation Oncology, Montefiore Medical Center, Bronx, USA
[2]Institute for Onco-Physics, Albert Einstein College of Medicine, Bronx, USA

Abstract Over a decade of oncology information system data were analyzed to investigate practice patterns and changes with respect to the treatment planning process across several sites. Evaluated metrics were physician conformity to submit target contours and prescriptions, available time for treatment planning, including quality assurance and physics checks, overall development of the number of plans treated, and treatment planning and physics check workload. Changes in practice patterns, e.g. the impact of the COVID-19 pandemic and its variation in the type of treatments performed at different sites, as well as the increasing number of patient plans treated with decreasing time available was identified. Interventions e.g., increasing time allotments and the implementation of a treatment planning progress monitoring system were identified in the data. Leveraging big data can generate insights and drive pro-active over reactive workflow changes.

1 Introduction

Radiation oncology practices rely on the collaboration of an interdisciplinary team to treat patients effectively and safely in a timely fashion. Between treatment simulation up to the delivery of radiotherapy physicians and physicists – in some countries supported by dosimetrists – work mostly in sequence to generate treatment plans using computerized treatment planning systems.

Treatment planning processes and workflows have been evolving over time to adjust to new technologies, organizational changes, and incrementally to improve efficiency and alleviate insufficiencies, actual or perceived. These changes are usually slow, making it difficult to measure their efficacy independent of short-term environmental changes. Moreover, changes may be initiated by temporary or short-term issues or subjective anecdotal evidence, ultimately failing to achieve their intended goal or even be detrimental to the overall evolution of the department's processes.

Before the advent of computerized oncology information systems (OIS), analyses of workflows and processes had been cumbersome and slow. Widespread implementations of OIS have enabled efficient aggregating of data relating to physician and physics tasking, patient scheduling, and many more over time frames of individual days or many years.

Our Medical Center provides radiation oncology services across six different sites, both hospital-based and freestanding, with four sites being integrated into the same OIS database. Of these, one hospital-based and freestanding site treat the majority of patients, each using two linear accelerators (one C arm, one ring gantry based). The other two sites treat an average of ten to fifteen patients per day using one C arm linac each.

Treatment planning is initiated by physicians by providing target and critical organs-at-risk (OAR) definitions, radiation dose prescriptions, and acceptable doses to OARs. External beam radiotherapy planning is centralized for the integrated sites and was performed by an on-site staff of six dosimetrists spread over the two main sites until the COVID-19 pandemic. Since then, dosimetry staff are primarily remote with staffing levels augmented through third party contracts. Towards the end of the planning process, treatment plans are submitted to be reviewed by

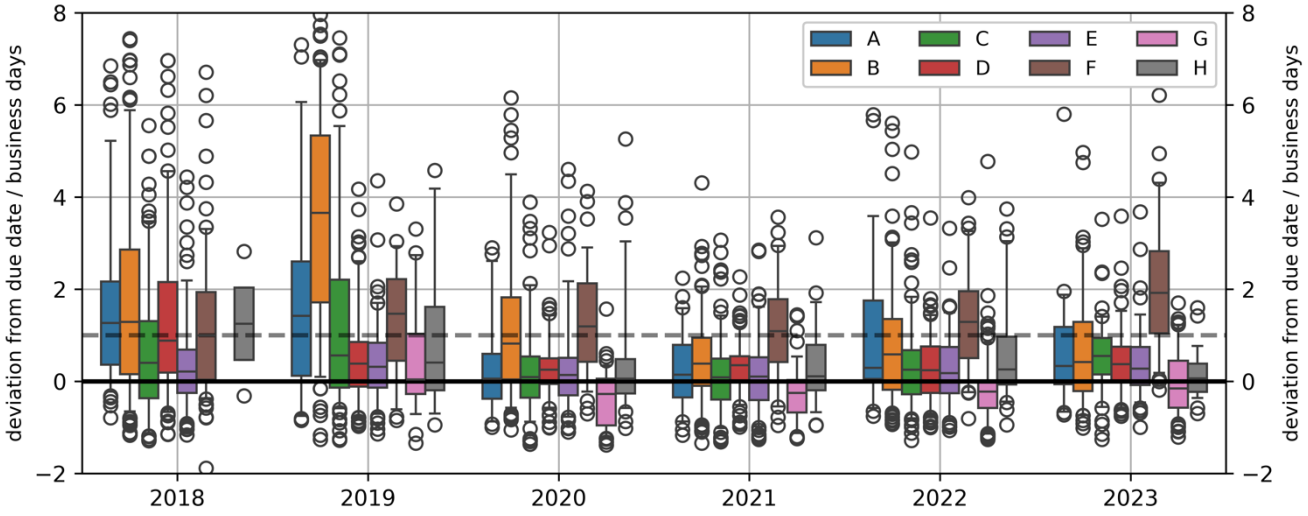


Figure 1. Deviations from physician input due date in business days for eight attending radiation oncologists over six years. Whiskers correspond to the 90% confidence interval.

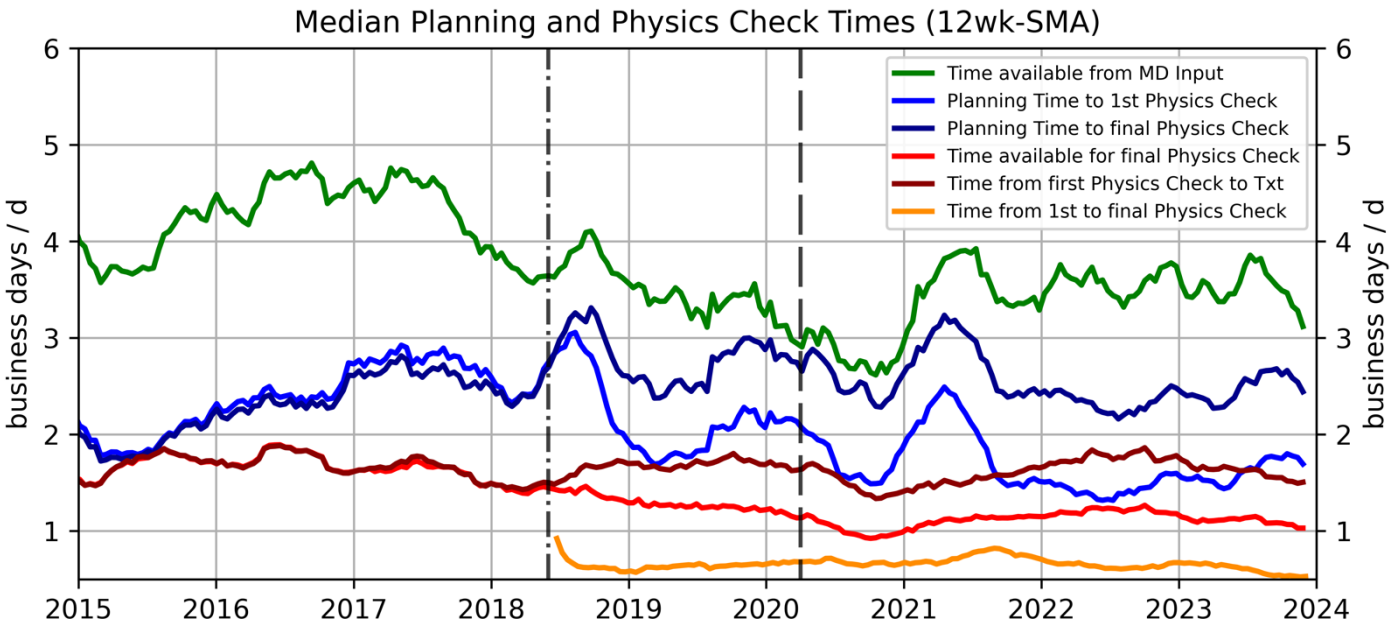


Figure 2. Evolution of planning and physics check times over nine years. In June 2018 physics checks before physician review (pre check) were implemented. Beginning with April 2020 (these time points are indicated by vertical lines), New York City was in a COVID-19 shutdown.

physicists to assess overall plan quality, safety, and consistency with the physician prescription.

We investigated treatment planning practice patterns across integrated facilities to evaluate treatment planning workloads, bottlenecks, differences among departments and staff groups, and the impact of planning interventions.

2 Materials and Methods

Physician and physics tasks, as well as patient appointments were queried from the Varian ARIA OIS database using SQL queries. Processing and analyses were performed using Python (pandas, seaborn, scipy). Dates and times were adjusted for US business days and hours (08:00 to 18:00) excluding federal holidays and weekends.
Physician conformance was evaluated as the deviation from physician input due date (including early completion) in business days. Boxplot whiskers extend from the 5% to 95% quantiles. Workload for the physics group, composed of dosimetrists and physicists is evaluated as the ratio of number of tasks and the average time available for completion in the same time span. Time available to complete a task is assumed to be inversely proportional with stress associated with its completion and hence task intensity. With this our definition follows the US Centers of Medicare and Medicaid Services and American Association of Physicists in Medicine Abt IV report definition for workload.

3 Results

Physician conformance is shown in figure 1 with the deviation from due date distributions from 2018 through 2023. From 2018 to 2019 several physicians consistently exceeded their allotted time by at least one day with one physician (B) reaching a median of almost four days in 2019 with 90% of all submissions being late and at least 25% overdue by more than 5 days. Coincident with the NCOV-19 pandemic in 2020, physician inputs became less overdue with the majority now submitted up to one day late. Since then, delays have increased slightly across all staff members. Notably, physician F's inputs did not decrease substantially during 2020 and have become increasingly late from a median of 1 day in 2021 to 1.5 days in 2022, and approximately 2 days in 2023.

Time series of available planning and physics check times are shown in figure 2. The available time for planning (green line) decreased from 4-5 days during 2016-17 to 3 days during 2019, extending into most of 2020, and has since increased to on average 3.5 days. With total planning time (dark blue line) remaining somewhat constant during that time, the time available to perform QA and perform physics checks decreased from a median of 1.5-1.9 days to 1 day in mid-year 2020. With the increase in available planning time in early 2021 both utilized planning time and the time available for QA and physics checks increased slightly.

The total number of plan check tasks performed across the OIS-integrated sites is shown in figure 3. Few tasks were

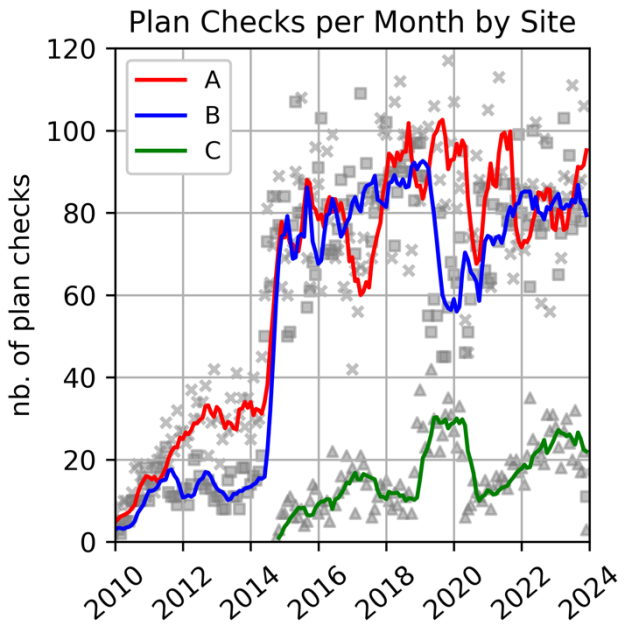


Figure 3. Number of physics plan check tasks submitted per month by department. C includes all secondary sites treating few patients cumulatively.

completed before 2015 with a total of approximately 50 per month after which the number of monthly plan checks jumped to approximately 75 for sites A and B. From 2015 through 2019 numbers increased to approximately 90 per month. Site-specific decreases in 2017 (A) and 2019 (B) of about 20-30 per month were compensated by increases in numbers for the other sites. All sites experienced a sharp decrease in monthly plan checks during the first half of 2020, from which A recovered rapidly within a few months with the recovery at sites B and C somewhat slower. The number of monthly plan checks at the end of 2023 has reached levels previously seen in 2018.

Planning and physics workloads are shown in figure 4. Planning workloads reached a minimum of 2.5 in 2017 before increasing in 2018 and varying between 3.5 to 5.5 since then. During 2022 planning workloads remained somewhat constant just above 5 and have been increasing since the beginning of 2023. Physics workloads reached a minimum of 5 in 2017 and have generally been varying substantially by up to a factor of two between 7.5 and 15. Since the beginning of 2023, physics workload again doubled from 10 to 20.

4 Discussion

Decreases in available planning times from 4-5 days during 2016-17 to 3 days during 2019 and extending into 2020 are consistent with physician input delays of one day or more (Fig. 1 & 2). Substantial improvements in physician conformance were observed for 2020 (Fig. 1), driven in part by the substantially decreased number of patients seen during the first three quarters that year due to the COVID-19 pandemic. Another factor was actuarial in that the scheduled time available to complete the physician input was increased from 3 to 5 days for most cases. During that year available planning time continuously deteriorated despite increased physician conformance (Fig. 2, green line). And while this did not translate to a substantial change in the total time utilized for planning (Fig. 2, dark blue line) compared to pre-pandemic levels it did decrease the time available for physics checks and QA times by almost 30%. This inconsistency between increased on-time submission and available time for planning is likely due to two factors: first, time-to-treatment was kept the same or did not increase as much as the time allotment for physician inputs increased, and second, physician tasks were completed despite no workable target structures or prescription available, circumventing the tasking system.

Available time for physics checks and QA was also reduced, perhaps unintendedly, by the implementation of a pre-physician review physics check (Physics Pre Plan Check) in 2018. Its intention is to catch planning errors and mistakes early in the process and to improve efficiency. However, most of these pre-checks are now being submitted at times that final checks used to be submitted. In cases where substantial changes had to be made due to planning errors or physician requests, this substantially increases the peak workload for physics staff.

With increasing patient numbers towards the end of 2020 and continuously deteriorating available planning time a planning progress monitoring system was implemented in November of that year. This system automatically sent out reminders about physician input due and overdue times. Almost immediately, the median available planning time increased and has since remained at around 3.5 days. The

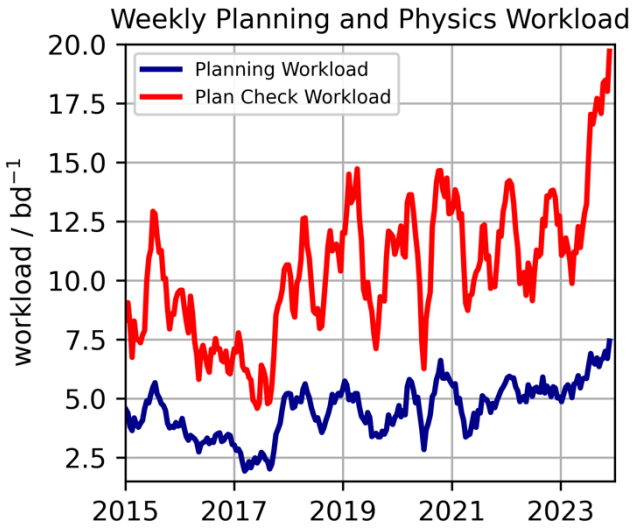


Figure 4. Weekly planning and physics workload measured as ratio of number of tasks to average time allotted for task completion.

time available for final physics checks had similarly recovered to 2019 levels by mid-year 2022.

Since mid-2022 actual planning times have consistently increased, while physics times for both pre and final checks have decreased by the same amount. During that time an automated whiteboard was implemented in 2022 to aid in reducing late plan submissions, however, it appears like it may have had the opposite effect. One potential reason is that dosimetrists may not be assigned to plans more than two days prior to scheduled start or dosimetrists may be delaying plan generation given the increased workload levels since 2022 (Fig. 4). This suggests that an intervention is warranted to identify cases where planning has not been initiated within a certain time of physician input. The need for such an intervention is even more evident given the sharp increase in physics workload during 2023 from 10-12 at the beginning of the year to almost 20 by December. This increase can only in part be explained by an approximately 20% increase in plans at site A (Fig. 3) where turnaround times are usually less than 3 days.

5 Conclusion

A centralized OIS database was leveraged to query and aggregate physician, dosimetry, and physics task data from more than 10 years to give insights into practice patterns and their evolution. Internal (machine replacements) and external (COVID-19 pandemic) impacts could clearly be identified. Correlating available task times and workloads can be used to inform and substantiate changes to workflows to increase efficiency.

Fast many-scenario robust IMPT planning through constrained dose-mimicking

Franziska Knuth[1], Michelle Oud[1], Wens Kong[1], Hazem Nomer[1], Joep van Genderingen[1], Linda Rossi[1], Steven Habraken[2,3], Ben Heijmen[1], and Sebastiaan Breedveld[1]

[1] Department of Radiotherapy, Erasmus MC Cancer Institute, University Medical Center Rotterdam, Rotterdam, the Netherlands
[2] HollandPTC, Delft, the Netherlands
[3]Department of Radiation Oncology, Leiden University Medical Center, Leiden, the Netherlands

Abstract
Patient setup variations and proton range uncertainties need to be taken into account in robust treatment planning for intensity modulated proton therapy (IMPT). In this study, a two-step approach is proposed that aims at fast generation of IMPT plans that are robust for many robustness scenarios. In the first step, automatic multi-criterial optimization (MCO) with Erasmus-iCycle is used for generation of a plan that is robust for 13 scenarios. In the following dose mimicking step, this plan is converted into a plan that takes into account 53 scenarios. The potential of this approach is demonstrated for 14 oropharyngeal cancer patients. Compared to the initial plans generated with the 13 scenarios, the converted plans show much improved target coverage robustness for the final 53 scenarios, similar to plans directly generated with 53 scenario based automatic MCO. The two-step approach with dose mimicking was also a factor 10.7 faster than direct automatic MCO, with dose mimicking times of on average 1.4 min. The proposed dose-mimicking is applicable to various treatment sites, providing a practical solution for enhancing IMPT plan robustness with significant time savings compared to direct robust optimization.

1 Introduction

Intensity modulated proton therapy (IMPT) is susceptible to patient setup variations and proton range uncertainties. To account for this, treatment planning is performed using a set of scenarios, i.e. patient shifts and range variations, to ensure adequate CTV coverage and adherence to maximum doses for critical organs at risk (OARs). However, optimization times increase with the number of included scenarios. In clinical practice, 21 scenarios are often used for planning, with more scenarios desirable for more accurate modeling of the impact of variations and uncertainties. For fast robust optimization with a large set of scenarios, we propose a two-step approach: 1) automated multi-criteria treatment planning using a small set of scenarios, 2) fast constrained dose-mimicking to make the plan robust for a much larger set of scenarios.

2 Materials and Methods

For 14 patients with oropharyngeal cancer, Erasmus-iCycle with large scale sparsity-induced spot selection (SISS) [1–3] was used to automatically generate two robust IMPT plans. The first, with only 13 robustness scenarios (iCycle13), was used as input for the proposed dose mimicking (details below), aiming at fast generation of the Mimic53 plan that should be robust for 53 uncertainty scenarios. The second Erasmus-iCycle plan (iCycle53) was compared to the Mimic53 plan for validation of the dose mimicking. Setup- and range robustness settings of 3 mm/3% were used. The scenarios were distributed on a unit sphere following Korevaar et. al. [4], where the 13 scenarios were a subset of the 53. Figure 1 illustrates the distribution of setup errors. Doses of 70 GyE and 54.25 GyE were prescribed to the primary and elective CTV, respectively. All three plan generations were constrained for adherence to minimum and maximum MU constraints and maximum number of energy layers, following the requirements in our clinic.

In the dose mimicking, a quadratic penalty function was minimized aiming at updated iCycle13 spot weights that closely reproduced all voxel doses of iCycle13 in the 13 scenarios, while enforcing obedience to the clinical CTV voxel-wise minimum dose constraints for all 53 scenarios.

Mimic53 plans were compared to ground truth iCycle53 plans and to the input iCycle13 plans. All plans were evaluated on 53 scenarios, via nominal, voxelwise minimum (VMin) and voxelwise maximum (VMax) dose distributions. The two-sided Wilcoxon signed-rank test was used to evaluate differences in dose-volume-histogram (DVH) metrics with a significance level of 0.05.

3 Results

Figure 2 compares plan parameters of iCycle13 and Mimic53 with the highly robust gold truth plans, iCycle53. Consid-

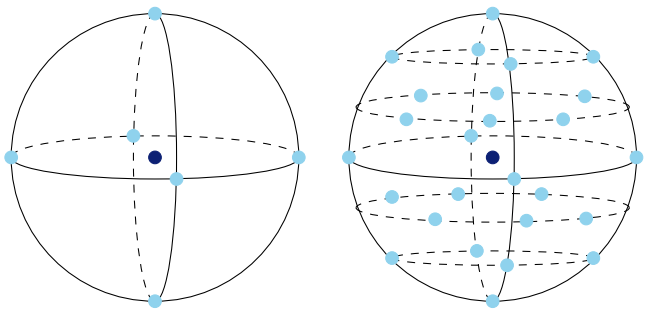


Figure 1: Illustration of setup uncertainties for 13 (left) and 53 (right) scenarios. For each setup scenario case (light blue), the positive and negative range error case is included, with exception of the nominal scenario (dark blue).

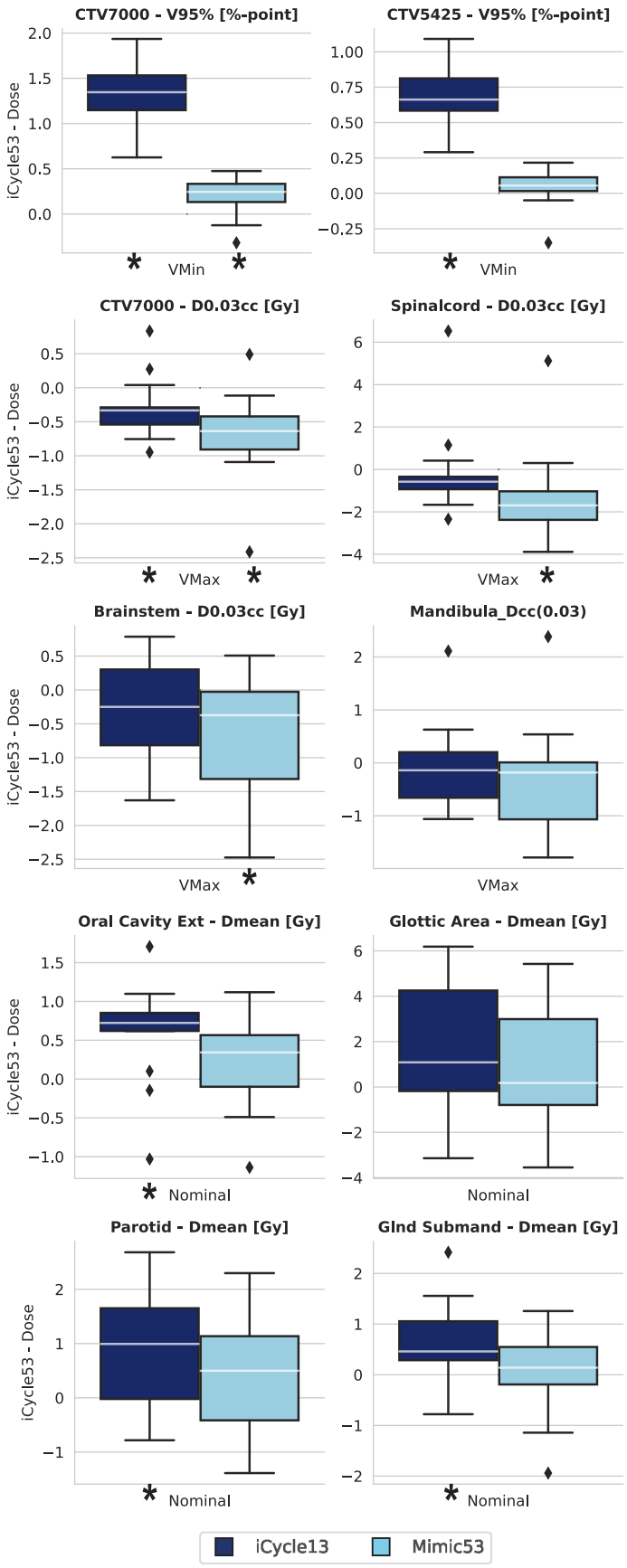


Figure 2: Boxplots showing the differences between both iCycle13 and Mimic53 to iCycle53, where all plans were evaluated on 53 scenarios. In the boxplots, the box indicates the median and interquartile range. Outliers are outside the 1.5 the interquartile range. Significant differences as denoted by the wilcoxon signed-rank test are indicated as *: p<0.05.

Table 1: Computation times (min) for the different plans

Step	iCycle13	Mimic53	iCycle53
PBC for SISS	54±8		869±118
SISS MCO	24±6		115±32
PBC for Mimicking†	-	12±3	-
Mimicking	-	1.4±0.3	-
Total	78±10	92±11	984±116

PBC: Pencil beam calculation; SISS: Sparsity-induced spot selection; MCO: Multi criteria optimization; †: Computation of pencil-beams in the 40 new scenarios.

ering all 53 robustness scenarios, iCycle13 plans show significant VMmin $V_{95\%}$ losses for both high (p=0.0001) and low (p=0.0001) dose CTV compared to iCycle53, with median differences of 1.3%-points and 0.66%-points, respectively. For the Mimic53 plans, these differences to the gold truth were reduced to 0.25%-points for CTV 7000 (p=0.009) and 0.055%-points for CTV 5425 with no statistical significance. Compared to iCycle13, mimicking also reduced differences with iCycle53 in nominal mean OAR doses (Dmean), eliminating the significant differences for iCycle13. For the parotids, the difference with iCycle53 in median Dmean reduced from 0.99 Gy (p=0.013) in iCycle13 to 0.5 Gy (p>0.05) in Mimic53. The same trend can be observed for the median Dmean of submandibular glands and the extended oral cavity. The difference in median Dmean also decreased for the glottic area, as well as the VMax near maximum doses (D0.03cc) of the mandibula, while the difference was already not significant for iCycle13. Differences in VMax D0.03cc for brainstem and spinal cord were slightly higher for Mimic53, compared to iCycle13, however well within the hard constraint of 35 GyE for all plans. Table 1 summarizes the optimization times in minutes for the different plans as well as the times for the spot dose calculations. Overall, the speed-up factor for the dose mimicking approach (Mimic53) compared to the direct highly robust optimization (iCycle53) was 10.7 including initial pencil beam dose calculations. Excluding times for pencil beam calculations, generation of a gold truth iCycle53 plan took 115 min, which reduced to 25.4 min for generating the Mimic53 plan. Dose mimicking to convert an iCycle13 plan into a Mimic53 plan only took 1.4 min.

4 Discussion

When robustly evaluated for 53 scenarios, iCycle13 plans, that were robustly generated considering only 13 scenarios, showed substantial plan parameter deviations from ground truth iCycle53, especially for target coverage. With the proposed dose mimicking, an iCycle13 plan could in 1.4 min be converted into a Mimic53 plan with no coverage differences with gold truth. This study includes only 14 patients, which

is a limitation, even though the proposed dose mimicking approach does not require fine tuning of parameters. The presented mimicking approach is not limited to the presented treatment site and prescription, and can be applied without special adjustments to other treatment sites or dose prescriptions. Oropharyngeal cancer cases were chosen as demonstration because of the challenging anatomy with many OARs in close proximity to the targets. The presented approach can also be easily extended to an even larger set of scenarios.

5 Conclusion

We have introduced a dose-mimicking approach to obtain IMPT plans which are robust for a large number of scenarios (53) with computational costs similar to full robust optimization for a low number of scenarios (13). The additional time for dose mimicking took on average 1.4 min.

References

[1] S. Breedveld, P. R. Storchi, P. W. Voet, et al. "iCycle: Integrated, multicriterial beam angle, and profile optimization for generation of coplanar and noncoplanar IMRT plans". *Med Phys* 39.2 (2012), pp. 951–63. DOI: 10.1118/1.3676689.

[2] W. Kong, M. Oud, S. Habraken, et al. "SISS-MCO: large scale sparsity-induced spot selection for fast and fully-automated robust multi-criteria optimisation of proton plans". *Phys Med Biol* (2024). DOI: 10.1088/1361-6560/ad1e7a.

[3] S. Breedveld, P. R. Storchi, and B. J. Heijmen. "The equivalence of multi-criteria methods for radiotherapy plan optimization". *Phys Med Biol* 54.23 (2009), pp. 7199–209. DOI: 10.1088/0031-9155/54/23/011.

[4] E. W. Korevaar, S. J. M. Habraken, D. Scandurra, et al. "Practical robustness evaluation in radiotherapy - A photon and proton-proof alternative to PTV-based plan evaluation". *Radiother Oncol* 141 (2019), pp. 267–274. DOI: 10.1016/j.radonc.2019.08.005.

DeepTuning: Tune Radiotherapy Plan Dose Distribution in Real Time by Deep Learning

Lin Ma[1], Huan Liu[2], Mingli Chen[3] and Weiguo Lu[3]

[1]Department of Radiation Oncology, University of Pennsylvania, Philadelphia, Pennsylvania, USA
[2]Department of Radiation Oncology, Mayo Clinic, Rochester, Minnesota, USA
[3]Department of Radiation Oncology, University of Texas Southwestern Medical Center, Dallas, Texas, USA

Abstract Radiotherapy treatment planning is a teamwork collaborated by physicians, dosimetrists and physicists. Physicians inform dosimetrists regarding their desired dose distributions and trade-offs, while dosimetrists provide physicians with the best plans can be achieved. It often requires multiple rounds of consultation between physicians and dosimetrists to tune plans and select one with desired trade-off. The back-and-forth between different parties prolongs planning process. To reduce the back-and-forth resulting from trade-off exploration, we proposed a new framework called DeepTuning. The framework utilized a deep learning model to predict spatial dose distributions with different trade-offs given a set of contours. The predicted dose distribution can be tuned in real time by manipulating the deepest layer of the model. We validated this method with a dataset consisting of prostate VMAT plans. The model can generate plans ranging from prioritizing PTV coverage to prioritizing rectum sparing. The DeepTuning framework has the potential to empower physicians to tune dose distributions immediately after they complete contours. A desired dose distribution can be sent to dosimetrists to inform dosimetry planning. As the trade-off selection occurs before dosimetry planning, the back-and-forth can be minimized, promoting the efficiency of treatment planning workflow.

1 Introduction

Radiotherapy treatment planning is a teamwork collaborated by physicians, dosimetrists and physicists. Physicians prescribe treatment dose to target and place constraints for normal organs. Dosimetrists provide feasible solutions to satisfy physician's orders by optimizing plans in treatment planning system (TPS). It often requires multiple rounds of consultation between physicians and dosimetrists to tune plans and explore trade-offs before plan approval. The back-and-forth between different parties prolongs planning process, because their schedules are not aligned.

Multi-criteria optimization (MCO) has been developed to allow real-time trade-off exploration and reduce the off-line back-and-forth. However, it wasn't adopted widely in the clinic because of its nature of being an optimization method. The optimization time is longer than usual planning, but accuracy is limited. Deep learning (DL) has shown promise in predicting dose distribution directly from contours without optimization [ref 1]. However, its prediction is fixed rather than adjustable. Physicians cannot tune the predicted dose and explore trade-off with DL tools.

We proposed a "DeepTuning" framework to reduce the back-and-forth in treatment planning. It is a DL model that can predict a distribution of dose distributions with different trade-offs from just contours. Users can tune the predicted dose in real time. It combines the advantages of DL and MCO because it can derive a dose in real-time and the doses can be adjusted for different trade-offs.

2 Materials and Methods

Theory: VMAT dose distributions are determined by two types of information: geometry and trade-off. On the first order, geometry information (contour location, shape and size) dictates the dose distribution because VMAT doses conform to the targets and carve around OAR contours. Furthermore, contour-to-dose is not a one-to-one mapping. Each patient can have multiple plans with different trade-offs, for example, between target coverage and OAR sparing. Therefore, on the second order, the trade-off information also affect dose distributions. Geometry information dominates dose distributions, while trade-off information represents dose tuning, whose influence is lesser.

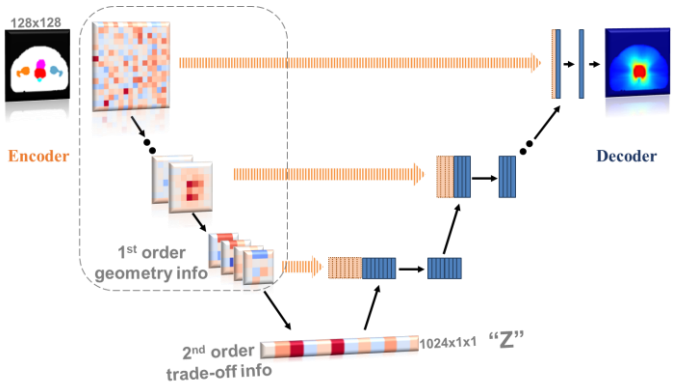


Figure 1. Network structure of a U-Net for dose prediction. It is composed of an encoder and a decoder. The feature maps at different depth encode geometry information and trade-off information.

A deep learning model (U-Net) takes contour as input and predict dose distribution. It comprised an encoder and a decoder. We hypothesized that the geometry information were store in the feature maps in shallow layers while the trade-off information were stored in the deepest feature map, which has been down-sampled to the resolution of 1 by 1 with 1024 channels. Hereafter, we refer to the deepest feature map as "trade-off encoding vector" by symbol Z. Given patient contours, the goal of DeepTuning is to generate dose distributions with different trade-offs by manipulating trade-off encoding vector Z.

Data: The models were trained by an anonymized dataset comprising 99 prostate VMAT cases. The structures of each case are PTV, bladder, rectum, left and right femur head. The PTVs were expanded from intact prostate or prostate plus seminar vesicle. For each case, we retrospectively optimized two VMAT plans with our in-house optimizer. One plan prioritized PTV coverage (referred to as "pro-ptv" plan

below), while the other plan prioritized rectum sparing (referred to as "pro-oar" plan below). The two plans were optimized with same objectives (80Gy to PTV, etc) except that the coverages in the PTV-rectum overlap were relaxed to 77Gy and rectum doses were penalized harder in the "pro-oar" plans. For each case, the two plans constitute a set plans with different trade-offs.

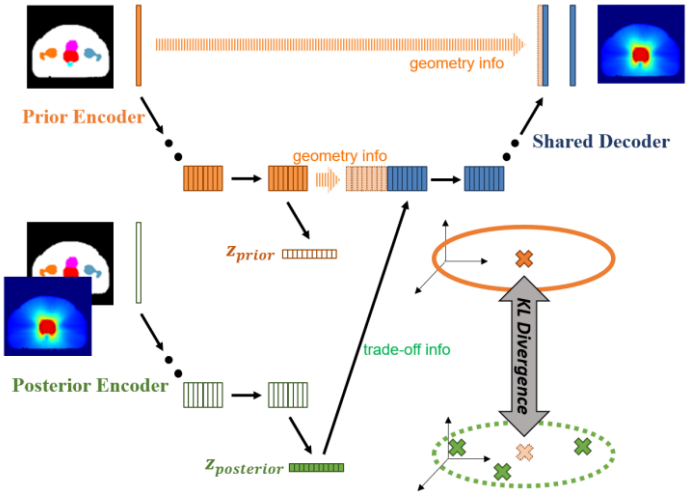


Figure 2. Training scheme for DeepTuning. Prior encoder (orange), posterior encoder (green) and shared decoder (blue) are trained together. The green crosses in the bottom right corner represent the Z_qs of different input doses. Orange cross represents the Z_p predicted by prior encoder, which is the average of Z_qs.

Model architecture: The DeepTuning framework has three components: prior encoder $\mathbb{E}_{prior}$, posterior encoder $\mathbb{E}_{post}$, and decoder $\mathbb{D}$. They serve for different purposes.

Prior encoder $\mathbb{E}_{prior}$ takes contour C as input. The shallow feature maps it generated represent geometry information G and the deepest feature map Z_p (p for prior) represents trade-off information. As contour is the only input, we expect Z_{prior} to encode a balanced trade-off, corresponding to the average of training dataset. An example of prior encoder is at the left side of figure 1.

$$G, Z_p = \mathbb{E}_{prior}(C)$$

Posterior encoder $\mathbb{E}_{post}$ takes contour C and dose D as input in two channels. It has the same structure as prior encoder except for the number of input channels at the first layer. We discard the shallow feature maps it generated and only use its deepest feature map Z_q (q for posterior). We expect Z_q to encode the trade-off of input dose.

$$\sim, Z_q = \mathbb{E}_{post}(C, D)$$

Decoder $\mathbb{D}$ can be used with a prior encoder alone to form a U-Net and predict dose distribution D. In this case, the inputs to $\mathbb{D}$ are the geometry information G and balanced trade-off Z_p extracted by encoder (from contour alone).

$$D = \mathbb{D}(G, Z_p)$$

Decoder $\mathbb{D}$ can also be used with both encoders. In this case, the inputs to $\mathbb{D}$ are the geometry information G extracted by encoder (from contour alone) and the trade-off Z_q extracted by posterior encoder from the input dose distribution.

$$D = \mathbb{D}(G, Z_q)$$

Model training: We trained the model with the prostate VMAT dataset described above. In each training epoch, we ran over all training cases. For each case, we randomly pick a plan, which could be pro-ptv or pro-oar. The contours were fed into prior encoder $\mathbb{E}_{prior}$ to extract geometry information G. The contour and the dose (pro-ptv or pro-oar) were fed into posterior encoder $\mathbb{E}_{post}$ to extract trade-off information Z_q. The shared decoder $\mathbb{D}$ combined the two information and predicted a dose. The three models were trained by minimizing two loss functions. First, we forced the output dose to be the same as input dose. Therefore, the posterior encoder can extract trade-off from input dose. Second, when posterior encoder was fed with plans with different trade-offs, it generated different Z_q(green points at the bottom right of figure 2). We forced the Z_p to be the average of the Z_qs by a KL divergence loss. Therefore, prior model can generate a Z_p corresponding to balanced trade-off, average of training plans. Please find training details in figure 2 and algorithm 1.

Algorithm 1: Model training

For i th case in training dataset

1. Get contours C^i. Randomly pick a plan (could be pro-ptv or pro-oar) and get its dose D^i
2. Feed contour C^i to prior encoder. Derive geometry information G and balanced trade-off encoding vector Z_p
$$G, Z_p = \mathbb{E}_{prior}(C^i)$$
3. Feed contour C^i and dose D^i to posterior encoder. Extract trade-off Z_q from input dose
$$\sim, Z_q = \mathbb{E}_{post}(C^i, D^i)$$
4. Feed geometry info G and the extracted trade-off info Z_q to decoder, which predicts dose $\widehat{D}^i$
$$\widehat{D}^i = \mathbb{D}(G, Z_q)$$
5. Calculate dose loss $||\widehat{D}^i - D^i||^2$
6. Calculate trade-off feature loss $KL\ Div(Z_p, Z_q)$
7. Optimize the parameters of models ($\mathbb{E}_{prior}$, $\mathbb{E}_{post}$ and $\mathbb{D}$) by minimizing the two losses

Model inference: After training, we obtained three models. The prior model can extract geometry information and derive Z_p representing balanced tradeoff from just contour. The posterior model can extract tradeoff from input dose distribution. The decoder can be used in conjunction with encoders to predict dose.

The application of models to new cases is consisted of two steps. First, extract desired trade-off ΔZ from template plans. Second, apply the desired trade-off ΔZ to new case and predict dose with desired trade-off.

In the first step, template plans refer to the plans of the patients treated before. The template plan has the desired trade-off (eg. trading a little PTV coverage for a lot OAR sparing). Physicians preferred this trade-off and wanted to replicate this trade-off in future cases. To extract trade-off from a template plan, we feed it into prior encoder and

posterior encoder separately, which generated a Z_p that represents balanced trade-off and a Z_q that encodes the trade-off of template plans. Their difference ΔZ can tune the trade-off of a dose distribution. Please find the details in Algorithm 2.

Algorithm 2: Extract trade-off from template plan

Input: Contours C and dose D of template plan with desired trade-off

Output: trade-off of template plan, ΔZ

Steps:

1. Calculate the balanced trade-off encoding vector Z_p by prior encoder
$$\sim, Z_p = \mathbb{E}_{prior}(C)$$
2. Calculate the trade-off Z_q of template plan dose by posterior decoder
$$\sim, Z_q = \mathbb{E}_{post}(C, D)$$
3. Extract trade-off of template plan by subtracting from balanced tradeoff
$$\Delta Z = Z_q - Z_p$$

In the second step, we use prior encoder to predict doses for new patients. When we feed just contour into prior encoder and run through decoder (step 1 and step 2 in Algorithm 3), it generates a dose with balanced trade-off, average of training dataset, functioning as a conventional dose prediction model. The novelty of DeepTuning roots in the next step. If we apply the trade-off ΔZ extracted to this new case, (step 3 in Algorithm 3), the output dose distribution changes, from balanced trade-off to the desired trade-off. Our DeepTuning framework tunes the trade-off of dose prediction by manipulating the trade-off encoding vector Z. Please find the details in Algorithm 3 and figure 3.

Algorithm 3: Predict doses for new case

Input: Contours C of the new case, the desired trade-off ΔZ extracted from template plan

Output: dose distribution for new case with desired tradeoff

Steps:

1. Feed contour C to prior encoder. Derive geometry information G and balanced trade-off encoding vector Z_p
$$G, Z_p = \mathbb{E}_{prior}(C)$$
2. Predict dose D with balanced trade-off by Z_p
$$D = \mathbb{D}(G, Z_p)$$
3. Or, predict dose D with desired trade-off by applying ΔZ
$$D = \mathbb{D}(G, Z_p + \Delta Z)$$

Training specifics: We split the dataset into training set (60 cases), validation set (10 cases) and testing set (29 cases). The models were trained by training set. The models in the iteration with lowest validation loss were selected as the final models. The DeepTuning project was implemented in Python language v3.7 and Pytorch deep learning framework v1.1.0. We trained the models by 200 epochs with Adam optimizer and a learning rate of 1e-4.

The contours and doses were resampled to an isotropic resolution of 0.25mm. The input and output are 128*128 2D patches, corresponding to 32cm*32cm, large enough to accommodate high and intermediate dose regions for dose prediction purpose.

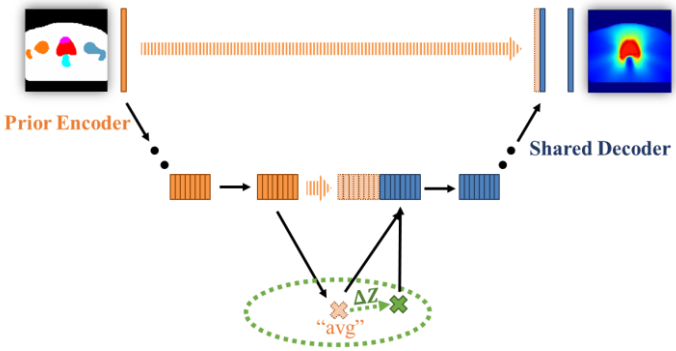


Figure 3. Prior model predicts dose for new case. It predicts average dose w/ balanced trade-off, or apply ΔZ to predict a dose with trade-off.

3 Results

In this section, we first evaluate the accuracy of posterior inference on testing set by comparing the output dose distribution to the dose input to the posterior encoder. This is a sanity check to validate that posterior encoder can extract the trade-off from input dose. Then we examine the prior inference on testing set by comparing the prior model prediction to the pre-optimized doses with trade-offs. The prior model should generate a plan with balanced trade-off, straddling between the pro-ptv plan and the pro-oar plan. Then we select a pro-oar plan from training set as "template plan", visualize the trade-off encoding vectors Z_p, Z_q and their difference ΔZ that can tune the trade-off. Finally, we apply the extracted ΔZ to testing set, to validate that our DeepTuning framework can rise above predicting average doses and succeed in predicting doses with the desired trade-off.

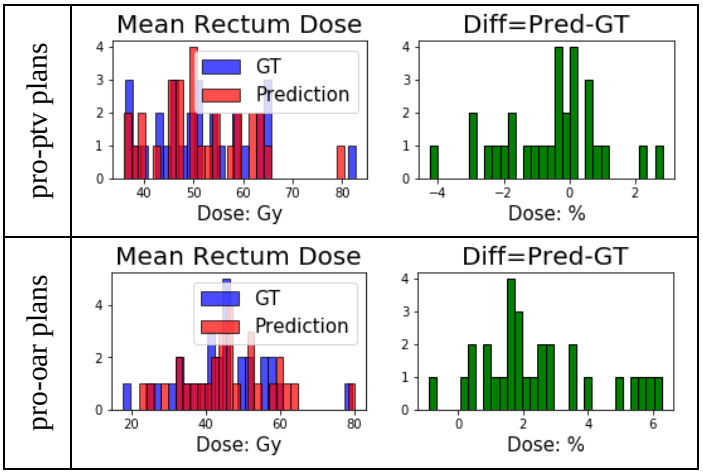


Figure 4. Histogram of predicted dose vs GT doses. It evaluates the accuracy of posterior model which replicates input dose distribution.

Posterior model inference: The results of posterior model in 29 testing cases were in Figure 4. The mean rectum doses in ground truth (GT) are 52.0±10.6Gy in pro-ptv plans and 44.2±12.2Gy in pro-oar plans. The predictions of posterior model match close to GT doses, with only differences of -0.6±1.5% and 2.4±1.8% in the two types of plans.

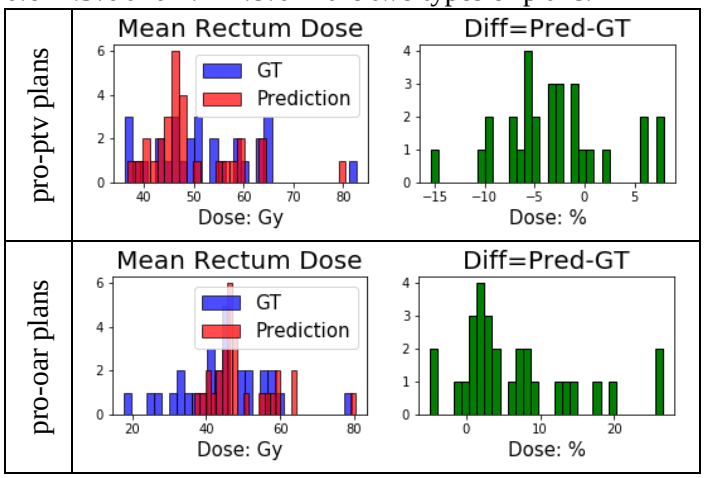


Figure 5. Histogram of predicted dose vs GT doses. It evaluates if prior model can predict an "average" dose with balanced trade-off.

Prior model inference: The results of prior model in 29 testing cases were in Figure 5. The prior model itself functions to predict an "average" dose with balanced trade-off. The mean rectum doses predicted by prior model are 49.5±9.0Gy, -3.1±5.3% colder than the pro-ptv GT plans and 6.6±7.9%colder than the pro-oar GT plans, which validated our hypothesis.

Trade-off extraction: We selected a case from training set and used its pro-ptv dose as template plan for pro-ptv trade-off and its pro-oar dose as template plan for pro-oar trade-off. The trade-off encoding vectors ΔZ were calculated for the two trade-offs. The Z_p for balanced trade-off (predicted by prior model), Z_q derived from pro-oar template plan and their difference ΔZ are shown in Figure 6.

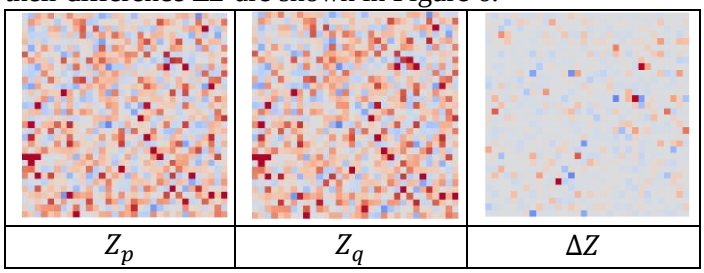


Figure 6. Example of the trade-off encoder vectors. Z_p is for balanced trade-off, Z_q here is for prioritizing rectum sparing, and ΔZ is their difference. Each Z is a vector with 1024 elements. We resize it into a 32x32 matrix for better visualization.

Apply trade-off to prior model inference: We applied the extracted trade-off ΔZ prioritizing PTV coverage and another ΔZ prioritizing OAR sparing to the 29 testing cases. Results were shown in Figure 7. With trade-off ΔZ for pro-oar applied, the mean rectum doses predicted by prior model dropped to 43.8±10.1Gy, from 49.5±9.0Gy in the average dose it predicted previously. The difference to GT pro-oar plans dropped from 6.6% to -0.6%. The same trend is observed for pro-ptv trade-off. This proves that the trade-off ΔZ can predict dose with trade-off for new patients.

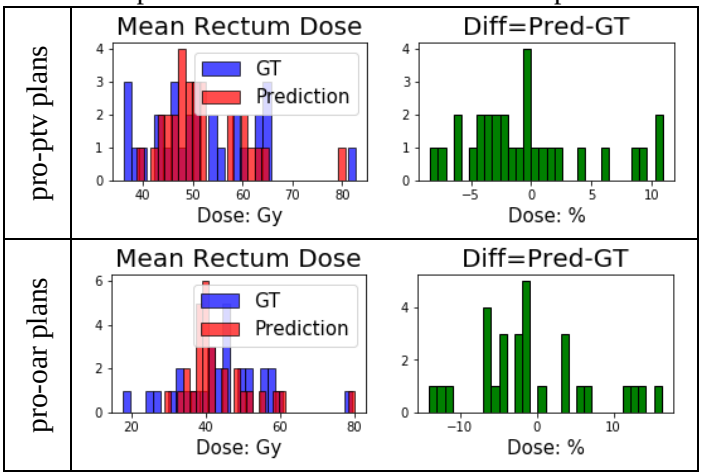


Figure 7. Histogram of predicted dose vs GT doses. It evaluates the accuracy of prior model with trade-off applied

An example of the input contour and predicted dose of prior model is shown in figure 8 below. Statistics of mean rectum doses are shown in table 1 below.

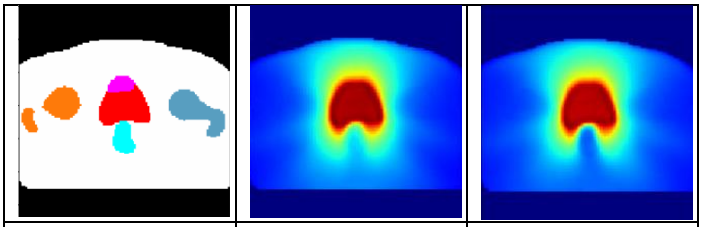


Figure 8. Model input (contours), the average dose predicted by prior model, and the rectum-sparing dose predicted by prior model with ΔZ

Table1:Mean rectumdose		Pro-ptv	Pro-oar
Ground truth doses		52.0±10.6Gy	44.2±12.2Gy
Post model inference	Prediction	51.6±9.9 Gy	46.2±12.4Gy
	Difference	-0.6±1.5%	2.4±1.8%
Prior model inference	Prediction	49.5±9.0Gy	
	Difference	-3.1±5.3%	6.6±7.9%
Prior model inference with ΔZ	Prediction	51.8±8.5Gy	43.8±10.1Gy
	Difference	-0.27±5.1%	-0.6±7.5%

4 Discussion

With pre-calculated ΔZs templated for different trade-offs, physicians can review doses with different trade-offs for a new patient immediately after contouring, and show the desired dose to dosimetrists. As trade-off selection occurs before dosimetry planning, back-and-force is minimized.

5 Conclusion

We proposed a DeepTuning framework that can predict doses with different trade-offs, showing great promise for improving planning efficiency

References

[1] Nguyen, Dan, et al. "A feasibility study for predicting optimal radiation therapy dose distributions of prostate cancer patients from patient anatomy using deep learning." Scientific reports 9.1 (2019): 1076.

Automating Physics Plan Checks with Monaco Scripting

Paul Hill PhD[1]

[1]Department of Medical Physics, Cork University Hospital, Cork, Ireland

Abstract The Scripting Application Programming Interface (API) of the Monaco Treatment Planning System (TPS) allows code to be written which accesses many of the capabilities of Monaco and much of the data of the radiotherapy treatment plans. Scripts are created using the C# programming language and Microsoft Visual Studio and then imported into Monaco using the Manage Scripts Dialog. The API utilizes User Interface (UI) automation to provide access to Monaco functionality. Many of the functions which can be accessed through the Monaco UI are available through the API. A Script been created to automate a number of elements of the Physics Check of radiotherapy treatment plans. It performs checks on the validity of the dose calculation parameters, the treatment beam parameters, details of the treatment couch assignment and the values of the inverse planning optimizer inputs. The Script is in clinical use and is executed by the Dosimetrists prior to plan approval by the Radiation Oncologist. It has led to efficiency improvements in treatment planning as defects can be identified and then rectified early in the workflow.

1 Introduction

Radiotherapy treatment plans are developed by a multidisciplinary team consisting of Radiation Oncologists (RO), Medical Physicists, Dosimetrists and Radiotherapists. One of the critical tasks undertaken by the Medical Physicist is to perform the Plan Check of the treatment plan after it has been approved for treatment by the RO and prior to being sent to the treatment unit for delivery to the patient. This involves confirming the patient identity and prescription, validating that the multiple parameters of the Treatment Planning System (TPS) are correct, checking the dosimetry and preparing the plan for delivery. Guidelines exist to assist the physicist design the Physics Check process [1], in order to ensure the Quality of the treatment plan.

We use the Monaco TPS (Elekta, Stockholm Sweden) at Cork University Hospital. This allows for the generation of complex treatment plans using Volumetric Arc Therapy (VMAT) and Dynamic Multi-leaf Collimator (dMLC) technics. Version 6 of Monaco introduced a Scripting function. This provides an Application Programming Interface (API) which allows code to be created outside of Monaco, and then executed from within. It enables access to many Monaco functions including [2]:

- DICOM file import and export,
- Contouring auto-margins,
- Image Fusion
- Plan creation, optimization and calculation,
- Data extraction, including plan and beam details,
- DVH information,
- QA Plan generation, and,
- Report printing.

The scripting API uses User Interface (UI) automation to allow the script to manipulate the Monaco UI and access the data contained therein. The details of the API are contained in a dictionary "Index.chm" file provided by Elekta.

This work investigates how the process of Physics Plan Check could be automated with the Scripting capability. While the overall plan quality, suitability and correctness has to be assessed by a Medical Physics Expert (MPE) there are elements of the process which are suitable for automation. Some parameters in the TPS, for example the dose calculation grid size, just need to be checked that the value is set to the appropriate clinical value.

2 Materials and Methods

Monaco version 6.1.1 was used in this work. The Script was written using the C# language in the Microsoft Visual Studio 2019 Software Integrated Development Environment (IDE). The IDE was installed directly on to one of the Monaco calculation servers. The Script was compiled and linked using the API headers and Dynamic Linked Libraries (DLLs) which enable the script to communicate with the instance on Monaco running on the server.

Monaco provides configuration and access control to the Script via the Manage Script Dialog [2]. This allows the script to be imported into Monaco by selection of all necessary files. This Script is then available to be executed on any Monaco calculation server. To run any Script a user has to be a member of the Monaco Scripters group. The Manage Scripts Dialog also allows Scripts to be approved. These may then be executed by any Monaco user. This enables the testing of scripts which are not universally available and only when the testing is complete is the script available for all users.

An analysis of the API was carried out to determine which elements of the Physics Check process could be automated and produced the following list:

- Calculation properties including grid spacing, dose deposition type (medium / water) and algorithm;
- For each beam the description (clockwise or anti-clockwise for VMAT and gantry angle for dMLC), Treatment Unit, treatment couch inclusion and electron density assignment , bolus consistency and minimum number of segments;
- Consistency between the Inverse Planning Optimizer Target Penalty and the plan prescription;
- Dose Reference Point (DRP) location;
- Fractional DRP dose should be within tolerance of the Prescription.

These items defined part of the software requirements specification for the Script. In addition the requirements

were defined for how the results would be presented to the user through the UI. The graphical elements were added via Windows Presentation Foundation (WPF) and Extensible Application Markup Language (XAML). For each item to be checked the result was displayed as a pass ✅, fail ❌, or item not found ⚠.
Once complete, the Script underwent unit testing by creating a number of test plans in which each item was set to cause a pass or fail result and, where possible, not found.
Upon completion of unit testing the Script was made available to other members of the Monaco Scripters via the Manage Scripts Dialog. Two Senior Dosimetrists acted as beta testers by incorporating the Script in to their workflow for a period of several weeks after which the Script was approved for Clinical use.
The requirements, design, testing and configuration management of the Script were documented in our in-house Software Design Document [3].

3 Results

The Script was executed by the Dosimetrists prior to the plan being presented to the RO for approval. This was so that items which fail can be found early in the planning workflow and therfore can be remedied prior to approval. Otherwise, if found at the Physics Check, then, in general, the plan would need to be returned to planning for re-calculation, re-approval and re-export, which is time costly.
The Script runs in Monaco by driving the UI. It can be seen opening and closing windows and selecting tabs. For this reason is it is rather slow to execute with a typical VMAT plan requiring about 50 seconds. The user is presented with a visual indication in Monaco that a Script is running with a green (approved) or red (unapproved) animated text message as show in Figure 1.



Figure 1: The visual indication in provided in Monaco for Script execution. The animated text is shown on a green background for an approved Script and red for unapproved.

There are some restrictions on how the Script should be run. Monaco must be in full screen mode and on the main screen if the user has a multi-screen set up. The user must have only one instance of Monaco running on the particular server and requires write access to the patient plan. Monaco, and not MonacoSim, must be used.
A sample of the results presented to the user is shown in Figure 2. The appearance of the UI was designed to be similar in appearance to Monaco, using the same font style and colours.



Figure 2: A sample display presented to the user after script completion.

4 Discussion

The Script has been in Clinical use since early 2023. The feedback from users has been very positive. There was some initial skepticism but is now considered to be generally popular since it has been demonstrated to catch defects early and therefore save time.
A key advantage of moving the checks upstream in the workflow from Physics Check to before the plan approval stage is that any defects are discovered early in the process

and are therefore easier to remedy. Running the Script is repeated at Physics Check to confirm the result.

There are limitations in how Elekta have implemented the API. Not all elements of Monaco are exposed. For example the contents of the Dose Reference Points tab is not available in Monaco 6.1.1. It is expected that more items will become available in future releases. Also, unlike other TPS's with a Scripting Interface, it is not a simple task to iterate through the entire patient database.

The choice of UI automation as the Scripting architecture results in the Script execution times being long. However it offers enormous potential for workflow automation. Combined with an auto-contouring tool then a Script which offers 'single-click' plan creation is only a small number of lines of code. Upon open a patient and running the script a plan can be automatically created, optimized, calculated and saved with relative ease.

We are currently investigating further developments using the API including planning creation automation and plan complexity analysis.

5 Conclusion

We have used the Monaco Scripting API to automate a number of elements of the Physics Check process. The Script is executed prior to the plan being approved by the RO therefore defects are detected early in the planning process. It has been released to Clinical use and has led to significant time saving and workflow efficiency improvement. We expect to further develop the Script with future releases of Monaco as the number of Monaco items exposed through the API increases.

References

[1] A Ford et al. "Strategies for effective physics plan and chart review in radiation therapy: Report of AAPM Task Group 275" Med. Phys. 47 (6) (2020). DOI: 10.1002/mp.14030

[2] Elekta Solutions AB. "Monaco Scripting Reference Manual". Document E011301/06 (2022).

[3] P. Hill. "A General Process for in-house Medical Physics Software Development". 18th International Conference on the use of Computers in Radiation Therapy.

Optimization in radiotherapy integrating a multidisciplinary framework with GEANT4

E. Sergi[1], M. Severgnini[2], E Rigoni[1], and F. Longo[3]

[1]ESTECO S.p.A., Trieste, Italy
[2]SC Fisica Sanitaria director, Azienda sanitaria universitaria Giuliano Isontina, Trieste, Italy
[3]Department of Physics, University of Trieste and INFN Sezione di Trieste, Trieste, Italy

Abstract Radiotherapy is a complex treatment that requires optimization to deliver the right dose to the target while minimizing side effects.
This work [1] proposes a novel approach that combines optimization techniques from other engineering fields with Monte Carlo (MC) simulations to optimize RT plans. The coupling of GEANT4, an MC simulation toolkit with modeFRONTIER, an optimization platform, enabled the efficient identification of optimal linac parameters for tumor treatment.
This approach demonstrated improved accuracy and personalized treatment plans for RT, offering promising advancements in medical physics.

1 Introduction

Radiotherapy treatment (RT) is one of the most important cancer treatment; mathematically, it is a constrained multi-objective optimization problem that involves balancing 10-30 highly correlated criteria and constraints.
The solution to the problem appears simple, but achieving the best radiotherapy plan for a specific patient is a difficult challenge.
No single optimal solution can be found in multi-objective optimization problems, but we can find a set of different trade-offs, mathematically equivalent, called *Pareto optimal solutions*. A perfect multi-objective solution is impossible; a reasonable solution is to investigate a set of equally good compromises.
When dealing with dose calculation, different algorithms are available. Monte-Carlo algorithms are more accurate than conventional techniques: they reduce uncertainty in dose calculation and the uncertainty in overall delivered dose.
Our aim is to introduce the knowledge of optimization techniques from other engineering fields into medical physics to optimize a treatment plan.

2 Materials and Methods

GEANT4 [2] and modeFRONTIER have both been used to optimize a RT plan. modeFRONTIER is an integration platform for multi-objective and multi-disciplinary optimization, developed by ESTECO [3], an independent Italian software company highly specialized in numerical optimization, simulation process, and data management. It provides a seamless coupling with third-party engineering tools, enables the automation of the design process simulation, and facilitates analytic decision-making.
GEANT4 is an open-source Monte Carlo simulation toolkit for precise dose calculation.
We created models for patient and medical linac based on the `medical_linac` example models [4] in the GEANT4 distribution, an IAEA compliant phase space calculation, and two different phantoms, water and lung tissue with an ellipsoidal tumor.
The optimization was executed using the `MOGA-II` [5] algorithm. The objective was to find the linac parameters that would maximize the tumor dose while minimizing the related side effects.

3 Results

The coupling of modeFRONTIER and GEANT4 worked well; we were able to find the optimal configuration with a reduced number of initial particles in the MC simulation.
We analyzed two different problems.
The first problem is to retrieve a given photon momentum distribution in a phase space by changing the flattening filter geometry. The "real" and "optimal" geometries are then compared in terms of Percentage Depth Dose (PDD) and lateral dose profile using gamma index 3%/3mm analysis. A water phantom is used for this study, with 100 sensitive cubic elements for each direction. This is a single objective problem, as the goal is to find the flattening filter geometry that produces the best PDD and lateral dose profile.

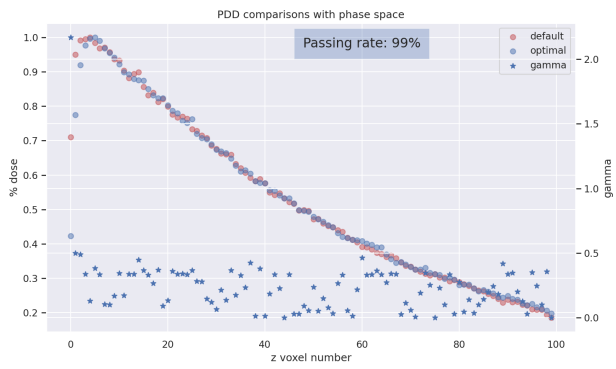


Figure 1: First problem: "real" and "optimal" geometries comparisons using γ-index 3%/3mm analysis for photon beam PDD.

The second problem involves a convex solid tumor inside lungs undergoing a simplified version of radiotherapy using four different gantry positions. The treatment is delivered all at once, without any dose fractionation. No ICRU (International Commission on Radiation Units and Measurements) volumes or dose constraints are defined. The problem is formulated as a two-objective optimization problem, with the objectives being to maximize tumor dose and minimize total dose to healthy lung tissue.
By utilizing a simplified TPS model, we retrieved the best configurations for tumor treatment. `MOGA-II` explored an area in the objective space not accessible in the initial designs. It also found a homogeneous Pareto front, with optimal designs distributed throughout the whole region.

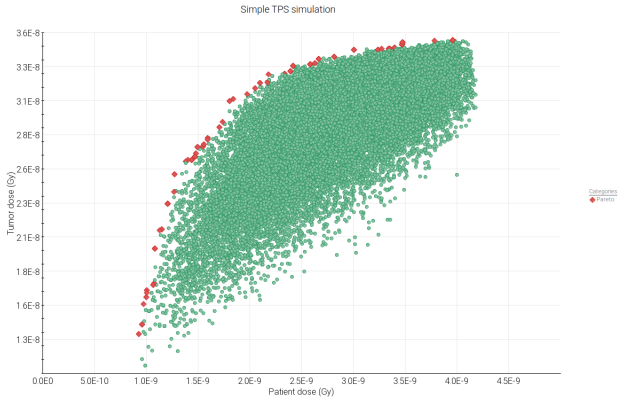


Figure 2: Optimization problem results with a well defined Pareto front (non-dominated results).

4 Conclusions

Medical physics benefits from the expertise in optimization derived from other engineering fields. Our procedure can be applied to optimize some linac components, such as the electron source.
RT plans benefit from the usage of MC algorithms, indeed they are more accurate than the conventional techniques and are suitable for a personalized treatment plan.
With its state-of-the-art algorithms for optimization, sensitivity analysis, and uncertainty quantification, modeFRONTIER is also a helpful tool for managing uncertainties in radiotherapy.

5 Forthcoming Research

To further enhance the efficiency and accuracy of the proposed radiotherapy optimization approach, the following research directions are planned:

- *Parallelization of* GEANT4: Leverage the parallel computing capabilities of GEANT4 to significantly reduce the computational time required for simulations, enabling the optimization of complex radiotherapy treatment plans more effectively.
- *Integration of GDML and DICOM*: Employ the GDML (Geometry Description Markup Language) and DICOM (Digital Imaging and Communications in Medicine) formats to seamlessly integrate linac and patient descriptions into the optimization framework. This will simplify the model setup and improve the accuracy of dose calculations.
- *Incorporating ICRU volumes and dose constraints*: Define the geometrical extent of ICRU volumes, representing organs at risk, and incorporate dose constraints into the optimization algorithm. This will ensure that the optimized treatment plan minimizes the risk of side effects to healthy tissues.
- *Investigating uncertainty quantification and robust optimization*: Address the issue of uncertainties in patient data and modeling assumptions by incorporating uncertainty quantification techniques into the optimization process. Additionally, explore robust optimization methods that can handle uncertainties and provide reliable solutions even in the presence of model imperfections.

These research endeavors aim to further advance the application of optimization techniques in radiotherapy, leading to more personalized, effective, and safe treatment plans.

6 Acknowledgments

Thanks to ESTECO for the provided support.

References

[1] E. Sergi, M. Severgnini, E. Rigoni, et al. “Investigating optimization in radiotherapy integrating a multidisciplinary optimization framework with GEANT4”. Master Thesis. University of Trieste, Department of Physics, 2021.

[2] G. Collaboration, S. Agostinelli, et al. “GEANT4–a simulation toolkit”. *Nucl. Instrum. Meth. A* 506.25 (2003).

[3] *ESTECO S.p.A.* company website. URL: https://www.esteco.com/.

[4] B. Caccia, C. Andenna, and G. A. P. Cirrone. “MedLinac2: a GEANT4 based software package for radiotherapy”. *Annali dell’Istituto Superiore di Sanità* 46 (2010), pp. 173–177.

[5] C. Poloni and V. Pediroda. “GA coupled with computationally expensive simulations: tools to improve efficiency”. *Genetic Algorithms and Evolution Strategies in Engineering and Computer Science*. Ed. by D. Quagliarella, J. Périaux, C. Poloni, et al. Wiley, 1997, pp. 267–288.

A Personalized Automatic Radiotherapy Plan Scoring Mechanism Based on Multi-organ Individually-objective Optimization

Zirong Li[+1], Zhenjiang Li[+2], X. Sharon Qi[3], Kuo Li[2], Huashan Sheng[1], Zhigang Yuan[1], Yong Yin[*2], and Qichao Zhou[*1]

[1]Department of Research Algorithms, Manteia Technologies Co., Ltd, Xiamen, China
[2]Shandong Cancer Hospital and Institute, Shandong First Medical University and Shandong Academy of Medical Sciences
[3]Department of Radiation Oncology, University of California, Los Angeles, Los Angeles, CA. USA

Abstract Radiation therapy plan optimization is a time consuming and very complex process, and may not always result in an ultimate optimal solution. It is difficult to uniformly define what kind of plan is a high-quality plan when multiple complex standards need to be balanced. Manual plan evaluation is generally time-consuming and subjective according to reviewers' experiences, preferences as well as consideration of patient characteristics. We proposed an individualized automatic and objective plan evaluation method based on multiple-organ individually objective optimization (MIO). The proposed plan evaluation method is possible without manual intervention and captures overall plan quality using a plan scoring system. The scoring system is constructed based on the achievable dose limit for relevant OARs. In double-blind comparative experiments, the consistency between the results of the manual evaluation (standard practice) and the automatic plan evaluation system was 0.963, indicating its feasibility and efficacy. The proposed mechanism significantly reduces manual review time, the automatic plan evaluation MIO takes an average of 33 seconds, and the evaluation process takes an average of 17 seconds.

1 Introduction

Radiotherapy is one of the main treatments for cancer[1]. The effectiveness of radiotherapy depends on the quality of the treatment plan. A higher target dose ensures the tumor control probability(TCP), and limiting the normal tissue dose within a tolerated range is the key to reducing the normal tissue complications probability(NTCP).

However, plan optimization is a very complex process, and there often does not exist a unique optimal solution[2, 3]. It is difficult to uniformly define what kind of plan is a high-quality plan when multiple complex standards need to be balanced. Manual plan evaluation is generally time-consuming and subjective according to reviewers' experiences, preferences, and complicated clinical situations.

For different patients tumor geometry and location result in large variations of achievable organs-at-risk(OAR) dosimetry in the radiotherapy treatment planning process. This greatly increases the difficulty of plan optimization because we have no way of knowing whether the currently selected dose is optimal.

In recent years, many studies have conducted dose prediction based on images[4–6]. However, deep learning models based on retrospective data often have poor generalization, and it is difficult to collect huge data for training due to the preciousness and scarcity of clinical data.

We proposed an individualized automatic and objective plan evaluation method based on multiple-organ individually objective optimization (MIO). The proposed plan evaluation method is possible without manual intervention and was summarized to form a plan scoring system based on the achievable dose limit for relevant OARs.

2 Materials and Methods

Individualized plan evaluation uses MIO to obtain each patient's unique OAR distribution lower-bound. For each plan, beam angles, prescription doses, and structures were kept the same for all planners. In the initial phase, target structures were optimized to achieve the desirable prescription dose using V95 for the targets without consideration of any OAR. Then, each OAR, as well as the target, was added to the optimization process to generate a single-objective OAR plan. The optimization process stop until the target V95 no longer met the prescription objective

Compared to traditional fluence map optimization (FMO)[7, 8], the proposed method considers a single OAR at a time in each optimization process. The optimization function can be described as follows:

$$f = \sum_{n=1}^{N} p_n^T f_n^T + p^{OAR} f^{OAR} \tag{1}$$

Whre f_n^T is the objective function for the nth target volume, and p_n^T is the weight for that target. Similarly, f^{OAR} is the OAR to be optimized, and p^{OAR} is the corresponding penalty factor. This process was repeated for all OARs resulting in optimal dose distribution for each OAR. The difference in the fluence map of MIO results for two OARs is shown in Figure 1.

The OAR score is defined as the ratio of the area between the OAR DVH curve of the reference (clinical) plan and the DVH curve obtained from the single-objective plan. The weighted average of OAR scores (OAR_S) were calculated as the weighted sum.. The weights were determined by the MIO results. The target dose was scored using the Homogeneity Index score (HI_S), maximum dose score (MAX_S), and

[+] The authors contributed equally to this work.
[*] Authors to whom correspondence should be addressed.

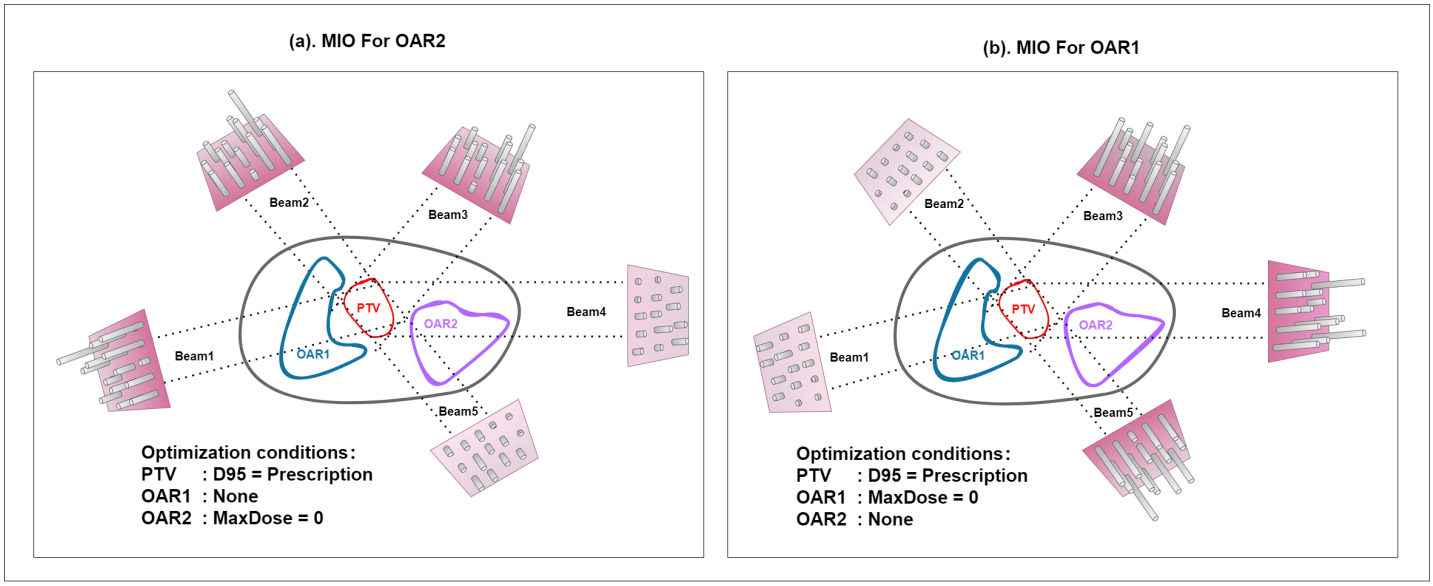


Figure 1: Difference in fluence map of MIO results for two OARs. The difference of the fluence map of the MIO plans when different OARs when included in the optimization process. If the beam passes through the optimized OAR, the fluence map will approach zero, which makes the surrounding dose larger, and the currently optimized OAR reaches the lowest dose.

minimum dose score (MIN_S). The final plan score was calculated as the average of the OAR_S for normal organs, and HI_S, MAX_S, and MIN_S for targets.

We evaluated the effectiveness of the algorithm through a set of head and neck cases. A double-blind comparison was designed. For the same patient, we asked three planners to generate three plans. These plans were evaluated by an expert oncology and ranked into the high-quality plans from low-quality plans. These plans were also evaluated by the MIO. The individual evaluation program did not include any knowledge of human evaluation results. The results of automatic evaluation were compared with the manual evaluation results.

3 Results

Figure 2 visualizes the MIO score results for two example patients. Longitudinal comparison of three plans from three planners in two patients. Differences in dose distribution may occur between different patients due to differences in the size and location of the target volume. Therefore, the MIO results of different patients must be different. This is the most important individualized index, and these differences can be included, in the OAR volume, target volume, and the positional relationship between OAR and target distinction. Finally, provides specific personalized quantitative indicators.

In a double-blind comparative experiment, the consistency between the results of the manual evaluation and the results of the automatic evaluation was 0.963. Most of the plan evaluation results were consistent with manual judgment, but manual evaluation of each plan took at least a few minutes to a few hours. However, the automatic plan evaluation MIO takes an average of 33 seconds, and the evaluation process takes an average of 17 seconds. It should be noted that only one MIO is required for N planned N assessments of one patient.

4 Discussion

We proposed an automatic plan evaluation method based on MIO, which can quickly and automatically score different plans. This algorithm can solve the problem of traditional manual evaluation being time-consuming and lacking objective standards and also allows the automatic evaluation of the merits of the plan in the automatic planning process. A scoring system consistent with experts can lead to higher-quality automated plans.

Although deep learning has developed rapidly in recent years, it still has many limitations, such as the need for huge data for training and poor personalization. In comparison, the method we proposed overcomes these shortcomings, but the application speed of the algorithm is slow. It is still slightly slower, and the advantages of both should be combined in the future.

5 Conclusion

The proposed plan evaluation method, based on multiple-organ individually objective OAR optimization, provided an objective, quantitative, and automated plan evaluation mechanism, which could greatly improve the quality and consistency of IMRT plans; streamline plan evaluation and evaluation efficiency in treatment planning workflow, and guide for planners to achieve optimal cancer treatment.

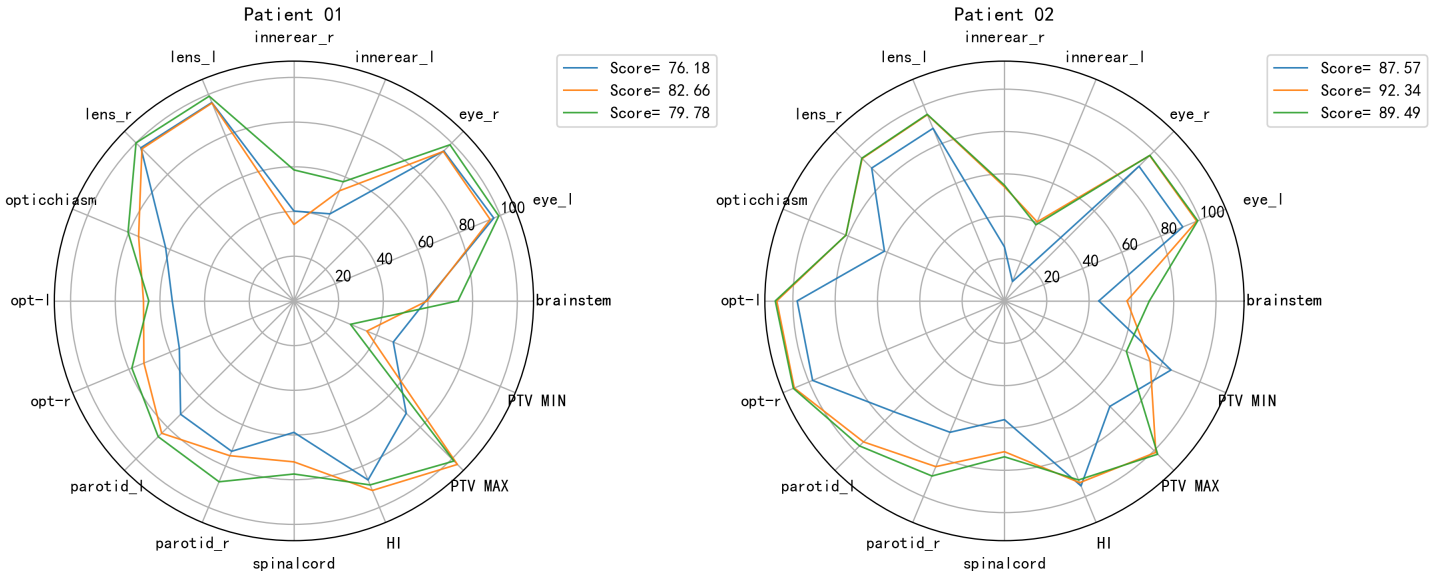


Figure 2: MIO results for each OAR in nine typical patients.

References

[1] R. Baskar, K. A. Lee, R. Yeo, et al. “Cancer and radiation therapy: current advances and future directions”. *International journal of medical sciences* 9.3 (2012), p. 193.

[2] T. Hong, M. Ritter, W. Tome, et al. “Intensity-modulated radiation therapy: emerging cancer treatment technology”. *British journal of cancer* 92.10 (2005), pp. 1819–1824.

[3] S. A. Leibel, Z. Fuks, M. J. Zelefsky, et al. “Intensity-modulated radiotherapy”. *The Cancer Journal* 8.2 (2002), pp. 164–176.

[4] N. Bakx, H. Bluemink, E. Hagelaar, et al. “Development and evaluation of radiotherapy deep learning dose prediction models for breast cancer”. *Physics and imaging in radiation oncology* 17 (2021), pp. 65–70.

[5] Z. Feng, L. Wen, P. Wang, et al. “DiffDP: Radiotherapy dose prediction via a diffusion model”. *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2023, pp. 191–201.

[6] Z. Liu, X. Chen, K. Men, et al. “A deep learning model to predict dose–volume histograms of organs at risk in radiotherapy treatment plans”. *Medical Physics* 47.11 (2020), pp. 5467–5481.

[7] H. E. Romeijn, R. K. Ahuja, J. F. Dempsey, et al. “A novel linear programming approach to fluence map optimization for intensity modulated radiation therapy treatment planning”. *Physics in Medicine & Biology* 48.21 (2003), p. 3521.

[8] Q. Wu and R. Mohan. “Algorithms and functionality of an intensity modulated radiotherapy optimization system”. *Medical physics* 27.4 (2000), pp. 701–711.

Dynamic Collimator Rotation Treatment Plan Optimization for O-Ring Treatment Systems

Michael K Fix, Werner Volken, Daniel Frei, Silvan Mueller, Gian Guyer, Daniel Schmidhalter, and Peter Manser

Division of Medical Radiation Physics and Department of Radiation Oncology, Inselspital,
Bern University Hospital, and University of Bern, Bern, Switzerland

Abstract In intensity modulated radiotherapy (IMRT) and volumetric modulated arc therapy (VMAT) static collimator angles are used per field or arc. While dynamic collimator rotation during beam on was investigated for C-arm linear accelerators, this was not investigated for O-ring ETHOS/Halcyon treatment systems. This work aims in the development of an optimization framework for treatment plans with dynamic collimator rotation during beam on for ETHOS/Halcyon systems. A previously developed in-house hybrid direct aperture optimization algorithm was first extended with a Monte Carlo (MC) model of the ETHOS/Halcyon treatment head including the double stack multi leaf collimator to calculate MC beamlet dose distributions and second with options for dynamic collimator rotation optimization for IMRT and VMAT. The framework was tested for a clinically motivated case showing improved dose sparing for organs at risk using dynamic collimator angle rotation compared with static collimator angle treatment plans, while the target coverage was the same.

1 Introduction

Currently, intensity modulated radiotherapy (IMRT) and volumetric modulated arc therapy (VMAT) are well-established state-of-the-art treatment techniques in radiotherapy [1,2]. More recently, there is an increasing interest in exploiting additional degrees of freedom in radiotherapy treatment planning to further improve treatment plan quality [3]. There is a variety of these techniques investigated ranging from using multiple arcs at different static table angles to dynamic trajectory radiotherapy (DTRT) using simultaneous dynamic gantry, table and collimator rotation during beam on leading to promising results in terms of treatment plan improvements [3–6]. While dynamic collimator rotation as a degree of freedom for dynamic collimator radiotherapy was investigated for C-arm linear accelerators, this was not yet investigated for O-ring ETHOS or Halcyon treatment systems.

Previously, we developed an in-house optimization framework within a treatment planning process to generate deliverable DTRT treatment plans for C-arm linear accelerators [7]. This framework consists of a hybrid direct aperture optimization (HDAO) algorithm, which combines a column generation algorithm with simulated annealing and is based on Monte Carlo (MC) generated beamlet dose distributions used to determine the aperture shape along with the Monitor Unit (MU) weight. However, this HDAO cannot be directly applied to ETHOS/Halcyon treatment systems, as those system consists of a different beam defining system including a unique double stack multi leaf collimator (MLC).

The aim of this work is the development of an optimization framework for VMAT and IMRT treatment plans with dynamic collimator rotation for ETHOS/Halcyon treatment systems to improve treatment plan quality compared to plans with static collimator rotation angels. For this purpose, first a MC model for an ETHOS/Halcyon treatment system was implemented and commissioned for an ETHOS unit to be used to generate MC beamlet dose distributions for the HDAO. Second, the HDAO algorithm was extended in order to handle the double stack MLC and the dynamic collimator rotation in the optimization. Finally, the developed framework was tested for a clinically motivated head and neck case.

2 Materials and Methods

In a first part of this work, an MC model of the beam defining system for an ETHOS/Halcyon treatment system consisting of the target, the primary collimator, the monitor chamber, the secondary collimator, the double-stack MLC and the leaf box cover as illustrated in Figure 1 was implemented into our in-house developed Swiss Monte Carlo Plan (SMCP) [8] following the information provided by the vendor (Varian, a Siemens Healthineers Company).

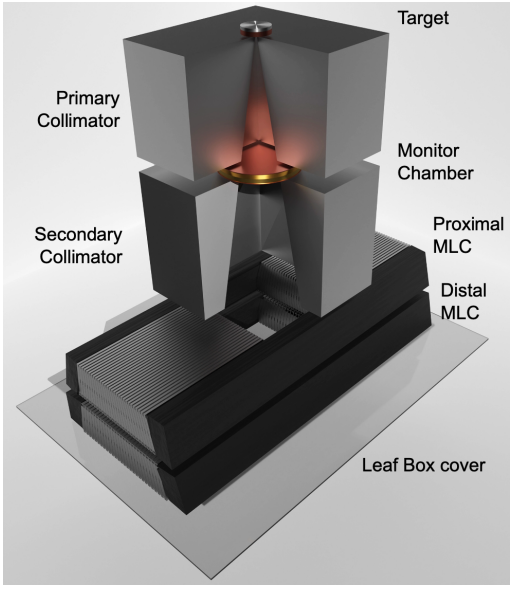


Figure 1: Illustration of the beam defining components for the ETHOS/Halcyon treatment system as implemented in the SMCP.

For the commissioning of the model, the energy and the lateral fluence distribution of the initial electron beam impinging on the target were tuned to match measured dose distributions. For this purpose, depth dose curves and lateral dose profiles in five depths ranging from 1.3 cm to 15 cm

were measured for square MLC shaped field sizes ranging from 1 x 1 cm^2 to 28 x 28 cm^2 at a source to surface distance (SSD) of 100 cm in units of cGy/MU on an ETHOS treatment system. For the validation the analogous set of dose distributions at an SSD of 90 cm were measured and compared with the corresponding calculated dose distributions using the commissioned MC model.

The implemented MC model was used in our HDAO algorithm. However, this HDAO so far is not able to handle the double stack MLC of an ETHOS/Halcyon system. This double stack ETHOS/Halcyon MLC consists of a proximal part with 29 leaf pairs and a distal part with 28 leaf pairs. In the following these are referred to as the physical MLC leaves. The width for each individual physical MLC leaf is 1 cm in the isocenter plane. The positions of the physical 57 MLC leaf pairs can be described by a set of 2 x 57 numbers. In general, the combination of two field shapes individually defined by the proximal and the distal MLC parts can be described with 56 so-called pseudo-MLC leaf pairs with a width of 0.5 cm in the isocenter plane, i.e. any combination of proximal and distal MLC leaf patterns (physical aperture) can be represented unambiguously by a pseudo leaf pair pattern (pseudo aperture). An example of such a conversion from a physical to a pseudo aperture is illustrated in Figure 2.

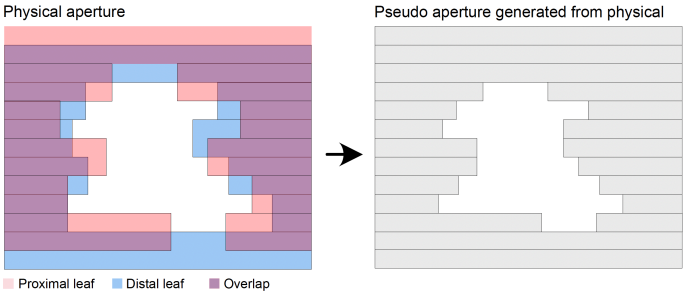


Figure 2: Illustration of the conversion from physical leaf positions to pseudo leaf positions. The graphics on the left shows the combination of a proximal and a distal MLC leaf pattern (physical aperture), while the graphics on the right-hand side shows its representation using the pseudo-MLC leaves (pseudo aperture).

Consequently, 56 x 56 beamlets with a size of 0.5 x 0.5 cm^2 each were generated using the commissioned MC model and serve as a base for the beamlet dose calculation resulting in the dose influence matrix applied during optimization. During the HDAO, the optimization algorithm keeps track of two sets of numbers representing the physical aperture (2 x 57 numbers) as well as the pseudo aperture (2 x 56 numbers). Both sets are always synchronized to guarantee that beamlet selection based on pseudo apertures is in compliance with the corresponding physical apertures.

In addition, the HDAO was extended to include the dynamic collimator rotation angle in the treatment planning process with different approaches for IMRT and VMAT.

In the case of IMRT, a set of 12 equi-angular beam directions are provided as input for the HDAO. This setup corresponds to our standard IMRT setup template of the ETHOS planning system. The number of apertures was limited to a total of 100, based on previous investigations with the HDAO. Next, beamlet dose distributions per beam directions were calculated for different collimator angles. This was done for a set of discrete collimator angles {-80°, -60°, -40°, -20, 0°, 20°, 40°, 60°, 80°} referred to as *'optimized collimator angle'* method (*oca*) leading to an increased search space of promising apertures for each beam direction. This newly developed method was compared with using a static static collimator angle of 10° for all apertures and all beam directions referred to as *'static collimator angle'* method (*sca*).

In the case of VMAT, the *oca* method would allow large collimator angle rotation changes between adjacent control points resulting in excessive collimator movement. In addition, this would also need an enormous amount of beamlet dose calculations. Thus, an approach was implemented in which the gantry-collimator path is predefined and provided as input to the HDAO. For this purpose, we implemented a '*predefined collimator angle*' method (*pca*). In this *pca* method a A* path finding algorithm is applied with a maximum gradient of 4°/5° (collimator/gantry) to a gantry-collimator whitespace area map to find the gantry-collimator path that maximally reduce the whitespace area. For this map, first the area between the PTV structure is extended by an isotropic margin of 0.5 cm. Then, the conformal MLC opening (whitespace area) is determined as a function of the gantry and collimator rotation angle settings leading to a two-dimensional gantry-collimator map. The resulting path was then passed to the HDAO for optimization.

Finally, this developed HDAO framework was tested for a clinically motived head and neck case and treatment plans for a 12 field IMRT and a two full arc VMAT technique with and without dynamic collimator rotation were generated. For the optimization the objectives for the intensity modulation optimization were based on those used for the clinically delivered treatment plan. The optimization utilizing a static or a dynamic collimator rotation used the same set and values for the objectives and priorities as well as the same objective function. For all these comparisons the normalization was such that 95% von the PTV ($D_{95\%}$) receives the prescribed dose. Resulting treatment plan qualities using either a dynamic collimator angle or a static collimator angle were compared by means of the dose distributions and dose volume histogram parameters.

3 Results

An MC model of the ETHOS/Halcyon radiotherapy beam was successfully implemented in the SMCP. The commissioned MC model for an ETHOS treatment system agreed in general within 1% (global) or 1 mm for all depth dose curves and lateral dose profiles in units of cGy/MU for

all square field sizes considered in the validation. Figure 3 shows some exemplary results of this validation.

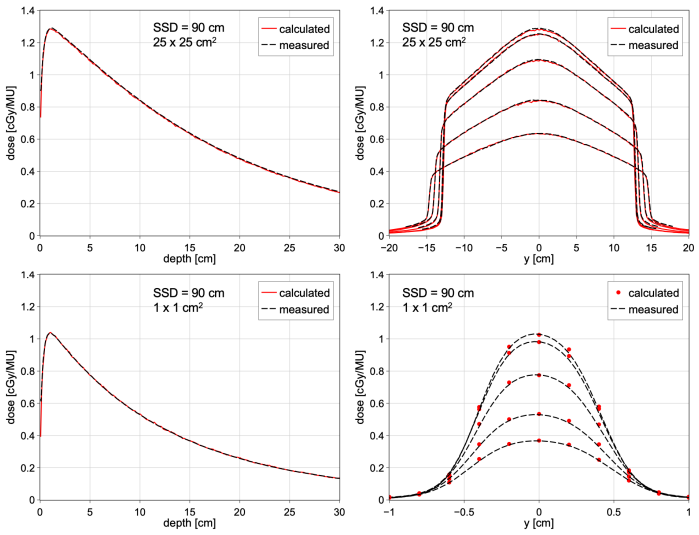


Figure 3: Absolute dose comparisons of calculated dose distributions with the corresponding measured dose distributions in homogeneous water at an SSD = 90 cm for field sizes of 25 x 25 cm² (top row) and 1 x 1 cm² (bottom row) using the commissioned MC model. Depth dose curves and lateral dose profiles in five depths (1.3 cm, 2 cm, 5 cm, 10 cm, 15 cm) of water in y-direction are shown for each field size. The agreement is generally within 1% or 1 mm.

In Figure 4 the results for the clinically motivated head and neck case are presented utilizing the 12 field IMRT technique.

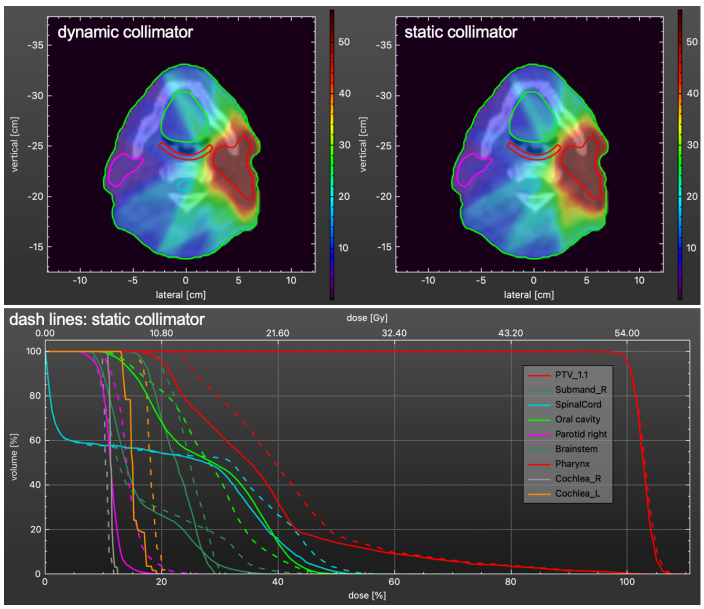


Figure 4: Dose distributions and DVH comparisons for IMRT treatment plans using either the dynamic collimator angle approach or the static collimator angle approach for a clinically motivated head and neck case. For this comparison the 'optimized collimator angle' (oca) method was used to determine the collimator angle for each individual aperture of the beam directions.

Both treatment plans result in a very similar target coverage, while improved dose distributions were achieved for the treatment plan with dynamic collimator rotation when compared with the treatment plan using a static collimator rotation of 10° for all apertures.

The results for the two-arc VMAT treatment plans are illustrated in Figure 5. Again, the target coverage was very similar for the two treatment techniques. For the plan using a static collimator rotation angle per arc the values of 9° and 29° were used (one value per arc). The OAR sparing improvements for the treatment plan including the dynamic collimator rotation vs. the treatment plan using static collimator angles was less pronounced for the VMAT technique when compared with the IMRT delivery modality. The dose volume histogram parameters for the comparisons are summarized in Table 1.

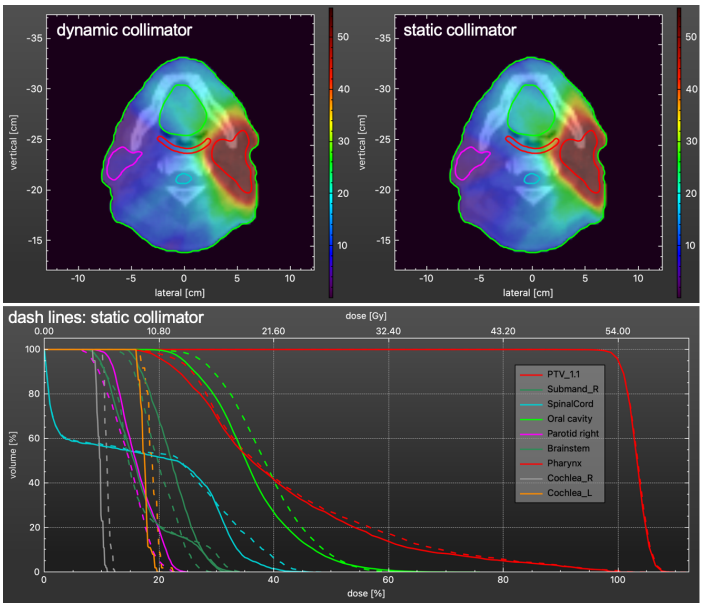


Figure 5: Dose distributions and DVH comparisons for the two-arcs VMAT treatment plans using either the dynamic collimator angle approach or the static collimator angle approach for the clinically motivated head and neck case already shown in Figure 4. For this comparison the predefined collimator angle (pca) method was utilized, where the A algorithm was used to determine the collimator rotation angle as a function of the gantry angle.*

Table 1: Comparison of dosimetric parameters for the dose distributions of treatment plans without (IMRT, VMAT) and with ($IMRT_{dyn}$, $VMAT_{dyn}$) dynamic collimator rotation for the clinically motivated head and neck case shown in Figures 4 and 5. Superior values are highlighted in bold (HI = homogeneity index; CI = conformity index).

PTV	IMRT	$IMRT_{dyn}$	VMAT	$VMAT_{dyn}$
$D_{98\%}$ [Gy]	53.6	53.6	53.5	53.5
$D_{2\%}$ [Gy]	57.3	**57.1**	57.8	**57.7**
HI	0.070	**0.065**	0.082	**0.079**
CI	0.695	**0.714**	0.718	**0.721**
Submadible right				
D_{mean} [Gy]	13.4	**12.1**	**10.8**	12.0
$D_{2\%}$ [Gy]	16.2	**15.5**	**14.5**	16.3
Spinal Cord				
Dmean [Gy]	12.3	**11.3**	9.8	**9.3**
$D_{2\%}$ [Gy]	27.5	**26.2**	23.1	**21.2**
Oral Cavity				
D_{mean} [Gy]	**15.0**	15.2	20.9	**19.5**
$D_{2\%}$ [Gy]	**23.9**	24.4	**29.6**	30.2
Parotid right				
D_{mean} [Gy]	7.7	**6.0**	**7.8**	8.8
$D_{2\%}$ [Gy]	11.4	**8.2**	**11.5**	12.3
Brainstem				
D_{mean} [Gy]	9.2	**8.8**	9.1	**9.0**
$D_{2\%}$ [Gy]	20.6	**17.5**	17.3	**16.2**
Pharynx				
D_{mean} [Gy]	22.6	**20.1**	23.0	**22.3**
$D_{2\%}$ [Gy]	48.0	**47.5**	50.4	**50.1**
Chochlea right				
D_{mean} [Gy]	**5.6**	6.1	5.9	**5.2**
$D_{2\%}$ [Gy]	**6.3**	6.7	6.5	**5.8**
Cochlea left				
D_{mean} [Gy]	9.8	**8.1**	10.1	**9.4**
$D_{2\%}$ [Gy]	10.8	**9.9**	11.0	**10.3**

4 Discussion

We have successfully implemented and commissioned an MC model of the beam defining components of an ETHOS treatment system. The validated MC model is now available within the SMCP for accurate dose calculation. The level of accuracy is similar to the results shown by Laakkonen et al. [9]. In addition, this model was successfully integrated in our in-house optimization framework, which was extended to allow optimization for an ETHOS/Halcyon system including the double stack MLC as well as for dynamic collimator angle rotation.

While the results of this work focused on an ETHOS treatment system, the developed framework is also suitable for Halcyon treatment systems, as the beam defining system components are the same. However, the MC model would have to be commissioned for the different systems.

The currently used clinical ETHOS/Halcyon treatment systems do not allow dynamic collimator rotation during beam on for VMAT. However, the generated treatment plans take into account all mechanical and dosimetric influencing parameters and contributions in the final dose calculation and consequently such VMAT treatment plans can be considered as potentially deliverable treatment plans. Although in this study only one clinically motivated case was investigated, the results demonstrate potential for dynamic collimator rotation treatment plans to improve OAR sparing when compared with treatment plans using static collimator rotation angles. However, a larger number of clinically motivated cases has to be investigated to determine the potential dosimetric benefit of dynamic collimator rotation radiotherapy for ETHOS/Halcyon treatment systems.

5 Conclusion

We have successfully implemented an optimization framework for dynamic collimator rotation optimization for ETHOS/Halcyon treatment systems based on MC generated beamlet dose distributions. The results suggest potential benefits for OAR sparing when dynamic collimator rotation is applied. The framework can now be used to investigate the potential benefit of dynamic collimator radiotherapy for a larger patient cohort.

Acknowledgements

This work was supported by Varian, a Siemens Healthineers Company. Dose calculations and optimizations were performed on UBELIX (www.id.unibe.ch/hpc), the high-performance cluster at the University of Bern.

References

[1] Yu CX. Intensity-modulated arc therapy with dynamic multileaf collimation: an alternative to tomotherapy. Phys Med Biol 1995;40:1435–49. https://doi.org/10.1088/0031-9155/40/9/004.

[2] Otto K. Volumetric modulated arc therapy: IMRT in a single gantry arc. Med Phys 2008;35:310–7. https://doi.org/10.1118/1.2818738.

[3] Smyth G, Evans PM, Bamber JC, Bedford JL. Recent developments in non-coplanar radiotherapy. Br J Radiol 2019;92:20180908. https://doi.org/10.1259/bjr.20180908.

[4] Fix MK, Frei D, Volken W, Terribilini D, Mueller S, Elicin O, et al. Part 1: Optimization and evaluation of dynamic trajectory radiotherapy. Med Phys 2018. https://doi.org/10.1002/mp.13086.

[5] Bertholet J, Mackeprang P-H, Mueller S, Guyer G, Loebner HA, Wyss Y, et al. Organ-at-risk sparing with dynamic trajectory radiotherapy for head and neck cancer: comparison with volumetric arc therapy on a publicly available library of cases. Radiat Oncol 2022;17:122. https://doi.org/10.1186/s13014-022-02092-5.

[6] Guyer G, Mueller S, Koechli C, Frei D, Volken W, Bertholet J, et al. Enabling non-isocentric dynamic trajectory radiotherapy by integration of dynamic table translations. Phys Med Biol 2022;67:175003. https://doi.org/10.1088/1361-6560/ac840d.

[7] Mueller S, Guyer G, Risse T, Tessarini S, Aebersold DM, Stampanoni MFM, et al. A hybrid column generation and simulated annealing algorithm for direct aperture optimization. Phys Med Biol 2022;67. https://doi.org/10.1088/1361-6560/ac58db.

[8] Fix MK, Manser P, Frei D, Volken W, Mini R, Born EJ. An efficient framework for photon Monte Carlo treatment planning. Phys Med Biol 2007;52:N425-437. https://doi.org/10.1088/0031-9155/52/19/N01.

[9] Laakkonen L, Lehtomäki J, Cahill A, Constantin M, Kulmala A, Harju A. Monte Carlo modeling of Halcyon and Ethos radiotherapy beam using CAD geometry: validation and IAEA-compliant phase space. Phys Med Biol 2023;68:044002. https://doi.org/10.1088/1361-6560/acb4d9.

Dosimetry for head and neck radiotherapy using personalized quantitative MRI

SAYAQUE Laura[1], LEPORQ Benjamin[1], BOUYER Charlène[2,3], PILLEUL Frank[1,3], GREGOIRE Vincent[2], BEUF Olivier[1]

[1]Univ Lyon, INSA-Lyon, Université Claude Bernard Lyon 1, UJM-Saint Etienne, CNRS, Inserm, CREATIS UMR 5220, U1294, F-69100, LYON, France
[2]CRLCC Léon Bérard – Département de Radiothérapie
[3]CRLCC Léon Bérard – Département de Radiologie

Abstract Radiotherapy for head and neck cancers is challenging, due to the complexity and heterogeneity of the anatomy. In order to improve this process and make it more accurate, MRI-based planning has been assessed in several studies. However, compared to brain and prostate, head and neck area may have anatomical and tumor type variations from one patient to another, leading to limitations in the existing solutions, mostly based on learning. To address this issue, a direct physical-based method is proposed in this work, to obtain quantitative information which would lead to a more personalized treatment, through the proton density. This method uses an ultra-short echo time sequence to measure the MR signal almost without any decay. The dosimetry on a reference CT and on the synthetic CT generated with this method were compared for nasopharyngeal, oropharyngeal, laryngeal and oral cavity squamous cell carcinoma patients treated by radiotherapy. A 125 HU MAE was obtained between the CT and the sCT and a dose difference of 1.5% of the 70 Gy prescription dose was computed as a mean of the Planning Tumor Volumes median dose. The maximum dose for several organs at risk showed a percentage of dose difference below 1% for 5 out of 6 organs. The mean global gamma index for a 3%/3mm criteria was 0.21 ± 0.05 with a 93 ± 2.8% gamma pass rate. The proton density method allowed to generate a synthetic CT with only one dedicated sequence, giving results comparable to the CT-based dosimetry and to the literature for head and neck radiotherapy.

1 Introduction

MRI's excellent contrast between soft tissues is of great interest for radiotherapy planning to improve normal tissue and tumor delineation. But for dose calculation electronic density is needed and is not directly accessible with MRI, unlike CT. This difficulty can be overcome with a synthetic CT created from MR images. Several methods allow to build a synthetic CT: statistical methods such as atlas registration[1] and deep learning[2] and physical-based methods derived from quantitative MRI such as density assignment[3] and proton density measurement[4].

The electronic density depends on tissue characteristics such as the mass stopping power and the mass attenuation coefficient, so these two values have been evaluated for different atoms and for a wide range of energies. It has been found[4] that these values have a different evolution for the hydrogen atom than the other atoms for energies between 0.2 and 8 MeV[4]. Quantitative MRI can provide hydrogen concentration through the measurement of proton density, which is possible with specific sequences.

This work presents a pilot study aiming to build a synthetic CT using MRI proton density measurement with a patient-specific workflow.

2 Materials and Methods

In order to compare a routine CT-based dosimetry with a dosimetry based on our proton density method, a protocol has been conducted on 26 patients with head and neck cancers, to be treated by radiotherapy at the comprehensive cancer Center Léon Bérard in Lyon.

This protocol is divided into two parts, a conventional injected CT scan performed on a Siemens Somatom Go Sim (Siemens Healthineers, Erlangen, Germany) and a 11 minutes MRI acquisition protocol performed on a Siemens 3T Magnetom Vida (Siemens Healthineers, Erlangen, Germany). The MRI protocol includes 7 minutes of clinical routine sequences for the contouring of the tumor volumes (manual) and the organs at risk (automatic) and an additional 4 minutes for the sequence dedicated to the sCT generation for the treatment planning.

To reduce large variations in the position of the patient, the same immobilization material was used for CT, MRI and radiotherapy: a five-points fixation mask and a radiotherapy table. To have high quality MR images and to fit with the radiotherapy table and mask, a flexible anterior Body coil and the lower part of a Head coil were used.

Proton density (PD) was obtain from the magnetization at the origin M_0, considering a very low T1 and a low flip angle, we can write:

$$M_0 = G \circ PD$$

with G the coils sensitivity maps[5] and ° the Hadamard product.

To determine the best algorithm for coil sensitivity correction, we used a calibration phantom (SpinSafety-T04, Rennes, France). The phantom tubes were filled with a D_2O dilution in water for which we knew the theoretical proton density and were able to draw linear regression with the pixel's intensities. In-line (Prescan and B1-filter) and post-processing (SENSE, ESPIRiT and adaptive reconstruction) algorithms were compared. The sequence used to obtain the magnetization at the origin, is an ultra-short echo time (30 µs) spiralVIBE. These regressions also showed the linear relationship between proton density and MRI pixels intensity. Before any other steps, the MR images were registered on the CT with a Matlab® algorithm.

The CT and the sCT were compared with the computation of the Mean Absolute Error (MAE). CT images-based dose calculation and distribution were computed using MONACO software (Elekta®) using a set of dose-volume constraints for the organs at risk (OAR) and Planning Target Volumes (PTV). Dose were then recomputed on

sCT images, keeping the same deliverable segments. The results are compared with several indexes.

Firstly, the dose-volume histograms were extracted for each patient, the range, mean, standard deviation and percentage of dose difference were computed for several OARs and for PTVs. Comparisons between sCT and CT-based dosimetry were done using a Bland Altman analysis.

Secondly, the gamma index[6] was computed. It evaluates the distance difference and the dose difference for each point of the reference dose planning on the CT with the sCT-based planning. A dose difference criterion δD (DD in %) and a distance to agreement criterion δr (DTA in mm) are set for the evaluation of the gamma index, in this work 1%/1mm, 2%/2mm and 3%/3mm gamma indexes were evaluated. The gamma pass rate $\gamma_{\%}$ was also computed and evaluates the percentage of gamma indexes bellow 1.

3 Results

From the images of 26 patients, the images of two patients were not considerate, because the MRI acquisition box was misplaced leading to a part of the face truncated.

The algorithm with the best results for the coil sensitivity correction was the Prescan, Figure 1 shows the correspondent linear regression, compared to SENSE algorithm.

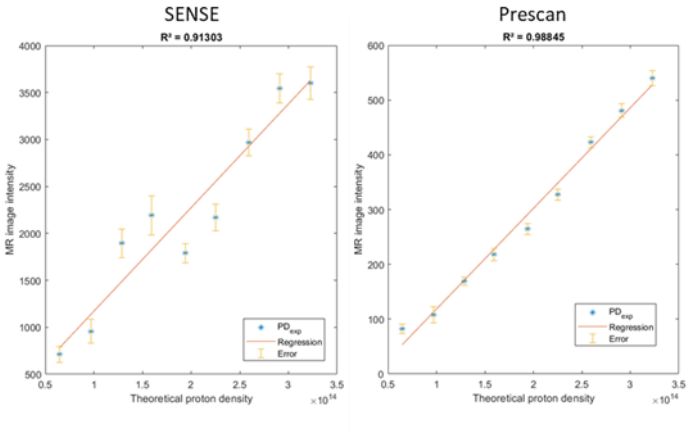


Figure 1: linear regression between MRI intensity corrected with SENSE algorithm (left) and Prescan algorithm (right) and theoretical proton density for 9 tubes with different D_2O concentrations

The mean MAE between the sCT and CT of the patients was 124.6 HU ± 23.1 HU which, as a percentage of the total range, is equivalent to 4.2 % ± 0.8 %. An example of a sCT is given Figure 2, as a comparison to the CT.

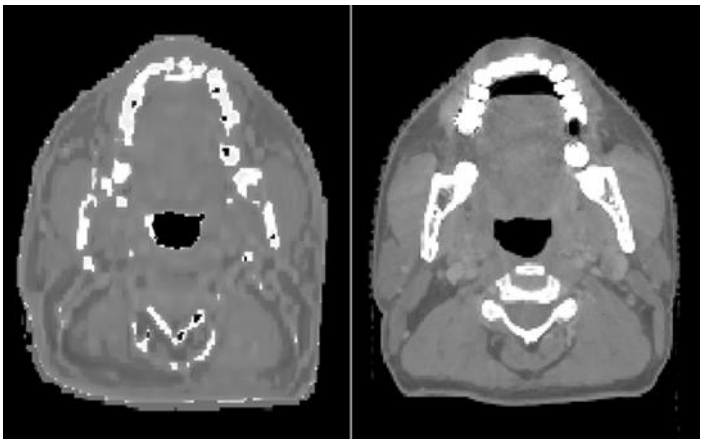


Figure 2: sCT (left) and CT (right) of a cavum patient

The dose-volume histograms showed a low difference between CT and sCT-based dosimetry. For the PTV 54.25 Gy (i.e the PTV for which the prescribed dose was 54.25 Gy) and the PTV 70 Gy, the difference between the mean median dose were 0.2 ± 0.2 Gy (0.57% of the prescription dose 70 Gy) and 0.0 ± 0.5 Gy respectively. For the OARs, the results are presented in the Table 1 as the CT-based dosimetry mean minus the sCT-based dosimetry mean of the maximum dose.

Organs	Brainstem	Spinal cord	Monolateral parotid
Dose difference in Gy (% of 70 Gy)	0,46 (0,66 %)	0,38 (0,54 %)	0,86 (1,23 %)

Table 1: absolute dose difference between CT-based mean of the maximum dose and sCT-based mean of the maximum dose for several OARs.

The Bland Altman analysis showed a low bias for most of the OARs (between -0.39 and 0.4 Gy) and for the PTVs (-0.04 Gy), the organ POST showed a bias of 1 Gy. The variability was very low for the tumor volumes (3.17 %), it was higher for some organs at risk such as the brainstem (20.94 %), the lips (22.70 %) and POST (21.05 %).

The mean gamma indexes were below 1 for every DD/DTA criteria, and with low standard deviation. For the tightest criteria of 1%/mm, the mean gamma was 0.21 ± 0.052, with a gamma pass rate of 93 ± 2.8 %. The maximum gamma pass rate was 98 ± 1.5 % for the 3%/3mm criterion. An example of a gamma index map is given Figure 3 for each criterion.

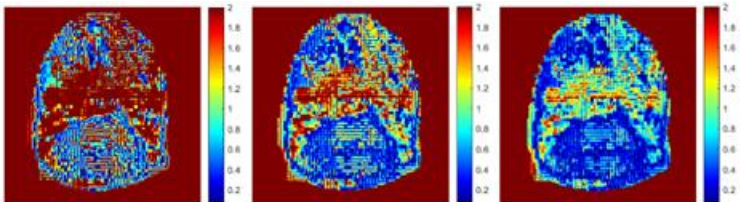


Figure 3: gamma indexes map for 1%/1mm (left), 2%/2mm (middle) and 3%/3mm (right) criteria, for a larynx patient

4 Discussion

The UTE spiralVIBE sequence is a sequence with a relatively short acquisition duration (4 minutes), allowing to generate a synthetic CT only based on the intensities of the pixels, the short echo time also allows to capture the signal short-T2 tissues such as the bone. This sequence is also less sensitive to B_1 field inhomogeneities. On the other hand, being a Work In Progress, this sequence may not be available yet on all systems and machines. Also, the fat inside the bone and the body fat proton densities are hard to dissociate, leading to low HUs for the bone, which is visible on the sCT. A solution would be to use a different regression model for bone or to use another MRI sequence to obtain a T2* map on which the segmentation would be easier.

The dosimetry demonstrated that the results were consistent, the differences with the reference CT-based dosimetry being very low. The MAE obtained are similar to the literature for the head and neck region, Guerreiro et al.[7] report a MAE of 116.3 HU with a multi-atlas-based method and Johansson et al.[8] report a MAE of 137 HU with a voxel-based method.

The acquisition protocol overall duration was only 11 minutes with the clinical sequences and the UTE sequence, which is compatible with the clinic. The use of the immobilization mask reduces the registration shifts, and allows to obtain images in the treatment position. The sCT generation is a fast process, even with the registration step, it required around 3 minutes per patient. The drawbacks of the radiotherapy immobilization material are the difficulty to fit within usual head and neck coils and the need for an adapted MR coil set up and patient elevation. In our specific case, the table on which the patient was laid was elevated of about 5cm to fit within the lower part of the head coil. The loss of quality compared to a regular fit inside the coil, due to the distance from the coil, was widely offset by the presence of shim coils. They improve a lot the quality of the images, compared to the use of a body coil only. Due to the immobilization mask, the patient cannot move too much but they still can swallow which induces artifacts. The spiraled acquisition of the UTE sequence in known to be less sensitive to this type of artifacts.

The dosimetry results obtained with our method can be compared to the literature. First, the dose difference of the PTVs is close to the results of the literature, with a mean of 0.27% of the prescription dose for the density assignment method of Karotki et al.[9]. For the OARs, Kazemifar et al.[10] obtained a 1% dose difference for the maximum dose of the brainstem with a deep learning method, which is higher than the 0.6% we obtained.

Then, the gamma index also gave results comparable to the literature. Guerreiro et al.[7] obtained 94 % ± 0.7 % and 98.4 % ± 0.3 % gamma pass rates for a 2%/2mm and a 3%/3mm criterion respectively, our gamma pass rates were very similar, with a slightly higher standard deviation. Johansson et al.[8] obtained a 94 % gamma pass rate for a 1%/1mm criterion, which is also very close to our 93 %.

MRI planning allowed equivalent dosimetry with to the CT and was not much longer for the patients. For this work, the organs at risk contouring was done automatically on CT images, as commonly used in the routine. The possibility to contour these volumes on an MR image could allow an MRI only workflow, which would be much easier for the patients: an 11 minutes protocol would be enough for volumes contouring and dose planning, avoiding extra-radiations for the patients and allowing more accuracy in delineation[11,12] for the medical staff. Compared to deep learning methods which cannot adapt to every anatomy, this work could also open the gate for a more personalized treatment with sequences allowing dose painting for example. The drawback is the pixels intensity-based segmentation, it is more sensitive to classes overlapping especially between bone and air.

5 Conclusion

This study shows that proton density quantification with MRI allows to build a synthetic CT with a direct method. The dosimetry shows results comparable to other methods in the literature and to the reference CT-based dosimetry. This method is a first step towards personalized treatment that could be adapted to other anatomical area with the proper UTE sequence and a coil setup compatible with the immobilization material.

Acknowledgments

This study was funded by the LabEx PRIMES (ANR-11-LABX-0063), the ANR HERESP (ANR-22-CE19-0027) and performed within the framework of the SIRIC LyriCAN+ grant (INCa INSERM DGOS 12563).

References

[1] Vanquin L, Boydev C, Korhonen J, Rault E, Crop F, Lacornerie T, et al. Planification de la radiothérapie du cancer de la prostate par l'imagerie par résonance magnétique. Cancer/Radiothérapie. 2019 Jul;23(4):281–9. https://doi.org/10.1016/j.canrad.2018.09.005

[2] Alvarez Andres E, Fidon L, Vakalopoulou M, Lerousseau M, Carré A, Sun R, et al. Dosimetry-Driven Quality Measure of Brain Pseudo Computed Tomography Generated From Deep Learning for MRI-Only Radiation Therapy Treatment Planning. International Journal of Radiation Oncology*Biology*Physics. 2020 Nov;108(3):813–23. https://doi.org/10.1016/j.ijrobp.2020.05.006

[3] Chin AL, Lin A, Anamalayil S, Teo BKK. Feasibility and limitations of bulk density assignment in MRI for head and neck IMRT treatment planning. Journal of Applied Clinical Medical Physics. 2014;15(5):100–11. https://doi.org/10.1016/j.canrad.2018.09.005

[4] Demol B, Viard R, Reynaert N. Monte Carlo calculation based on hydrogen composition of the tissue for MV photon radiotherapy. Journal of Applied Clinical Medical Physics. 2015 Sep;16(5):117–30. https://doi.org/10.1120/jacmp.v16i5.5586

[5] Mezer A, Rokem A, Berman S, Hastie T, Wandell BA. Evaluating quantitative proton-density-mapping methods: Proton-Density Mapping. Hum Brain Mapp. 2016 Oct;37(10):3623–35. https://doi.org/10.1002/hbm.23264

[6] Hussein M, Clark CH, Nisbet A. Challenges in calculation of the gamma index in radiotherapy – Towards good practice. Physica Medica. 2017 Apr;36:1–11. https://doi.org/10.1016/j.ejmp.2017.03.001

[7] Guerreiro F, Burgos N, Dunlop A, Wong K, Petkar I, Nutting C, et al. Evaluation of a multi-atlas CT synthesis approach for MRI-only radiotherapy treatment planning. Physica Medica. 2017 Mar;35:7–17. https://doi.org/10.1016/j.ejmp.2017.02.017

[8] Johansson A, Karlsson M, Nyholm T. CT substitute derived from MRI sequences with ultrashort echo time. Medical Physics. 2011;38(5):2708–14. https://doi.org/10.1118/1.3578928

[9] Karotki A, Mah K, Meijer G, Meltsner M. Comparison of bulk electron density and voxel- based electron density treatment planning. J Appl Clin Med Phys. 2011 Nov 15;12(4):97–104. https://doi.org/10.1120/jacmp.v12i4.3522

[10] Kazemifar S, Barragán Montero AM, Souris K, Rivas ST, Timmerman R, Park YK, et al. Dosimetric evaluation of synthetic CT generated with GANs for MRI- only proton therapy treatment planning of brain tumors. J Appl Clin Med Phys. 2020 May;21(5):76–86.

[11] Njeh CF. Tumor delineation: The weakest link in the search for accuracy in radiotherapy. J Med Phys. 2008;33(4):136–40. https://doi.org/10.4103/0971-6203.44472

[12] Rasch C, Keus R, Pameijer FA, Koops W, de Ru V, Muller S, et al. The potential impact of CT-MRI matching on tumor volume delineation in advanced head and neck cancer. Int J Radiat Oncol Biol Phys. 1997 Nov 1;39(4):841–8. https://doi.org/10.1016/s0360-3016(97)00465-3

IOeRT conventional and FLASH TPS implementation exploiting fast GPU Monte Carlo: the case of breast cancer

G. Franciosini[1,2], A. De Gregorio[3,2], M. Di Francesco[4], F. Di Martino[5,6,7,8], G. Felici[4], J. Harold Pensavalle[4,5,6], M. Marafini[9,2], A. Muscato[2,10], V. Patera[1,2], P. Poortmans[11,12], A. Sciubba[13], A. Schiavi[1,2], M. Toppi[1,2], G. Traini[2], A. Trigilio[13,3], and A. Sarti[1,2]

[1]Sapienza, University of Rome, Department of Basic and Applied Engineering Sciences, Rome, Italy
[2]National Institute of Nuclear Physics, INFN, Section of Rome I, Rome, Italy
[3]Sapienza, University of Rome, Department of Physics, Rome, Italy
[4]S.I.T. Sordina IORT Technologies S.p.A, Aprilia, Italy
[5]Centro Pisano Multidisciplinare sulla Ricerca e Implementazione Clinica della Flash Radiotherapy (CPFR), Pisa, Italy
[6]University of Pisa, Department of Physics, Pisa, Italy
[7]Azienda Ospedaliero Universitaria Pisa (AOUP), Fisica Sanitaria, Pisa, Italy
[8]National Institute of Nuclear Physics, INFN, Section of Pisa, Pisa, Italy
[9]Museo Storico della Fisica e Centro Studi e Ricerche "E. Fermi", Rome, Italy
[10]Specialty School of Medical Physics, La Sapienza University of Rome, Rome, Italy
[11]Department of Radiation Oncology, Iridium Netwerk, Antwerp, Belgium
[12]University of Antwerp, Faculty of Medicine and Health Sciences, Antwerp, Belgium
[13]National Institute of Nuclear Physics, INFN, Frascati National Laboratories (LNF), Frascati, Italy

Abstract Breast cancer is widely treated with IOeRT, (Intra Operative electron RadioTherapy) in the case of partial breast irradiation treatments, yielding satisfactory results; one of the main limitations of this technique is the absence of a Treatment Planning System (TPS) that could significantly enhance proper coverage of the target volume during irradiation. In this contribution, we present an IOeRT TPS that has been developed using an in-house fast GPU-based Monte Carlo (MC), and an ultrasound imaging system. This tool aims to provide the best irradiation strategy (electron beam energy, applicator position, and bevel angle) within the limited amount of time allowed to keep the patient on the surgery bed, and facilitate the optimization of dose prescription and delivery to the target volume. The study was conducted in-silico, exploiting MC simulations of a breast cancer treatment. Ultrasound-based input was employed to compute absorbed dose maps for different irradiation strategies, and a quantitative comparison was performed analysing dose maps and the corresponding Dose Volume Histograms. The system demonstrated the capability to identify the optimal strategy within a few minutes. The overall computation time was found to be compatible with clinical implementation. The impact of ultra high dose rate irradiations on a possible additional sparing of the Organs at Risk was modelled as well, when optimizing the biological dose, integrating the benefits from minimally invasive surgery techniques.This long awaited planning tool will boost the treatment efficacy, overcoming the limits identified within the clinical trials carried out so far, and pave the way for the first implementation of the FLASH effect into the IOeRT clinical workflow.

1 Introduction

Among the different cancer treatment modalities available today, Intra Operative Radiation Therapy, delivered with electrons (IOeRT), allows to achieve a very high dose conformity to the target volume maximising the local treatment efficacy and thus decreasing the risk of local cancer recurrence [1]. It is a technique that involves precise delivery of a dose of ionizing radiation directly to the tumour bed after the surgical tumor mass removal. Currently, IOeRT irradiations are mainly performed by a dedicated mobile linear accelerator close to the surgical beds, with energies between 4 and 12 MeV. Within IOeRT, whenever needed and possible, temporary beam modifiers, such as the Radio Protection Disc (RPD) used for the treatment of Partial Beast Irradiation (PBI), are used to protect the underlying healthy tissues during the irradiation and the beam is passively collimated by means of a PMMA hollow tube (called applicator).

Despite the technique high efficacy, the planning technology has not evolved since its conception, being outdated in comparison to the current state of the art in other external Radiotherapy (RT) techniques and therefore slowing down the IOeRT clinical potential.

The absence of proper imaging, treatment planning and applicator docking or position checking is directly related to the underdosage or geographical miss of the Clinical Target Volume (CTV) as reported by the ELIOT trial [2]. The treatment planning, up to now, was not foreseen in the clinical workflow for two main fundamental reasons: the lack of 'post-surgery' imaging, needed to allow for the dose calculation, and the time needed to produce dose maps that are needed in the optimisation process. Such planning tool becomes additionally interesting in view of the recent experimental evidence of Organs At Risk (OARs) sparing in presence of treatments delivered at Ultra High Dose Rates (UHDR) (>40 Gy/s, FLASH effect [3]). The use of monoenergetic high-intensity pulses of electrons (needed to deliver as fast as possible the whole dose) makes IOeRT the current best candidate for the first clinical implementation of the FLASH effect. In this contribution we explored the possibility to use a fast GPU-based Monte Carlo (MC) software, FRED (Fast paRticle thErapy Dose evaluator) [4–6], in combination with an ultrasound (US) imaging input (PTV) to perform a full IOeRT treatment plan optimisation in the limited amount of time allowed by the surgical procedure. The case of PBI was taken as an example for this feasibility study, as it represents the largest fraction of patients treated with IOeRT nowadays.

The obtained results are discussed also in view of the possible applications to FLASH therapy.

2 Materials and Method

The IOeRT planning involves rapid intra-operative imaging and fast MC simulations, performed with FRED, for the computation of absorbed dose maps when exploring various irradiation configurations.
Within FRED, particles interactions are described with the precision level available in bleeding edge MC tools used for medical physics applications, with the advantage of reducing the simulation time up to a factor 1000. In this way, it is possible to performa full MC plan recalculation in few minutes on a single GPU (Graphics Processing Unit) cards.
S.I.T. Sordina IORT Technologies S.p.A., provided an intra-operative imaging acquired using 3D real-time ultrasound system, resulting in the clinical workflow illustrated in Fig. 1. This system is able to provide the positions of the tissues, post-tumor resection and of the protective disk whenever implanted.
The acquired imaging is the starting block for the exploration of different treatment configurations, such as beam energies, applicator positions, dimensions, and bevel angles. MC simulations are then compared for target volume coverage and optimal OARs sparing through dose maps and Dose Volume Histograms (DVHs) analysis.
As patient related US imaging data are not yet available, to simulate real conditions, we used a CT image of an anthropomorphic RANDO male phantom with a silicone breast prosthetic application, modified to account for the expected US resolution. All the voxels belonging to the patient were assigned to the water Hounsfield unit (HU) while outside the patient the air HU was used. For the plan optimisation and evaluation it is crucial to have a proper definition of the PTV and of the OARs. We defined the ROIs for the study of the PBI optimisation according to the clinical workflow characteristic of these treatments. According to the US expected spatial resolution, Regions Of Interest (ROIs) with a depth of no more than 7 cm (equal to the US optimal viewing) have been defined. The PTV was approximated as a spherical region with a mean radius of 2.5 cm, positioned at a depth of 1.3 cm from the skin, while the RPD as a cylinder (radius of 3.9 cm) of a 3 mm thick layer of PTFE (HU=950), followed by 3 mm of stainless steel (HU=10000). The skin was introduced with a thickness of 5 mm. The Pectoral Muscle was structured as a cylindrical-shaped region (1.2 cm thick, mean radius of 3.9 cm) positioned just after the RPD, while the breast region was defined to encompass all the tissues between the Skin and the RPD surrounding the PTV. All these ROIs are shown in Fig.2, panel A. In addition to that, also the surgical breach and the tissue deformation needed to insert the applicator were accounted for in the geometrical setup.

3 Results

The best plan selection involved computing DVHs, as well as maximum and mean dose values absorbed by each ROI. Based on the dimension, position and thickness of the PTV considered in our simulations and on the clinical experience of PBI treatments, we defined 8, 10 and 12 MeV as options to be investigated for the beam energy and applicator diameters ranging from 40 to 70 mm to simulate the possibility to consider a small or large safety margin. An automated applicator position scan around the surgical breach was also performed to determine the best applicator position. For this study, the dose map calculation and optimization assumed a dose prescription of 20 Gy to 90% of the PTV according to the standard IOeRT procedure. Fig.2, panel B and C, shows the dose map distributions obtained for the energy and applicator dimensions scan. All the dose maps were normalized to fulfill the dose prescription. As can be observed looking at the dose maps, 8 MeV beam does not provide an adequate PTV coverage defined as, here and in the following, the fraction of the PTV volume that receives 95% of the goal dose (D95). The values of D95 for 8, 10 and 12 MeV are respectively 92%, 96% and 97%. At 12 MeV a significant increase of the dose absorbed by the pectoral muscle region is observed, while at 10 MeV a proper PTV coverage is obtained, with a reduced dose to the pectoral muscle. These results have been confirmed by the DVHs and mean/max dose values analysis. For what concern the applicator diameter scan the D95 values obtained are 93%, 96%, 96% and 97% for the 40, 50, 60 and 70 mm respectively demonstrating that when using a 40 mm internal diameter, it is not possible to achieve the required PTV coverage, as the applicator diameter is approximately equal to the PTV mean radius. The dose absorbed by the Breast region and to the skin grows as expected when using larger applicators, accordingly with the geometry and positioning of the ROIs. Finally, the optimisation of the applicator position was also performed by means of an automated scan around the surgical breach. The beam centre, impinging on the phantom along the y direction, was moved from the centred position in a (x,z) grid with 10 mm steps. Fig. 2, panel D, shows the dose maps in Gy units for the centered and worst-case plans.
Once all the MC simulations are performed the best plan selection algorithm implements the calculation of mean and maximum doses in the OARs and returns the optimal configuration ensuring the proper PTV coverage while lowering the dose to the OARs as much as possible. For what concerns the optimization process time we were capable to optimize the treatment in few seconds, and thus in a time window completely compatible with the few minutes available during the surgery. As an example, the simulation time needed to explore three different beam energies and, for each energy, nine different applicator positions, is $\sim$ 1 minute. PBI treatment aligns well with FLASH effect requirements, being compatible with the high-dose rates used to deliver

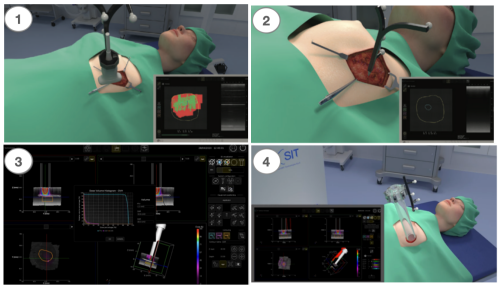


Figure 1: IOeRT procedure: after the tumor mass removal and positioning of the RPD disc (1) by using a US probe, which is referenced in the space by means of an optical tracker, the surgeon acquires US images; (2) the surgeon identifies the surface to be treated and regions of interest (PTV and OARs); (3) different treatment configurations are explored looking for the optimal one. The whole process takes less than 2 minutes. Image guided docking happens at the end and the dose is recalculated in the final position (4).

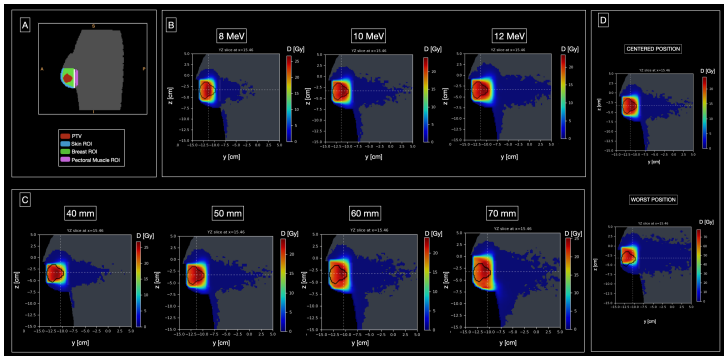


Figure 2: Panel A: Visualisation of anthropomorphic phantom CT. The CT has voxels dimension of 0.2 cm in all the directions. The ROIs are also highlighted. Panel B and C: Plan comparison with different beam energies and applicator dimensions. Dose maps obtained from the IOeRT plan with an 8, 10 and 12 MeV electron beam collimated by a 50 mm diameter applicator are reported on panel B, whereas the ones obtained by changing only the applicator dimension (ranging from 40 to 70 mm) with a fixed beam energy equal to 10 MeV are shown on panel C. Panel D: Dose maps comparison between the centered (upper) and worst-case (bottom) plans, performed with 10 MeV electron beam collimated by 50 mm applicator.

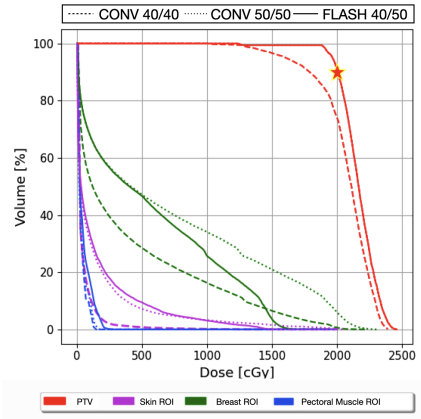


Figure 3: Biological effective dose volume histograms for the PTV (red), Breast ROI (green), Pectoral Muscle ROI (blue) and Skin (purple) for the simulations performed in conventional regime (CONV) with a 40 mm (dashed line, 40/40) and 50 mm (dotted line, 50/50) surgical breach and applicator dimension, and in FLASH regime with a small surgical breach 40 mm, and a larger applicator diameter 50 mm (solid line, 40/50).

20 Gy in a single fraction. In our study we implemented the FLASH-modifying factor (FMF) model that was introduced by Bohlen et al. in [7]. On the basis of the results obtained from mini-pig skin treatments [8] we selected values of FMF^{min} and D_{Th} equal to 0.65 and 6 Gy, respectively. According to the dose absorbed by each voxel in the full treatment, an FMF re-weighed dose has been calculated to evaluate the FLASH potential. In IOeRT the FLASH effect can be exploited to combine a minimally invasive surgery technique, which avoids postoperative complications, with the very precise delivery of IOeRT irradiation, thus reducing the probability of tumor local recurrence. Fig.3 shows the comparison of 3 different DVHs: the dashed line represents the "current clinical workflow", i.e. a small applicator exactly matching the surgical breach of 40 mm (CONV 40/40), the dotted line represents the correct PTV treatment, i.e. large surgical breach and thus applicator dimension equal to 50 mm (CONV 50/50), while the solid line describes the FLASH IOeRT minimally invasive treatment, i.e. small surgical breach (40 mm) and a larger applicator dimension equal to 50 mm (FLASH 40/50). As can be seen, all the OARs, in FLASH configuration, have reduced effective doses, when compared to the conventional one, with the additional advantages of minimizing the surgery impact without causing significant damage on the patient skin thanks to the FLASH effect sparing, while ensuring the correct PTV coverage.

4 Discussion

The obtained results confirm the feasibility of the developed tool for an online treatment optimization. Absorbed dose maps can be quickly generated and studied selecting a plan that achieves the necessary PTV coverage while minimizing OAR doses. The results emphasize the importance of an automated procedure to identify the best treatment position. Through image-guided docking of the applicator, ensuring the treatment alignment with the plan, the impact of geographical misses will be reduced, enhancing the treatment efficacy. This approach can also be used to explore and assess the advantage of delivering beams at ultrahigh dose rates, leveraging the FLASH effect finally helping in avoiding invasive surgical procedures while maintaining treatment efficacy [9].

5 Conclusion

The ability to compute and optimize an IOeRT plan in a few minutes using a single GPU card has been demonstrated. This streamlines the preparation, recording, and verification of IOeRT procedures, bringing it closer to the level used for external beam RT. IOeRT plan optimization addresses a current limitation of the technique, specifically the risk of underdosage of the PTV due to the restricted surgical breach when RPD is necessary. The proposed clinical workflow enables radiation oncologists and breast surgeons to assess the actual need for RPD insertion, evaluate dose distribution, and ensure PTV coverage by adjusting the applicator diameter. The software tools also support exploring the potential of UHDR irradiations. For PBI, the FLASH effect presents an opportunity to redefine the workflow for IOeRT treatments, allowing direct applicator placement on part of the patient's skin or eliminating the need for RPD and finally improving the patient's quality of life.

References

[1] F. Roeder, V. Morillo, L. Saleh-Ebrahimi, et al. "Intraoperative radiation therapy (IORT) for soft tissue sarcoma – ESTRO IORT Task Force/ACROP recommendations". *Radiother Oncol* 150 (2020), pp. 293–302. DOI: https://doi.org/10.1016/j.radonc.2020.07.019.

[2] G. Fastner, C. Gaisberger, J. Kaiser, et al. "ESTRO IORT Task Force/ACROP recommendations for intraoperative radiation therapy with electrons (IOERT) in breast cancer". *Radiother Oncol* 149 (2020), pp. 150–157. DOI: 10.1016/j.radonc.2020.04.059.

[3] M.-C. Vozenin, J. Hendry, and C. Limoli. "Biological Benefits of Ultra-high Dose Rate FLASH Radiotherapy: Sleeping Beauty Awoken". *Clin Oncol* 31.7 (2019), pp. 407–415. DOI: 10.1016/j.clon.2019.04.001.

[4] A. Schiavi, M. Senzacqua, S. Pioli, et al. "Fred: a GPU-accelerated fast-Monte Carlo code for rapid treatment plan recalculation in ion beam therapy". *Phys Med Biol* 62.18 (Sept. 2017), p. 7482. DOI: 10.1088/1361-6560/aa8134.

[5] M. De Simoni, G. Battistoni, A. De Gregorio, et al. "A Data-Driven Fragmentation Model for Carbon Therapy GPU-Accelerated Monte-Carlo Dose Recalculation". *Front Oncol* 12 (2022). DOI: 10.3389/fonc.2022.780784.

[6] G. Franciosini, G. Battistoni, A. Cerqua, et al. "GPU-accelerated Monte Carlo simulation of electron and photon interactions for radiotherapy applications". *Phys Med Biol* 68.4 (Feb. 2023), p. 044001. DOI: 10.1088/1361-6560/aca1f2.

[7] T. T. Bohlen et al. "Normal Tissue Sparing by FLASH as a Function of Single-Fraction Dose: A Quantitative Analysis". *Int J Radiat Oncol Biol Phys* (2022). DOI: 10.1016/j.ijrobp.2022.05.038.

[8] M.-C. Vozenin, P. De Fornel, K. Petersson, et al. "The Advantage of FLASH Radiotherapy Confirmed in Mini-pig and Cat-cancer Patients". *Clin Cancer Res* 25.1 (Jan. 2019), pp. 35–42. DOI: 10.1158/1078-0432.CCR-17-3375.

[9] J. Bourhis, W. J. Sozzi, P. G. Jorge, et al. "Treatment of a first patient with FLASH-radiotherapy". *Radiother Oncol* 139 (2019). FLASH radiotherapy International Workshop, pp. 18–22. DOI: 10.1016/j.radonc.2019.06.019.

Exploring VHEE treatment efficacy through in-silico investigations in conventional and FLASH irradiation regimes

A. Muscato[1,2], D. Carlotti[4,3], A. De Gregorio[3,2], M. Fiore[4], G. Franciosini[1,2], T. Insero[4], V. Marè[4], S. Ramella[4], A. Sarti[1,2], A. Schiavi[1,2], M. Schwarz[5] and V. Patera[1,2]

[1]Dipartimento di Scienze di Base e Applicate per l'Ingegneria, Sapienza Università di Roma, Rome, Italy
[2]Istituto Nazionale di Fisica Nucleare (INFN) Sezione di Roma I, Rome, Italy
[3]Dipartimento di Fisica, Sapienza Università di Roma, Rome, Italy
[4]Radiation Oncology, Fondazione Policlinico Universitario Campus Bio-medico, Rome, Italy
[5]Radiation Oncology Department, University of Washington & Fred Hutchinson Cancer Center, Seattle, CA, United States

Abstract Radiotherapy (RT) is crucial for treating deep-seated tumors; options include photon RT, particle therapy (PT) with protons or heavier ions like ^{12}C. Very High Energy Electron (VHEE) beams (100-200 MeV) were proposed for deep tumors, but clinical implementation faced challenges. Recent advancements in compact C-band electron acceleration and FLASH radiotherapy have renewed the interest in this option. FLASH, demonstrated in pre-clinical studies, reduces healthy tissue toxicity while maintaining cancer-killing efficacy whenever the dose rates are radically increased (~40 Gy/s) with respect to current conventional radiotherapy (~0.1 Gy/s or less). This study explores the potential impact of electron beam therapy for the treatment of deep-seated tumors focusing on the stereotactic treatment of pancreatic cancer. Results are presented and compared against conventional radiotherapy performed with a VMAT approach.

1 Introduction

Electrons of 'very' high energy (VHEE, E above 50 MeV) have been already explored in the past for the treatment of deep-seated tumors [1]. The obtained results showed a tumour coverage and Organs At Risk (OARs) toxicities comparable to advanced proton (PT) or photon radiotherapy (RT) only at the cost of implementing multiple fields and/or very high electron energies (above 100 MeV). Challenges in deploying such beams, combined with limited therapeutic gains, hindered the technique development. Recent advancements in compact, high-gradient electron acceleration and experimental evidence of organ sparing in Ultra-High Dose Rate (UHDR) irradiation modalities, the so-called "FLASH" regime, have the potential to reshape this landscape. Compact accelerators, a few meters long, could deliver required electron energies for stereotactic treatments, capitalizing on FLASH's organ-sparing benefits. This potential could allow the simplification of irradiation geometry, also reducing the number of fields needed for patient treatment.

2 Materials and Methods

To evaluate the potential of VHEE treatments implementing the FLASH effect, we compare the results obtained optimizing treatment plans with electrons of energy below 130 MeV in the case of three patients (PT1, PT2, PT3) affected by a pancreatic cancer, with the plans delivered using photons according to Volumetric Modulated Arc Therapy (VMAT) technique, implemented using 2-3 arcs. All the patients underwent a treatment with very high dose per fraction, the so-called Stereotactic Body Radiation Therapy (SBRT) condition, with 6 Gy or more absorbed dose per session delivered in target lesions that are of few centimeters in size. Treatments, carried out at the Campus Biomedico Hospital, were carried out using a TrueBeam Linear Accelerator, produced by Varian.

In this study, we employed FRED [2] Monte Carlo (MC) software, specifically designed for GPU (Graphics Processing Unit) operation. This facilitated the simulation of absorbed dose in electron treatment with full MC accuracy, all accomplished within a few hours.

FRED requires CT information, the electron's origin points relative to CT position, and beam specifications (energy, lateral and angular spread) at the production point to calculate the absorbed dose maps. An Intensity-Modulated Radiation Therapy (IMRT) -like configuration was employed to plan the geometry of the electron fields, in order to define their number and their respective directions.

In our simulations, we adopted beam characteristics consistent with conventional electron linear accelerator solutions with energies exceeding 50 MeV [3]. The beam exhibits a transverse size of ~ mm and a divergence of ~mrad. Furthermore, we limited our analysis to the relatively low energy range of VHEE, specifically, 70 to 130 MeV; we assumed an active scanning-like beam delivery technology [4], akin to the approach presently employed in proton therapy (PT).

Once absorbed dose maps were obtained with the full MC simulation approach for each PB in the treatment plan, the fluence of each PB was optimized to ensure coverage of the required Planning Target Volume (PTV) while sparing organs at risk (OAR). Specifically, an optimization was implemented to do this by minimizing a voxel-based objective function, defined below:

$$\chi^2 = \sum_{i \epsilon PTV} \omega_i \frac{(d_i - D_{PTV})^2}{d_i^2} + \sum_{i \epsilon OAR} \omega_i \frac{(d_i - D_{OAR})^2}{d_i^2} g(d_i - D_{OAR})$$

where the weights are adjusted euristically to match the dosimetric constraints.

To reduce the computational cost, we adopted two different minimization methods: Steepest Descent, as observed in [5] and Quantum Simulated Annealing [6].In the end, the evaluation of the FLASH effect potential was performed after the treatment optimization. Once the dose maps were obtained, the Flash Modifying Factor (FMF) [6] was introduced to assess the reduction in radiation-induced toxicity.

3 Results

To evaluate and compare the effectiveness of VHEE treatment plans, we used the common tools available and used in radiotherapy: Dose Volume Histograms (DVHs) and isodose maps have been produced to verify the dosimetric constraints on OAR and PTV and finally ensure the clinical acceptability of the plan.
The first patient (PT1) underwent homogeneous stereotactic treatment, requiring a V(95%) coverage greater than 95% for plan approval, accepting "hot spots" below 106% of the prescribed dose. In contrast, the second and third patients (PT2, PT3) received heterogeneous stereotactic treatment, accepting target coverage of 100% of the dose in at least 95% of the volume and allowing "hot spots" up to 126%, promoting the destruction of tumor cells through necrosis. The common goal for all three cases was to achieve adequate PTV coverage while simultaneously limiting the dose to various OARs, including the duodenum, as shown in Figures 1 and 2 and referring to Table 1, we can observe the dosimetric constraints.
From these preliminary studies, it has been observed that VHEEs are competitive compared to advanced techniques such as VMAT, demonstrating good target coverage while limiting the dose to various organs at risk, such as the duodenum. The selection of these three cases was not random, as they involved the use of high doses over a few sessions, making them suitable candidates for studying the effectiveness of the FLASH effect in sparing various OARs. Specifically, by selecting parameters to trigger the FLASH effect, modeled according to [7], an increase in OARs dose sparing, especially in the duodenum, has been be observed. Such outcome has the clear potential of enhancing the treatment's efficacy leading to either a dose escalation in the PTV or improving the duodenum sparing.

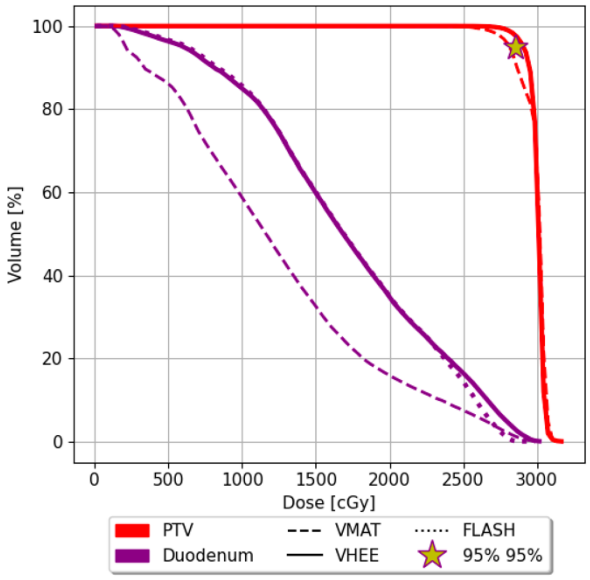


Figure 1. Analyzing the variations in PT1 treatment utilizing VMAT and VHEE, both with and without the FLASH effect. Specifically, we are studying the FLASH effect as outlined in [7], employing a D_{th} of 22.5 Gy and an FMF_{min} of 0.8. The reduction in dosage, facilitated by the implementation of the FLASH effect, is depicted by the dotted curve.

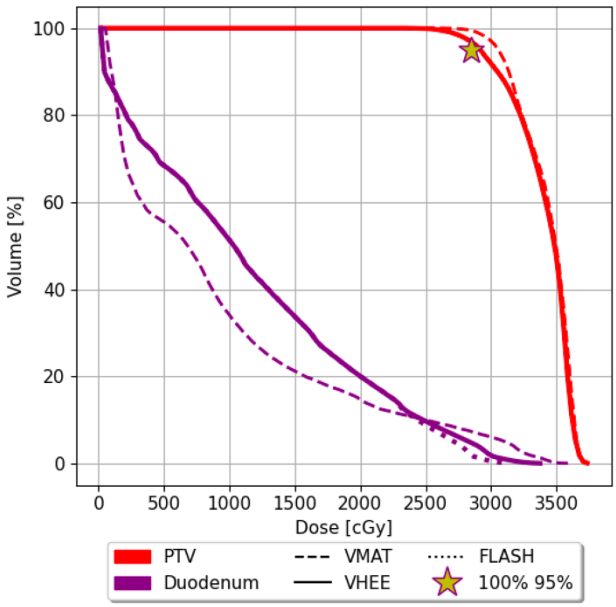


Figure 2. Analyzing the variations in PT2 treatment utilizing VMAT and VHEE, both with and without the FLASH effect. Specifically, we are studying the FLASH effect as outlined in [7], employing a D_{th} of 22.5 Gy and an FMF_{min} of 0.8. The reduction in dosage, facilitated by the implementation of the FLASH effect, is depicted by the dotted curve.

PT1				
Organ	Dosimetric constraint	VMAT	VHEE	FLASH
PTV	$V_{95\%}$>95%	97 %	98 %	98 %
Duodenum	D_{max}<33 Gy	30 Gy	30 Gy	28 Gy
PT2				
Organ	Dosimetric constraint	VMAT	VHEE	FLASH
PTV	$V_{100\%}$>95%	96%	95%	95%
Duodenum	D_{max}<35 Gy	33 Gy	34 Gy	31 Gy

Table 1. Dosimetric endpoints are either expressed as the fraction of volume (V_{XX}) that absorbs a given amount of dose (XX Gy), or as the value of the maximum absorbed dose (and the associated volume) and are shown for all the OARs considered in a pancreas treatment.

4 Discussion

The results presented in this contribution demonstrate how it could be possible to obtain clinically acceptable plans with VHEE with a limited number of fields and energy below 130 MeV. The implemented field configurations, including the number of fields and directions, are similar to those used in techniques such as IMRT. The analyzed cases were treated with stereotactic radiotherapy techniques, which are widely used today to treat small-sized lesions, utilizing high doses over a few treatment sessions. Given these considerations, it was observed that by using only VHEE, without introducing the FLASH effect, it is possible to achieve adequate coverage of the PTV while simultaneously limiting the dose to surrounding OARs. This makes the treatment plan competitive compared to plans obtained with commonly used techniques such as VMAT.
Furthermore, since the analyzed cases are excellent candidates for studying the effectiveness of the FLASH effect, it was noted that delivering the plan using UHDRs allows for better sparing of OARs compared to photons. In all three cases, it was crucial to preserve the OAR closest to the PTV, the duodenum, as an excess dose to it would have rendered the treatment ineffective. As observed, implementing the FLASH effect resulted in a lower dose to the duodenum whose size will finally depend on the experimental characterization of the FLASH effect sparing that is currently ongoing in a worldwide campaign.

5 Conclusion

The research here-before reported, explores the treatment of pancreatic tumors using VHEE. A comparison between VHEEs and VMAT plans revealed that VHEE can achieve performance levels similar to state-of-the-art irradiation techniques, even without additional sparing from the FLASH effect.
The findings suggest that, assuming reasonable conditions for triggering the FLASH effect, it could be effectively utilized to mitigate the impact on OARs surrounding the PTV. Across all cases, the FLASH effect demonstrated a clear potential for significantly improving the therapeutic window of External Beam Radiation Therapy (EBRT) treatments of pancreatic cancer.

References

[1] Bazalova-Carter M, Qu B, Palma B, Hårdemark B, Hynning E, Jensen C, et al. Treatment planning for radiotherapy with very high-energy electron beams and comparison of vhee and vmat plans. Med Phys (2015) 42:2615–25. doi:10.1118/1. 4918923

[2] Franciosini, G., et al. "GPU-accelerated Monte Carlo simulation of electron and photon interactions for radiotherapy applications." *Physics in Medicine & Biology* 68.4 (2023): 044001. DOI: 10.1088/1361-6560/aca1f2.

[3] Kokurewicz K, Brunetti E, Curcio A, Gamba D, Garolfi L, Gilardi A, et al. An Experimental Study of Focused Very High Energy Electron Beams for Radiotherapy. Commun Phys (2021) 4:33. DOI: 10.1038/s42005-021-00536-0

[4] Ronga MG, Cavallone M, Patriarca A, Leite AM, Loap P, Favaudon V, et al. Back to the Future: Very High-Energy Electrons (Vhees) and Their Potential Application in Radiation Therapy. Cancers (2021) 13:1–19. DOI: 10.3390/ cancers13194942

[5] Sarti, Alessio, et al. "Deep seated tumour treatments with electrons of high energy delivered at FLASH rates: the example of prostate cancer." *Frontiers in Oncology* 11 (2021): 777852. DOI:10.3389/fonc.2021.777852

[6] Eliahu Cohen e Boaz Tamir. “Quantum annealing–foundations and frontiers”.In: The European Physical Journal Special Topics 224.1 (2015), pp. 89–110.

[7] Bohlen TT, Germond JF, Bourhis J, Vozenin MC, Ozsahin EM, Bochud F, et al. Normal tissue sparing by flash as a function of single-fraction dose: A quantitative analysis. Int J Radiat Oncol Biol Phys (2022) 114:1032–44. A Red Journal Special IssueOligometastasis, Part 2. :10.1016/j.ijrobp.2022.05.038

Reconstruction Toolkit (RTK) v2, an Insight Toolkit (ITK) module for tomographic reconstruction

Simon Rit[1], Sébastien Brousmiche[2], Julien Finet[3], Gregory C. Sharp[4], and Philipp Steininger[5]

[1]INSA-Lyon, Université Claude Bernard Lyon 1, CNRS, Inserm, CREATIS UMR 5220, U1294, F-69373, Lyon, France
[2]Ion Beam Applications SA Louvain-la-Neuve, Belgium
[3]Kitware SAS, Villeurbanne, France
[4]Massachusetts General Hospital and Harvard Medical School, Department of Radiation Oncology, Boston, USA
[5]Research & Development, medPhoton GmbH, Salzburg, Austria

Abstract Since its first presentation at ICCR 2013, the Reconstruction Toolkit (RTK) has become a reference software for tomographic reconstruction of computed tomography (CT) images. In this period, RTK has become a remote module of the Insight Toolkit (ITK), thus enabling improved continuous integration and Python wrapping. New functionalities have also been developed, including simulation of Forbild phantoms, handling of cylindrical detectors and new iterative reconstruction algorithms for conventional CT, single photon emission CT (SPECT), four-dimensional (4D) CT and spectral CT. This article summarizes these eleven years of developments to today's RTK v2.

1 Introduction

The Reconstruction Toolkit (RTK) was first presented at the XVII-th international conference on the use of computers in radiotherapy (ICCR 2013) in Melbourne, Australia [1]. Initially, RTK was an open-source and cross-platform software for fast circular cone-beam CT reconstruction based on the Insight Toolkit (ITK) [2], "an open-source, cross-platform library that provides developers with an extensive suite of software tools for image analysis" (https://www.itk.org). RTK has been and keeps being a tool for researchers working in radiotherapy and radiology as well as a sofware used by companies for their cone-beam CT scanners. This paper summarizes the new features of RTK which have been developed in the past eleven years.

2 Software design

2.1 ITK remote module

The change of major version number for RTK, from version 1 to version 2, was motivated by a complete refactoring of RTK's code to make it compatible with ITK's modular architecture. The architecture of each module is defined by the ITKModuleTemplate repository (https://github.com/InsightSoftwareConsortium/ITKModuleTemplate). ITK is split into several modules with a dependency graph. Core modules are distributed in the `Modules` directory of the main ITK source repository, https://github.com/InsightSoftwareConsortium/ITK, *external* modules can be manually added to an `Modules/External` directory of ITK source directory and *remote* modules are external modules which are described in the ITK source repository and which are automatically downloaded when activated. The description of remote modules is provided in a `CMake` text file in the `Modules/Remote` directory of ITK's main source repository. RTK's module file was added in February 2019 and it is since one of ITK's remote modules. From the user perspective, RTK compilation has to be activated during ITK compilation by turning on the CMake option `Module_RTK`.

2.2 CudaCommon module

More recently, in June 2022, the modular architecture of ITK has also enabled the extraction from RTK of the code to manage ITK images on the graphics processing unit (GPU) using the Compute Unified Device Architecture (CUDA, NVidia). The CudaCommon module, available at https://github.com/RTKConsortium/ITKCudaCommon, can now be used by all users without RTK even if it is still managed by the RTK consortium. The main feature of this module is to offer an `itk::CudaImage` class which inherits from `itk::Image`. An `itk::CudaImage` object has a GPU buffer, in addition to the inherited central processing unit (CPU) one, to or from which the image data is transferred depending on which processing unit is used for the calculation. The data transfer from one unit to the other is automatically managed by triggering a data synchronization when calling a class member which may modify the data buffer.

2.3 Continuous integration

RTK relies on *GitHub actions* for testing every change to the repository before integrating it to the `master` branch. To that end, ITK provides reusable GitHub Actions workflows in a dedicated repository https://github.com/InsightSoftwareConsortium/ITKRemoteModuleBuildTestPackageAction/. Compilation and testing are executed by Linux, Windows and MacOS runners provided by GitHub's infrastructure. For CUDA, the RTK consortium currently manages self-hosted Linux and Windows runners but RTK will switch to CUDA-enabled Github runners as soon as they are made available.

2.4 Python wrapping

ITK offers the infrastructure to wrap selected parts of remote modules for their use in Python software. Since ITK heavily relies on templates to be able to, e.g., manage images of any dimension using pixels of any type (int, float, etc.), the wrapping code specifies which image types are compiled for Python. A large part of the CudaCommon and RTK modules has been wrapped. The user can either compile the wrapping or download pre-compiled binaries distributed via the Python Package Index (PyPI, https://pypi.org/). Both CPU only and CPU+GPU packages of RTK are distributed. The distributed CUDA packages are wrapped for a specific CUDA version number which is appended to the package name. Currently, the distributed packages are `itk-rtk`, `itk-rtk-cuda116` and `itk-rtk-cuda121` and the latter depends on `itk-cudacommon-cuda116` and `itk-cudacommon-cuda121`. New versions of the packages are distributed after every release of ITK base packages to ensure compatibility. The last ITK version is v5.4.0 and the corresponding CudaCommon and RTK releases are v1.1 and v2.6, respectively.

3 New features in RTK v2

3.1 Forbild phantoms

Tomographic reconstruction is often evaluated on numerical phantoms which are described by simple geometric shapes (ellipsoids, cones, parallelepipeds, etc.). The most famous one is the Shepp-Logan phantom, made of ellipsoids only, which was already available in RTK v1. RTK has since implemented the use of a larger set of geometrical shapes described by the Forbild file format. A separate GitHub repository (https://github.com/RTKConsortium/Forbild) pre-process the original Forbild files using gcc's preprocessor (to apply `#include` and `#define` directives) and Python (to calculate mathematical expressions), thus producing a single file understandable by RTK. These files (currently `Abdomen`, `Thorax` and `Head`) can be input to RTK to compute projection and reference CT data, as used for example in [3] to evaluate several reconstruction algorithms. Figure 1 illustrates the Forbild head and thorax phantoms.

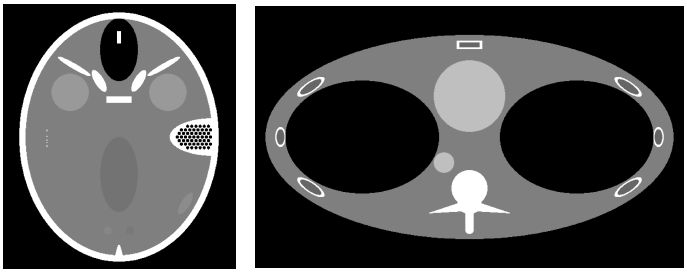


Figure 1: Central axial slice of the head (left) and thorax (right) Forbild phantom generated with RTK. Gray windows: $[0, 100]$ HU (left) and $[-100, 100]$ HU (right).

3.2 Cylindrical detector

RTK was initially developed for x-ray cone-beam CT with flat panel detectors which are commonly used in image-guided radiation therapy. Diagnostic CT scanners, used for acquiring treatment planning CT images, use cylindrical detectors with an axial axis (Figure 2). RTK is now capable of simulating and reconstructing projection data of cylindrical detectors, as illustrated in, e.g., [4].

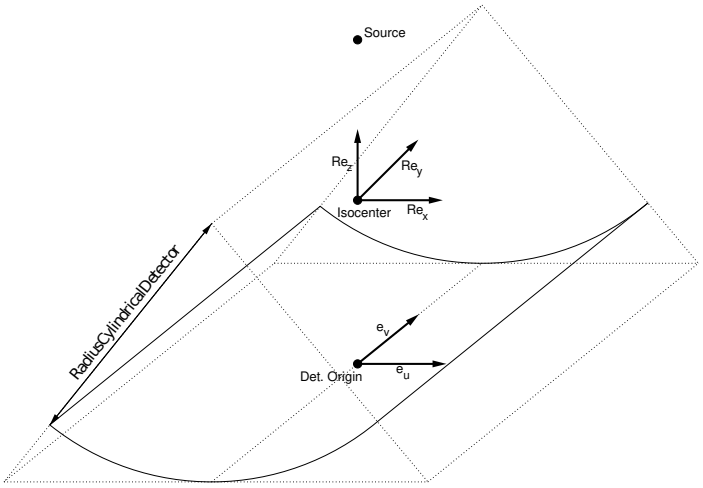


Figure 2: Schematic illustration of RTK's geometry with a cylindrical detector (from RTK's documentation, http://www.openrtk.org/Doxygen/DocGeo3D.html).

3.3 Iterative reconstruction

Several iterative reconstruction algorithms have been developed in RTK since v1. Those include conventional least square minimization with quadratic Tikhonov regularization using conjugate gradient minimization as well as less smooth regularizers, e.g., total variation, L1-norm of the image or L1-norm of the wavelet decomposition of the image. Several of these algorithms are also used by four-dimensional (4D) CT reconstruction algorithms to improve image quality, as illustrated in Figure 4.

3.4 Single photon emission CT (SPECT)

RTK has also been extended to handle projection data acquired with a SPECT scanner. New projectors (and adjoint backprojectors) have been implemented to model the patient attenuation (given a known attenuation map) and the resolution of the collimator [5]. For example, these projectors have been used to reconstruct respiration-correlated 4D SPECT images using ordered subset expectation maximization (OSEM) in [6]. Figure 3 shows sinograms of a sphere in the left lung of the Forbild thorax phantom without and with attenuation, where the attenuation map has been scaled such that the attenuation of water is $0.2\ \mathrm{cm}^{-1}$.



Figure 3: Parallel projections of a 1 cm sphere located at the center of the left lung of the Forbild thorax (Figure 1) without (left) and with (right) attenuation.

3.5 Four-dimensional CBCT

Four-dimensional (4D) CBCT has been a major axis of development of RTK since its debut. Early versions included filtered backprojection respiration-correlated 4D CBCT using the breathing signal extracted from the projections [7]. Several reconstruction algorithms have been developed since RTK v1 for reconstructing higher quality 4D CBCT images using iterative reconstruction techniques. The most advanced algorithms are the RecOnstructiOn using Spatial and TEmporal Regularization (ROOSTER) [8] and its motion-aware version [9]. The two algorithms alternate between conjugate gradient minimization and several constraints or regularizations: positivity, time-averaging outside a motion mask, 3D spatial total variation (TV) denoising and one-dimentional (1D) temporal TV denoising. The motion-aware version used a priori knowledge of breathing motion to improve the 1D TV denoising. Figure 4 shows coronal slices of several 4D CBCT images reconstructed with RTK.

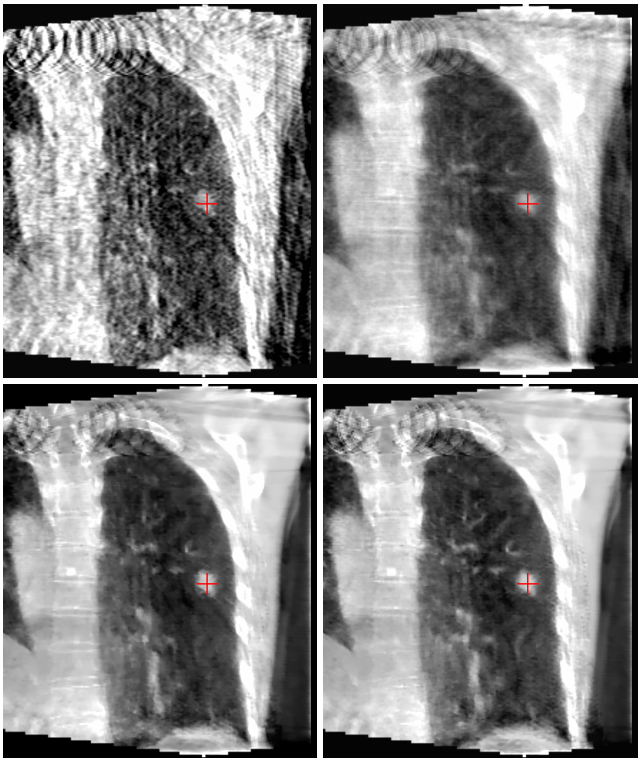


Figure 4: Coronal slice of the end-exhale CBCT frame of 4D CBCT images reconstructed with RTK from the same dataset, using filtered-backprojection (top left), unregularized conjugate gradient optimization (top right), ROOSTER [8] (bottom left) and MA-ROOSTER [9] (bottom right). Figure adapted from [9].

3.6 Spectral CT

Spectral CT is becoming increasingly popular in diagnostic CT. Dual-energy CT scanners are commercialized by all major vendors and scanners with photon counting detectors are also commercialized by some vendors. Spectral CT images can be processed in many ways. The first solution is to reconstruct each energy bin CT image and do image-based processing. However, the bin CT images will suffer from beam-hardening. An alternative is to process the acquisition in the projection domain if projection images are acquired with the same geometry, e.g. with a dual-layer detector or a photon counting detector. For example, material decomposition [10] can result in decomposed sinograms (Figure 5) and each sinogram can be reconstructed separately in material CT images using filtered backprojection or iterative algorithms.

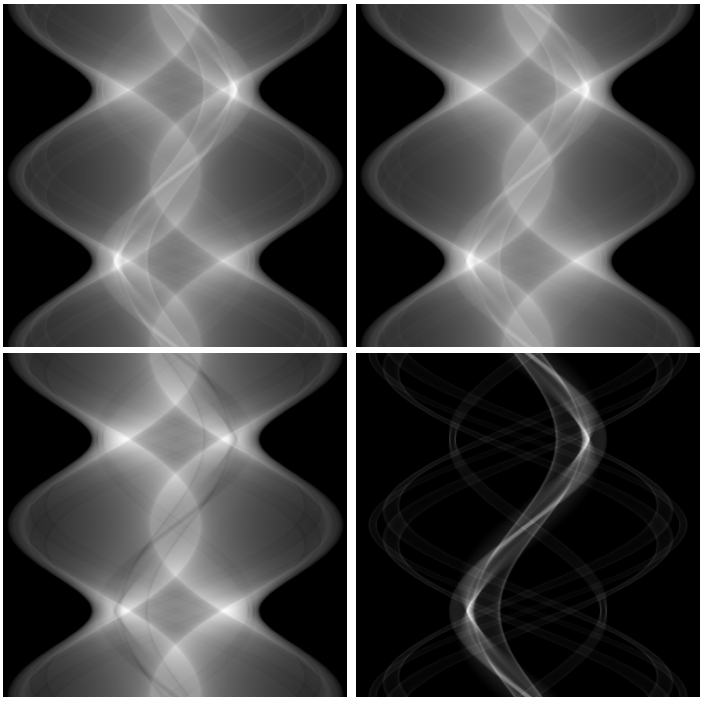


Figure 5: Log-transformed simulated X-ray projections of the Forbild thorax phantom (same slice as in Figure 1) at 80 kV (top left) and 120 kV (top right) and the corresponding water (bottom left) and bone (bottom right) sinograms of the material decomposition.

Alternatively, projection images can be decomposed and reconstructed jointly, in one-step. Such processing has a high computational cost but can significantly improve image quality, particularly if the number of energy bins of the scanner is larger than the number of sought material CT images. Two-step and one-step reconstruction algorithms have been implemented in RTK and evaluated on patient images obtained on a prototype photon counting scanner [11], as illustrated in Figure 6. Image quality improvement is a trade-off between spatial resolution and image noise can be improved depending on the selected regularization parameters.

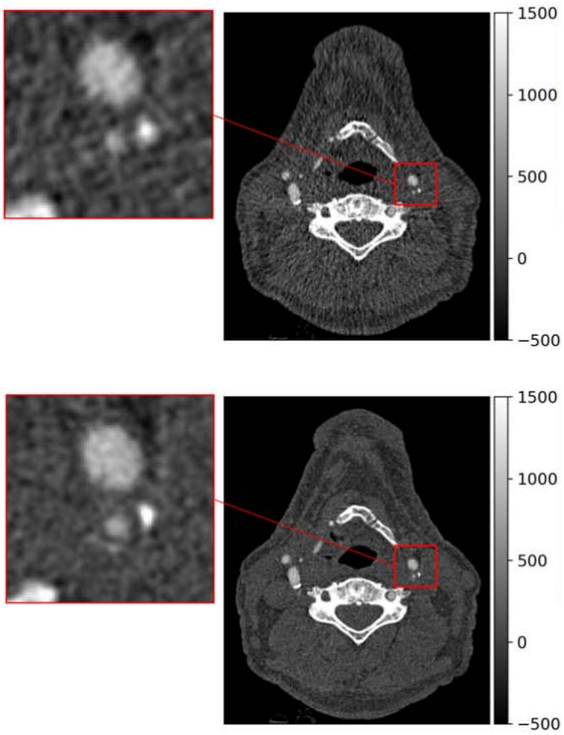


Figure 6: Virtual monoenergetic CT image at 40 keV obtained from RTK's two-step projection-based (top) and one-step (bottom) material decomposition of a prototype photon-counting CT angiography patient dataset. Figure adapted from [11].

4 Conclusion

Since its first public release in 2012 and its presentation at ICCR 2013, RTK has become a reference software for tomographic reconstruction of CT images. Thanks to funding from research projects and industrial partners, RTK keeps being improved and will remain an open-source research and development tool used and developed by an ever-growing international community.

References

[1] S. Rit, M. Vila Oliva, S. Brousmiche, R. Labarbe, D. Sarrut, and G. Sharp. "The Reconstruction Toolkit (RTK), an open-source cone-beam CT reconstruction toolkit based on the Insight Toolkit (ITK)". *J. Phys.: Conf. Ser.* 489 (Mar. 2014), p. 012079. DOI: 10.1088/1742-6596/489/1/012079.

[2] M. McCormick, X. Liu, L. Ibanez, J. Jomier, and C. Marion. "ITK: enabling reproducible research and open science". *Frontiers in Neuroinformatics* 8 (2014). DOI: 10.3389/fninf.2014.00013.

[3] M. Charles, R. Clackdoyle, and S. Rit. "Image reconstruction from truncated fan-beam projections along a circular trajectory using the virtual fan-beam method". *Phys Med Biol* 68.24 (2023), p. 245004. DOI: 10.1088/1361-6560/ad00fd.

[4] M. Mouchet, J. Létang, J. Lesaint, and S. Rit. "Cone-beam pairwise data consistency conditions in helical CT". 42.10 (2023), pp. 2853–2864. DOI: 10.1109/TMI.2023.3265812.

[5] G. Zeng, C. Bai, and G. Gullberg. "A projector/backprojector with slice-to-slice blurring for efficient three-dimensional scatter modeling". 18 (8 Aug. 1999), pp. 722–732. DOI: 10.1109/42.796285.

[6] A. Robert, S. Rit, T. Baudier, J. Jomier, and D. Sarrut. "Data-driven respiration-gated SPECT for liver radioembolization". *IEEE Trans Radiat Plasma Med Sci* 6.7 (Sept. 2022), pp. 778–787. DOI: 10.1109/TRPMS.2021.3137990.

[7] J.-J. Sonke, L. Zijp, P. Remeijer, and M. van Herk. "Respiratory correlated cone beam CT". *Med Phys* 32.4 (Apr. 2005), pp. 1176–1186. DOI: 10.1118/1.1869074.

[8] C. Mory, V. Auvray, B. Zhang, M. Grass, D. Schäfer, S. Chen, J. Carroll, S. Rit, F. Peyrin, P. Douek, and L. Boussel. "Cardiac C-arm computed tomography using a 3D + time ROI reconstruction method with spatial and temporal regularization". *Med Phys* 41.2 (Feb. 2014), p. 021903. DOI: 10.1118/1.4860215.

[9] C. Mory, G. Janssens, and S. Rit. "Motion-aware temporal regularization for improved 4D cone-beam computed tomography". *Phys Med Biol* 61.18 (Sept. 2016), pp. 6856–6877. DOI: 10.1088/0031-9155/61/18/6856.

[10] R. Alvarez and A. Macovski. "Energy-selective reconstructions in X-ray computerized tomography". *Phys Med Biol* 21.5 (Sept. 1976), pp. 733–744. DOI: 10.1088/0031-9155/21/5/002.

[11] P.-A. Rodesch, S. Si-Mohamed, J. Lesaint, P. Douek, and S. Rit. "Image quality improvement of a one-step spectral CT reconstruction on a prototype photon-counting scanner". *Phys Med Biol* 69.1 (2024), p. 015005. DOI: 10.1088/1361-6560/ad11a3.

Validation of an Open Source Contouring Tool

Haley Clark[1,2]

[1]Department of Radiation Oncology, BC Cancer, Surrey, Canada
[2]Department of Physics and Astronomy, University of British Columbia, Vancouver, Canada

Abstract While radiotherapy increases in sophistication, and software complexity continues to grow, there remains a lack of validated open source software for routine radiotherapy tasks. This work presents a validation of the geometric accuracy, positional precision, and interoperability of the open source DICOMautomaton manual contouring subsystem. Using 2D and 3D contour arrangements, we demonstrate that DICOMautomaton is sufficiently accurate for radiotherapy purposes.

1 Introduction

The complexity of treatment planning systems (TPSs) continues to increase, with commercial proprietary TPSs now commonly supporting specialized use-cases including adaptive radiotherapy, multi-criteria optimization, and automated segmentation [1]. Despite escalating sophistication, manual contouring or manual contour refinement nevertheless remains a crucial step in the radiotherapy process, even for the state-of-the-art [2].

In contrast to commercial offerings, non-commercial open source TPSs as a whole tend to be more focused in scope and, for various reasons, do not report validation of the implementation. This is especially true for 'infrastructure' software, where scientific novelty is deprioritized and opportunities to report validation are limited. Compared with autosegmentation tooling, which has exploded in popularity in recent years [3], the number of robust, validated, and open source manual contouring tools are limited.

The aim of this work is to validate the manual contouring subsystem of the DICOMautomaton [4, 5] open source project. While the set of features provided by the contouring subsystem is modest, it is capable of providing high-quality contours; can be composed with other subsystems (e.g., image registration, surface mesh extraction, 3D constructive geometry, and dosimetric analysis); supports reading and writing of contours in standards-compliant formats; and is extensible. The DICOMautomaton contouring subsystem is designed to interoperate with, and complement, other software, thus providing a validated open source tool to complete a crucial step in the radiotherapy process.

2 Materials and Methods

For region-of-interest (ROI) contouring to be useful for radiotherapy planning or dosimetric analysis, the provided tools must be versatile enough to contour a wide range of tissue (c.f. optic nerves vs liver), and the resulting contours need to be both dimensionally accurate and precise.

DICOMautomaton contouring supports any planar image modality (e.g., CT, MR, PET, US). Images are backed by a volumetric image mask from which contours are extracted. The underlying mask resolution (and thus minimum supported contour feature size) is user configurable. Increasing mask resolution increases the time need to extract contours, so the user is encouraged to adjust resolution to accommodate the nature of the specific target(s). For this work, synthetic CT images with mask images comprising 512x512 pixels are used. All masks used isotropic voxels with width 0.5 mm.

Contouring brushes with 2D and 3D extent are supported. Brushes can be manually dragged across an image to draw/erase contours, and both straight and right-angle sweeps are supported. Brushes support 'dynamics' which cause a contour to grow or shrink while lingering, mimicking the dynamics of paint spray. Brush sizes can range in size/extent continuously. The ability to use the available contouring tools to produce geometrically accurate contours is measured. To simulate realistic conditions when manually contouring, brush dynamics are enabled, and a visual guide of the brush is provided.

Circular/spherical targets with radii ranging from 5-50 mm are created and evaluated for concentricity against ideal circles/spheres using orthogonal projection fitting. All fits were allowed 200 iterations to converge and were considered fully converged when a tolerance of 0.001 mm was reached. Radii and maximum radial deviations are assessed against the idealized target radii. Brush sweeps are performed for circular and spherical brushes with 10 mm radius along the cardinal directions and diagonally; radii are similarly assessed. Sweeps can be used to provide interpolation in 3D, but dedicated contour interpolation is also provided. The accuracy is confirmed by interpolating between two 2D circles on different image planes to produce a cylinder; the interpolated portion is compared to the manually contoured portions. This is done once with 30 mm radius and once with 10 mm radius.

2D boundary contours are extracted from masks using a *marching squares* algorithm [6]. By inverting this algorithm, resulting in a new algorithm which we refer to as '*receding squares*,' it is possible to convert between mask and contour representations bidirectionally. Relaxing the requirement for the conversion to be exact, it is possible to convert from contours to a mask with a single pass using an error-correcting algorithm as follows: (1) for each pixel ($v_{i,j}$) in an image, starting in the top-left corner, determine the position and intensity of the adjacent pixels to the left (v_{i-}

$_{i,j}$), top ($v_{i,j-1}$), and top-left diagonal ($v_{i-1,j-1}$); (2) for each contour, find the locations where contours intersect the line segments connecting $v_{i,j}$ to the left, top, and diagonal pixels; (3) assuming linear interpolation will be used for marching squares, and that the left, top, and diagonal pixel intensities are already fixed, determine three separate estimates of the $v_{i,j}$ pixel intensity which will exactly reproduce the contour intersections, (4) finally, take the median of all estimates to yield a final estimate for the $v_{i,j}$ pixel intensity. Note that various boundary conditions can be applied when pixels have no top, left, or top-left diagonal neighbours; contour orientation and overall can be selectively honoured by counting crossings and transitions; and the estimate can also be clamped to the mask intensity range (e.g., [0:1]).

Accuracy of the receding squares algorithm is characterized by manually generating a contour mask (a circle with 50 mm radius), performing marching squares to extract the contour, performing receding squares on the contour to recover the mask, extracting a contour from the recovered mask, and then comparing the two contours geometrically.

Positive and negative margins (i.e., expansions and contractions) are supported via extensible morphological operations [7]. Accuracy is confirmed by applying -0.5, 0.5, 1.5, 3, and 5 mm margins to a spherical region-of-interest (ROI) with 30 mm radius and comparing the radial change vs expected.

The primary export format is DICOM with RTSTRUCT modality. Because individual contour sizes can easily exceed format length restrictions, e.g., >65534 bytes for explicit value representation (VR) transfer syntax of the contour data 3006,0050 element, many clinical systems now default to an *implicit* VR transfer syntax. This imposes additional consistency constraints across implementations, which need to share a common dictionary of elements. Furthermore, consistency and alignment of the coordinates systems used by images and contours is critical for radiotherapy purposes. Therefore, interoperability with other implementations needs to be explicitly confirmed; we do so here by importing into the Varian Eclipse TPS (v18.0) and MIM Maestro (v7.2.10), confirming the DICOM file interpretation is consistent, and then confirming contour geometry. DCMTK [8] (v3.6.7) is used to provide additional verification that files are in accordance with the DICOM standard.

Finally, an 'undo-redo' buffer is provided which allows the user to quickly jump. The validity of this mechanism is confirmed by comparing checksums of the underlying masks.

3 Results

Contour geometrical accuracy measurements were made for 2D and 3D circles/spheres with nominal radii 5, 10, 15, 30, and 50 mm. All fits converged, and fitted radii for 2D and 3D brushes were all within 0.002 mm of expected. ROIs with 3D extent correctly extended in- and out-of-plane to the expected image slice. Effective radii extracted from areas and perimeters computed directly from contours (i.e., using a vertex and edge representation) agreed with the nominal radii within 0.005 mm.

Cardinal axis-aligned brush sweeps were all accurate to within 0.005 mm, whereas diagonal sweeps were accurate within 0.02 mm. Axial interpolation between two 2D circles of either 10 mm or 30 mm radius separated by 5 empty image slices resulted in a perfect cylinder in both cases, with all interpolated 2D contours sharing the exact radii and location of the bounding contours.

For receding squares, the 'edit contours' feature was used to invoke receding squares, followed by marching squares to attempt to recover the original contour. Contour measurements are listed in table 1. Measured radii were equivalent to < 0.01 mm; areas differed by 0.003%, equivalent to roughly ¼ of the volume of a single voxel; and perimeters differed by 0.8%. The fitted centres differed by 0.01 mm.

Contour	Fitted Radius (mm)	Area (mm^2)	Perimeter (mm)
original	50.00	7853.81	314.18
recovered	50.00	7853.55	316.88

Table 1: Contour dimensions, original vs recovered via receding squares algorithm.

Results of applying margins to a 2D contour with 30 mm radius are shown in table 2. Though the input was a single contour, margins correctly expanded in 3D axially an appropriate number of slices. The maximum discrepancy in all cases was within ± 0.1 mm.

Margin (mm)	Expected Radius (mm)	Fitted Radius (mm)	Discrepancy (mm)
-0.50	29.50	29.54	0.04
0.50	30.50	30.46	-0.04
1.50	31.50	31.43	-0.07
3.00	33.00	32.93	-0.07
5.00	35.00	34.96	-0.04

Table 2: Geometric accuracy when applying margins.

Synthetic CT images and contours produced by DICOMautomaton were imported into Eclipse and Maestro TPSs. 2D and 3D circles/sphere contours with nominal radii 5, 10, 15, 30, and 50 mm were imported, and radii and centroids were measured and compared between software to confirm geometric accuracy and positional precision. All three systems agreed within 0.05 mm for radii (n.b. limited resolution due to the use of manual measurement tools) and

within 0.01 mm for centroids. Contours were then exported and imported back into DICOMautomaton, confirming bidirectional compatibility.

DCMTK's 'dcmpschk' and 'dcmdump' tools were used to confirm that encoding, syntax, structure, and content of the contour and image files were consistent with the DICOM standard. No errors or warnings were produced.

Finally, the 'undo-redo' buffer was confirmed to be functional and lossless.

4 Discussion

Validation is necessary to ensure software accuracy and functionality, but is often either not performed, or not reported for open source projects. The aim of this work was to validate the open source DICOMautomaton contouring subsystem for the purposes of radiotherapy.

All geometric accuracy and precision tests exhibited discrepancies that are considered to be clinically insignificant, including basic use of 2D and 3D brushes, sweeps, and axial interpolation between 2D contours.

The receding squares algorithm was found to recover the original contour's radius and area to within 0.003%. Due to the relaxed constraint for exact inversion, the recovered contour's perimeter was 0.8% larger. This is an expected result since the single-pass algorithm is not able to simultaneously satisfy three independent constraints (left, top, and top-left diagonal). The small amount of extra perimeter represents sub-millimetre variations or surface ripple and is acceptable for clinical radiotherapy. Note that increasing mask resolution will both reduce this variation, and also improve geometric accuracy for all discrepancies reported in this work.

Application of positive and negative margins resulted in accurate growth or reduction, with a maximum discrepancy of less than 0.1 mm from expected. Use of morphology-based margins is more straightforward compared to directly applying margins to contours (i.e., in vertex and edge representation) due to the topological complexities, including computing intersections and handling degenerate situations (e.g., abutting or thin contours). The morphology-based margins implementation provided by DICOMautomaton was found to provide accuracy sufficient for use for radiotherapy purposes.

Two third-party commercial TPSs were used to confirm DICOM interchange conformance. Geometrical accuracy and consistency of the coordinate systems was also confirmed. Open source tools from the DCMTK suite were used to provide verification against the DICOM standard.

The rationale for using predominantly circular or spherical contouring targets for geometric assessment is as follows: (1) the mask underlying contours is rectilinear and line segments that comprise circular contours span 360° angles, so circles/spheres exercise local-neighbour pixel intensity arrangements with non-zero gradient; (2) circles/spheres have a extreme relationship between area/volume and perimeter/surface area, simplifying detection of discrepancies; (3) circles/spheres have a well-defined position and extent, which can be used to simultaneously evaluate accuracy and precision, and the centroid corresponds to the geometric centre, simplifying the process of position extraction; and (4) the sweep of a circle/sphere in any direction has a consistent diameter, which simplifies assessment. While the tools provided by DICOMautomaton are capable of making intricate and sharp contours, they are comparatively harder to evaluate.

The receding squares algorithm demonstrated accurate inversion of the marching squares algorithm and reproduction of the underlying mask. Like marching squares, it can be accomplished in a single scan over image pixels and still yield accurate results. However, because the algorithm carries forward discrepancies from left-to-right and top-to-bottom (and thus compensating discrepancies arising from inconsistent intensity estimations), it is not 'embarrassingly parallel' like marching squares. In practice, multiple scans may be preferable to handle contour overlap (e.g., deciding whether an inner concentric circle is 'within' or 'outside' of a contour). Re-running receding squares in a reverse orientation (i.e., from bottom-right to top-left) can also improve accuracy. Both marching and receding squares algorithms trivially generalize to 3D voxels and surface meshes.

ROI volume was not compared in this work. Accurate estimation of the volume of a ROI requires extrapolation, and the spacing and resolution of adjacent mask image planes can impact the extrapolation. Furthermore, the volumes reported by the benchmark software (Varian and MIM) differ. While DICOMautomaton supports volumetric analysis, a detailed comparison is beyond the scope of this study.

5 Conclusion

The contouring subsystem provides rudimentary but effective tooling, generates geometrically sound contours, interoperates according to widely supported standards, and can thus fulfill a crucial step in the radiotherapy process. DICOMautomaton is free/libre and open source software. Contributions are welcome.

References

[1] Kamperis E, Kodona C, Hatziioannou K, Giannouzakos V. Complexity in radiation therapy: it's complicated. International Journal of Radiation Oncology, Biology, Physics. 2020 Jan 1;106(1):182-4. doi:10.1016/j.ijrobp.2019.09.003

[2] Harrison K, Pullen H, Welsh C, Oktay O, Alvarez-Valle J, Jena R. Machine learning for auto-segmentation in radiotherapy planning. Clinical Oncology. 2022 Feb 1;34(2):74-88. doi:10.1016/j.clon.2021.12.003

[3] Cardenas CE, Yang J, Anderson BM, Court LE, Brock KB. Advances in auto-segmentation. in Seminars in radiation oncology 2019 Jul 1 (Vol. 29, No. 3, pp. 185-197). WB Saunders. doi:10.1016/j.semradonc.2019.02.001

[4] Clark H, Thomas S, Moiseenko V, Lee R, Gill B, Duzenli C, Wu J. Automated segmentation and dose-volume analysis with DICOMautomaton. in Journal of Physics: Conference Series 2014 Mar 24 (Vol. 489, No. 1, p. 012009). IOP Publishing. doi:10.1088/1742-6596/489/1/012009

[5] Clark H, Doebeli C, Ringstrom C, Ho A, Rosenrauch J, Gayda J, et al. hdclark/DICOMautomaton: 20220203.1. Zenodo; 2022. doi:10.5281/zenodo.4088796

[6] Lorensen WE, Cline HE. Marching cubes: A high resolution 3D surface construction algorithm. InSeminal graphics: pioneering efforts that shaped the field 1998 Jul 1 (pp. 347-353).

[7] Chanda B. Morphological algorithms for image processing. IETE Technical Review. 2008 Jan 1;25(1):9-18.

[8] Eichelberg M, Riesmeier J, Wilkens T, Hewett AJ, Barth A, Jensch P. Ten years of medical imaging standardization and prototypical implementation: the DICOM standard and the OFFIS DICOM toolkit (DCMTK). InMedical Imaging 2004: PACS and Imaging Informatics 2004 Apr 19 (Vol. 5371, pp. 57-68). SPIE. doi:10.1117/12.534853

Harmonizing multicentric T1-weighted magnetic resonance imaging of children with ependymomas: can deep learning based image synthesis help?

Soumia Bengoufa[1], Fatima Tensaouti*[2,3], Pedro J Soto Vega*[1], Stéphanie Bolle[4], Alexandre Escande[5], Xavier muracciole[6], Claire Alapetite[7], Line Claude[8], Julie Leseur[9], Aymeri Huchet[10], Jérôme Doyen[11], Georges Noël[12], Stéphane Supiot[13], Bernier-Chastagner Valerie[14], Julien Welmant[15], Pierre Leblond[16], Dimitris Visvikis[1], Vincent Jaouen[1], Anne Laprie[2,3], Mathieu Hatt[1], and On behalf of GFRP (groupe français de radiothérapie pédiatrique) and of Ependymomics project consortium"[1]

[1]Laboratory of medical information processing (LaTIM), INSERM UMR 1101, Univ Brest, Brest, France
[2]Department of Radiation Oncology, Oncopole Claudius Regaud, Institut Universitaire du Cancer de Toulouse – Oncopole, France.
[3]Toulouse NeuroImaging Center, Université de Toulouse, Inserm, UPS, France
[4]Department of Radiation Oncology, Institut Gustave Roussy, Paris, France
[5]Department of Radiation Oncology, Centre Oscar Lambret, Lille, France
[6]Department of Radiation Oncology, CHU La Timone, Marseille, France
[7]Department of Radiation Oncology, Institut Curie, Paris, France
[8]Department of Radiation Oncology, Centre Léon Bérard, Lyon, France
[9]Department of Radiation Oncology, Centre Eugène Marquis, Rennes, France
[10]Department of Radiation Oncology, CHU Bordeaux, France
[11]Radiation oncology, Centre Antoine-Lacassagne, University of Côte d'Azur
[12]ICANS - Radiation Oncology Strasbourg, France
[13]Department of Radiation Oncology, Institut de cancérologie de l'ouest, Nantes, France
[14]Department of Radiation Oncology, Centre Alexis Vautrin, Vandoeuvre, Nancy, France
[15]Department of Radiation Oncology, Institut Cancer de Montpellier, France
[16]Department of Pediatric Onco-Hematology, IHOP, Lyon, France
*contributed equally

Abstract This work was carried out within the Ependymomics project, focusing on the analysis of multi-omics data of children and teenagers with epndymoma. In this paper, we present preliminary results of the harmonization of multicentric brain T1 MRIs from 14 sites in France. We aimed to mitigate image properties heterogeneity using an unsupervised deep learning GAN-based CUT model. Using T1-weighted MRI (n=217 subjects), we tested the model (i) by using different references, and (ii) with a single or multiple slices as input references. Qualitative evaluation of harmonized images, as well as the use of image quantitative metrics, show that the method efficiently reduced the inherent variability of images, with before and after variance values of $2,81e^{11}$ and $1,15e^{3}$ respectively. Our next step will be to evaluate the impact of these harmonization strategies on tumor-derived radiomics, and the resulting improvements in the performance of radiomics models predicting patient outcomes (progression free survival).

1 Introduction

Projects including multimodal medical images as a data source (amongst others histopathology, genomics, or clinical variables) to build models predictive of patient outcomes, typically face challenges associated with recruiting larger cohorts of patients in numerous clinical centers over a long period, to reach a reasonable number of patients for model training and evaluation. The main advantage of such a recruitment is that a larger cohort of patients is available to support stronger statistical significance of findings, and can allow for more efficient training and validation of models. The drawback is that collected images can display a high level of heterogeneity in terms of overall quality and properties (e.g., signal-to-noise ratio, contrast levels, axial versus sagittal, voxel size, textures...). This is due to the fact that the various centers in which patients were imaged relied on different scanners (vendor, model, generation) and with different acquisition and reconstruction protocols. Scanner and/or protocol variability can affect both the stability and the discriminating power of radiomic features and predictive models[1]. The Ependymomics project aims at building models predictive of the risk of relapse and outcomes in pediatric ependymoma patients,(the third most common brain tumor affecting children [2]), with a high rate of relapse of approximately 40 % after surgery and radiotherapy [3, 4], by exploiting imaging, dosimetry, clinical, genomics, and histopathology data. The imaging datasets collected in this context consist of MRI sequences including T1, T2, and FLAIR. The cohort exploited in the project contains 345 patients and was collected in 14 different centers in France. The goal of this paper is to present our preliminary evaluation of a method for harmonizing the images of the ependymomics cohort, focusing on the post-contrast T1-weighted images. The method uses a modified Generative Adversarial Network (GAN)[5] image synthesis, namely the Contrastive Unpaired Image to Image Translation (CUT) algorithm [6]. The evaluation of the performance of the harmonization will be reported here using a set of overall image metrics. Future work will consist of evaluating the harmonization using tumor radiomics features, and the potential impact on the predictive models, including all sequences.

2 Materials and Methods

2.1 Harmonization procedure

Based on the domain adaptation approach, our objective is to increase the similarity of characteristics between the images of the data set, through a style translation, while preserving their clinically relevant informative content. Thus, in the context of our multi-center study, we aim to evaluate the performance of the generative model for mitigating the effects of the acquisition shift. Based on two subsets of data, an input source domain X (raw images) and a target domain Y (reference images), the harmonization model used here learns to translate an image from one domain to the other. In other words, it aims to translate images from X to appear like an image from Y generating a "fake y" = $\bar{y}$, i.e. the harmonized images. In our application, we use this method to transform images so they appear as if they had been acquired in the same configuration (scanner model, acquisition protocol, reconstruction settings) as the chosen reference.

2.2 Dataset and pre-processing

The data were collected from the NETSPARE (Network to Structure and Share Pediatric data to Accelerate Research on Ependymoma) database, which stores all relevant data from 345 children and adolescents with ependymoma since 2000 in France. After the curation of the images, 217 patients from 14 imaging centers were included in the present analysis. For each patient, the post-gadolinium T1-weighted sequence was selected. These images were acquired using scanners from Siemens, Philips, and GE, using different acquisition protocols at each center with varying acquisition parameters (1.5 or 3 Tesla, slice thickness ranging from 1.5 mm to 4 mm, and matrix sizes from 68×65 to 512×512 pixels).
We applied some commonly performed pre-processing steps[7], such as conversion from DICOM to NIFTI format, cropping to a size of 512×512, normalization to [-1, 1] based on minimum and maximum intensity values, and matching to the percentile of the quantile range [1%, 99%].

2.3 CUT Model

The CUT model is based on a CycleGAN [8] used in the context of unpaired data. Unlike CycleGAN, the CUT model requires learning the mapping in one direction and avoids using inverse auxiliary generators and discriminators. This can simplify and shorten the training. To summarize, the training needs a set of images from each domain. Given a dataset of unpaired instances $X = \{x \in X\}$, $Y = \{y \in Y\}$. The objective is to learn a pixel-level adapter $G(x \in X)$ to a reference subset Y and generate a new set of data $\bar{Y}$. The generator function G is broken into two components, an encoder *Genc* followed by a decoder *Gdec*, which are applied sequentially to produce output image $\bar{y} = G(z) = Gdec(Genc(x))$. To achieve our goal of content conservation and style transfer, two loss functions are used: the content conservation of the source domain (X => $\bar{Y}$) is ensured through the PatchNCE loss ((L_{NCE}), and the style translation from the reference domain (Y => $\bar{Y}$) is ensured through the GAN loss ((L_{GAN}).

Strategy 1	Strategy 2	Strategy 3	Strategy 4
Reference 1		Reference 2	
3 slices	1 slice	3 slices	1 slice

Table 1: 4 strategies for harmonization performance assessment

2.4 Comparison of harmonization strategies

Four different strategies were compared to assess the performance of each method. The first main parameter concerns the choice of an image of the reference domain. A typical reference should be a good-quality image within / external of X or can be synthesized images, with desirable properties, i.e. higher contrast, greater sharpness, and lower noise or blur levels. Therefore, for our work, we defined certain criteria for selecting and testing several references within the dataset, to train the model, namely: recently acquired, isotropic, 3 Tesla, and with high contrast. The second parameter was to use either a single slice or several (in our case, 3 at different locations, middle, top, and bottom of the patient's brain) in the reference volume to feed the network. Therefore, the model was trained for each chosen reference using either 1 or 3 slices (**Table 1**). All experiments were implemented in PyTorch 1.13 with the Adam optimizer using the same training hyper-parameters, and they were trained and tested on an NVIDIA GeForce RTX 3090GPU (24G) with a batch size of 1. The training was performed for 400 epochs, and a learning rate of 0.0002 using a ResNet-9 inspired architecture.

2.5 Evaluation

First, a qualitative (visual) assessment was carried out to appreciate the magnitude of the harmonization and to ensure that the harmonized images were visually satisfying, with no obvious artifacts, noise, or distortions of the clinically-relevant contents. A statistical analysis (quantitative evaluation) was then carried out to assess the model's performance in reducing variability after harmonization, relying on 15 image metrics (including first-order statistical measures as well as second-order statistical and filter-based measures) extracted before and after harmonization, as it was proposed previously to quantify the batch effect in MRI and to detect artifacts [9]. t-distributed stochastic neighbor embedding (t-SNE) [10] was then used to project the 15 metrics into a 3D space and to visualize the difference of image properties before and after harmonization. The reduction of variance was measured to compare the four different strategies.

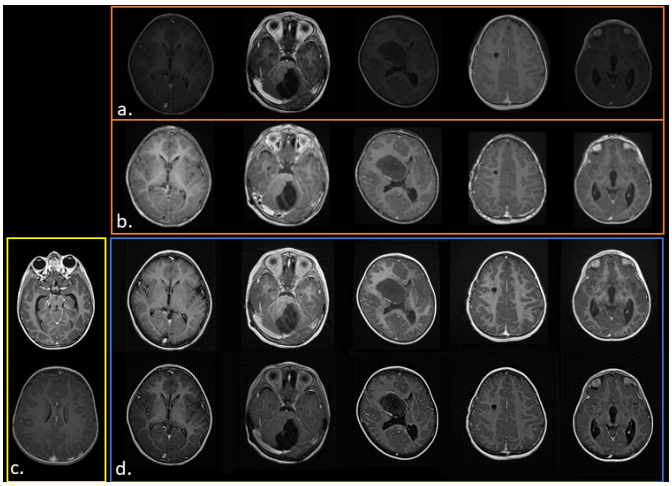


Figure 1: Qualitative analysis of the CUT model on T1-weighted source images of 5 different patients towards the two references (c.) from different centers. (a.) Raw images. (b.) The pre-processed non-harmonized images. (d.) The harmonized images.

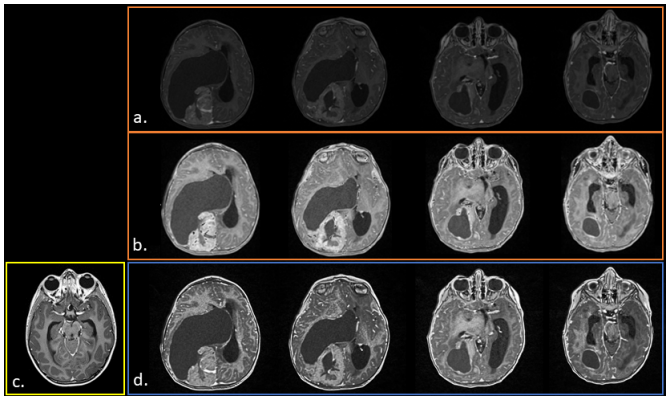


Figure 2: Qualitative analysis of CUT model on 4 slices of a single patient. (c.) The reference, (a.) the raw source image, (b.) pre-processed images, (d.) the harmonized images.

3 Results

Figure 1 gives an overview of the model results, showing five 2D axial slices of five different patients from different centers. The two first lines represent the raw and the pre-processed images, respectively, and the last two lines correspond to the generated (harmonized) images with each of the two chosen references. It is worth noting that the harmonized images are generated without artifacts or distortions, for both references. It can also be appreciated visually that all generated images have a more similar appearance to the chosen reference than before harmonization. According to clinician collaborators, the style transfer was successfully achieved without altering the clinically relevant content of the images (position tumor, organs...). The same observation is made for the results represented in **Figure 2**, which displays the harmonized images for a single patient with a large tumor over 4 slices of the volumes. The t-SNE results are shown in **Figure 3**, highlighting a clear separation of the data in three-dimensional space, for all the strategies using different references and one (middle of the head) or more slices of

Strategies	1	2	3	4
σ^2 before harmonization	$2.81e^{11}$			
σ^2 after harmonization	$2.38e^{8}$	$7.44e^{4}$	$2.23e^{4}$	$1.15e^{3}$

Table 2: Variance (σ^2) after/ before harmonization by strategy.

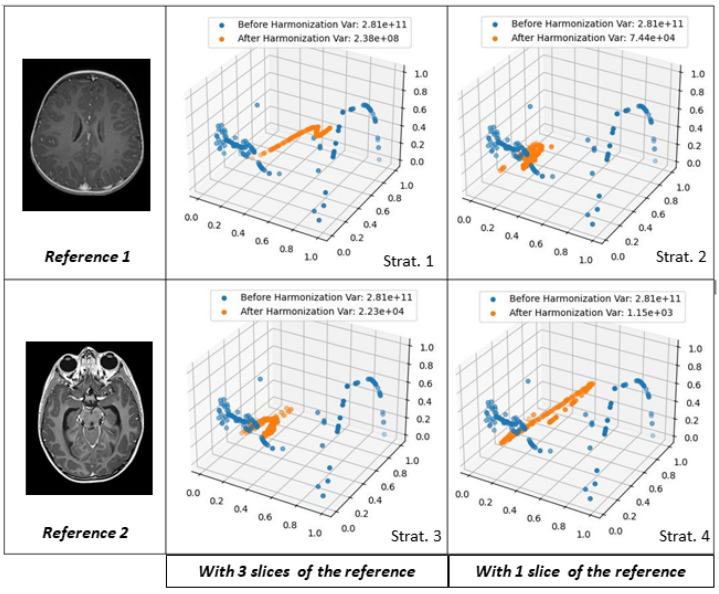


Figure 3: T-SNE analysis based on the 15 metrics extracted from images, before and after harmonization.

each. It can be seen that the values of the metrics extracted before harmonization (blue points, each point is an image) have a much greater dispersion. After harmonization (orange points), the observations are much more clustered, displaying reduced dispersion. In other words, there is less variation and more similarity in the data after harmonization. For each strategy, the variance (σ^2) before and after harmonization was calculated to assess the model performance (**Table 2**). Strategies 2 ($\sigma^2 = 7.44e^4$) and 4 ($\sigma^2 = 1.15e^3$) led to a high decrease in variance, suggesting that the use of one slice from a reference as an input for training the model is more efficient than using 3. Furthermore, the use of reference 2 led to stronger variance reduction (strategies 3 ($\sigma^2 = 2.23e^4$) and 4 ($\sigma^2 = 1.15e^3$)) than reference 1.

4 Discussion

In this paper, we presented preliminary results of our contribution to the Ependymomics project where the major challenge is the large heterogeneity of data gathered from different sites, subjecting it to the problem of acquisition shift. Research shows that this heterogeneity adversely affects radiomic studies and predictive models[11] supporting the hypothesis that an image harmonization step could help in training predictive models if the information is contained in the images but is "hidden" due to the heterogeneity of image properties. The purpose of our work was to evaluate the ability of a harmonization model based on image synthesis to reduce variability

within a heterogeneous MRI dataset. The CUT model used here, was implemented for two objectives: style translation and content preservation, ensured thanks to a GAN and a PatchNCE loss functions, respectively. To assess the performance of the CUT model in terms of reducing variability within a dataset, a quantitative analysis was carried out using image quality measurements extracted before and after the harmonization. This was done for 4 strategies (i.e., 2 different references chosen, and for each reference, using 1 or 3 slices as input) for model training. In all experiments, the variance values decreased by several orders of magnitude (from $2.81e^{11}$ before to between $2.38e^{8}$ and $1.15e^{3}$ after harmonization), showing that the CUT model can reduce the variability of image characteristics. This applies even to very small reference domain training sets consisting of a single 2D reference image slice (Figure 3: strategies 2 and 4), where the model achieved its best performance. The choice of reference is also an important factor in achieving optimum model performance. Given the variance values, we found that the model performed better with reference 2 ($\sigma^2 = 1.15e^{3}$) than with reference 1 ($\sigma^2 = 7.44e^{4}$). This may be explained by the fact that the model performs better with a reference whose image characteristics are closest to those of the dataset.

A very recent study [12] has developed a similar approach and added a gradient difference loss (GDL) to reinforce the structure preservation of source images while allowing flexible contrast translation from the reference. To improve our results, we intend to investigate the addition of this loss. Furthermore, one of the limitations of our work is that the evaluation of the CUT model has been done just at the level of the image, in future work, we will include an evaluation at the level of tumor radiomic characteristics, as well as a study of the potential impact of harmonization on the predictive modeling stage, including other MRI sequences.

5 Conclusion

This paper presents preliminary results of image harmonization of multi-center imaging data from the ependymomic cohort, focusing on T1-weighted sequences. This is an image-to-image translation approach transferring the style of a reference image from a target domain to heterogeneous images from a source domain. We evaluated the model performance in reducing variability within a dataset through a t-SNE analysis of image quality metrics extracted before and after image harmonization. This analysis showed that variability was significantly reduced after harmonization, with better efficiency when using a single slice of mid-brain volume as a reference than when using multiple slices. This approach showed promising results and could be used as image pre-processing for radiomic studies. In future work, we aim to add further losses, which could increase the structural correspondence between generated and source images, as well as apply the model to different MR image sequences in the database. We will also evaluate the impact of the CUT harmonization on radiomics predictive modeling of outcomes.

6 Acknowledgments

This work was funded by the French National Cancer Institute (INCa) "PRT-K2020, N° 2021-019, the Unicancer group "Prix Unicancer de l'Innovation: Intelligence Artificielle et Data Sciences 2020", the CHIST-ERA grant [CHIST-ERA-19-XAI-007] with project acronym INFORM, by General Secretariat for Research and Innovation (GSRI) of Greece, National Science Centre (NCN) of Poland [2020/02/Y/ST6/00071] and Agence Nationale de la Recherche (ANR) of France and carried out within the Ependymomics project (NCT05151718).

References

[1] S. A. Mali, A. Ibrahim, H. C. Woodruff, et al. "Making radiomics more reproducible across scanner and imaging protocol variations: a review of harmonization methods". *Journal of personalized medicine* 11.9 (2021), p. 842. DOI: 10.21037/qims-19-885.

[2] S. T. Jünger, B. Timmermann, and T. Pietsch. "Pediatric ependymoma: an overview of a complex disease". *Child's Nervous System* 37.8 (2021), pp. 2451–2463.

[3] A. Ducassou, L. Padovani, L. Chaltiel, et al. "Pediatric Localized Intracranial Ependymomas: A Multicenter Analysis of the Société Française de lutte contre les Cancers de l'Enfant (SFCE) from 2000 to 2013". *International Journal of Radiation Oncology* Biology* Physics* 102.1 (2018), pp. 166–173.

[4] F. Tensaouti, A. Ducassou, L. Chaltiel, et al. "Patterns of failure after radiotherapy for pediatric patients with intracranial ependymoma". *Radiotherapy and Oncology* 122.3 (2017), pp. 362–367.

[5] I. Goodfellow, J. Pouget-Abadie, M. Mirza, et al. "Generative adversarial networks". *Communications of the ACM* 63.11 (2020).

[6] T. Park, A. A. Efros, R. Zhang, et al. "Contrastive learning for unpaired image-to-image translation". *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*. Springer. 2020, pp. 319–345.

[7] Y. Mallya, J. Vijayananda, M. Vidya, et al. "Automatic delineation of anterior and posterior cruciate ligaments by combining deep learning and deformable atlas based segmentation". *Medical Imaging 2019: Biomedical Applications in Molecular, Structural, and Functional Imaging*. Vol. 10953. SPIE. 2019, pp. 471–477.

[8] J.-Y. Zhu, T. Park, P. Isola, et al. "Unpaired image-to-image translation using cycle-consistent adversarial networks". *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2223–2232. DOI: 10.1109/ICCV.2017.244.

[9] A. R. Sadri, A. Janowczyk, R. Zhou, et al. "MRQy—An open-source tool for quality control of MR imaging data". *Medical physics* 47.12 (2020), pp. 6029–6038. DOI: 10.1002/mp.14593.

[10] G. E. Hinton and S. Roweis. "Stochastic neighbor embedding". *Advances in neural information processing systems* 15 (2002).

[11] A. Cameron, F. Khalvati, M. A. Haider, et al. "MAPS: a quantitative radiomics approach for prostate cancer detection". *IEEE Transactions on Biomedical Engineering* 63.6 (2015).

[12] C. Hognon, P.-H. Conze, V. Bourbonne, et al. "Contrastive image adaptation for acquisition shift reduction in medical imaging". *Artificial Intelligence in Medicine* 148 (2024), p. 102747.

Performance evaluation of a Pix2Pix synthetic Xray image generator for preclinical applications

Vasilis Eleftheriadis[1], Paneta Valentina[1], Eleftherios Fysikopoulos[1] and Panagiotis Papadimitroulas[1]

[1]BIOEMISSION TECHNOLOGY SOLUTIONS, BIOEMTECH, Athens, Greece

Abstract In this work, we evaluate a pix2pix conditional generative adversarial network as a potential solution for generating adequately accurate synthesized morphological X-ray images by translating standard photographic images of mice. Continuing our previous work, we expanded our dataset to 940 paired photographic/X-ray mice images and used data augmentation and transfer learning techniques to improve the performance and generalization of our model. We explore how the expansion of the training dataset affects the models' performance and use five image quality metrics to quantitatively evaluate the models' performance in the present dataset and compare the results to visual inspection by experts to assess the metrics' ability to evaluate image-to-image translation tasks. In conclusion, a combination of metrics is essential for generated image quality evaluation, and we propose an ensemble of the used metrics as an essential methodological step to evaluate the performance of image-to-image deep learning models.

1 Introduction

Scintigraphy is a high throughput alternative to 3D high-end molecular imaging systems and biodistribution ex vivo studies in monitoring the interaction of radioisotope- or fluorescence-labeled pharmaceuticals on animal subjects. It can effectively be used to track a new tracer and study the accumulation from zero point in time post-injection.

Image-to-image translation is a generative deep learning (DL) technique that utilizes generative adversarial networks (GANs), conditional generative adversarial networks (cGANs), and convolutional neural networks (CNNs) to learn complex mapping functions between input and output images to translate a source image into a target image while preserving certain visual properties of the original.

In our previous studies [1, 2] we presented a potential solution for generating adequately accurate synthetic morphological X-ray images by translating standard photographic images of mice using a well-known cGAN for image-to-image translation (pix2pix) [3]. Such an approach would benefit both imaging and targeted radionuclide therapy studies. The results showed that the network predicts an X-ray image with sufficient accuracy and that the calculated metrics are comparable with those presented in literature. In this study, we expanded our dataset, and we used data augmentation and transfer learning techniques to improve the generalization of our method. In addition, we performed a thorough evaluation of the performance comparing different metrics.

2 Materials and Methods

A dataset of 940 input/ground truth image pairs consisting of a photographic image and the corresponding X-ray scan of the same anesthetized mouse has been acquired. The photographic images were captured using commercial imaging devices by BIOEMTECH (eyes-series – www.bioemtech.com).

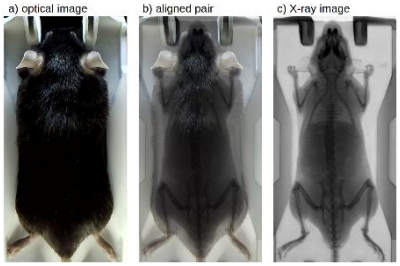


Figure 1: a) Indicative optical image acquired by BIOEMTECH eyes-series imaging devices, b) Corresponding X-ray image, c) Aligned pair.

These scanners, featuring a standard photographic sensor, provide an optical image of the animal (Fig. 1a) [4]. Both systems are adequate for mice imaging providing a useful field of view (FoV) of 50×100 mm^2.

The study involved 86 white and 88 black Swiss albino mice. We acquired 5 input/ground truth images of each animal in different poses upon the hosting bed, leading to a total of 940 input/ground truth images. The mice were divided into two groups to form the training and testing datasets. A group of 76 white and 78 black mice was used to collect 823 paired images for training, while a group of 10 white and 10 black mice was used to collect 117 paired images for testing. Except mouse color, the method was evaluated against different animal hosting beds (plastic bed with white color; plastic bed with black color), which leads to different backgrounds in the input image. The image pairs were properly aligned before training, having a resolution of 512×1024 pixels, corresponding to the 50×100 mm^2 FoV (Fig. 1b). Additionally, to ensure the validity of the training and testing sets, the mice used for training were separate from the individual mice used in the test set.

In this study, we utilized the PyTorch implementation of the pix2pix algorithm [3] and used the cross-entropy loss function. We trained all models for 200 epochs. For our transfer learning experiments, we first trained models on a subset of the dataset and then finetuned those on data from specific scanners using the weights of the pretrained models. For data augmentation we used the open-source Python library Albumentations [5]. By using different geometric and photometric transformations we managed to expand our training dataset up to 5175 paired images. We evaluated the performance of the pix2pix models in the present dataset using five commonly used image quality metrics: (a) mean squared error (MSE), (b) normalized root mean squared error (NRMSE), (c) peak signal-to-noise ratio (PSNR), (d) structural similarity index measure (SSIM) and (e) Fréchet inception distance (FID). Furthermore, we propose and evaluate our models' performance with a novel

weighted average ensemble performance evaluation metric (WAEM) which combines these five metrics.

$$WAEM = \frac{a}{MSE} + \frac{b}{NRMSE} + c \cdot PSNR + d \cdot SSIM + \frac{e}{FID}$$

3 Results

Model performance for increasing dataset size, starting from 50 to 5175 images, is presented in Fig. 2 through the 5 metrics used in this study. By visual inspection of the generated images, none of these metrics was adequate to individually be used to identify the best performing model. The weighted combination of the used metrics (WAEM), relying on a greater extent on FID, was found to be the metric which most consistently agrees with visual inspection and is also included in Fig. 2.

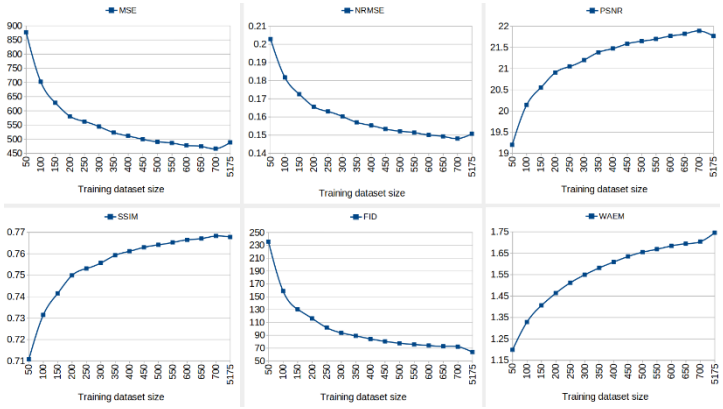


Figure 2: Graphs of the 6 metrics used in this study for the performance evaluation of the pix2pix models. The X-axis represents the size of the training dataset.

Fig. 3 shows indicative optical to X-ray translations of the same optical image in the test dataset for different training set sizes for visual comparison. The increase of the dataset using data augmentation (Fig. 2 & 3) was found to boost performance, while transfer learning techniques did not improve the generalization of our model, in our case study.

4 Discussion

Increasing the training dataset size with real input/ground truth pairs of images (up to 700 in the graph) the performance of the pix2pix models is improved according to all evaluation metrics. Nevertheless, when we further increase the training dataset size using data augmentation to 5175 images, four out of five metrics (MSE, NRMSE, PSNR and SSIM) show a drop in the model's performance. In contrast, FID metric gets even lower values implying that the final model with the augmented training dataset performs better. Fig. 3 depicts the visual resemblance of the generated images to the ground truth that drastically improves by increasing the training dataset size either with real images and/or using data augmentation. We identified FID as the most indicative metric for image-to-image translation. However, we observed that when 2 models have similar FID scores the models that translate the photographic images into synthetic x-rays that resemble the ground truth better are the ones that score better on the rest of the performance evaluation metrics (MSE, NRMSE, PSNR and SSIM), independently of which model had the best FID score. Our proposed weighted metric (WAEM) was found to clarify these cases and gave a straightforward measure of performance.

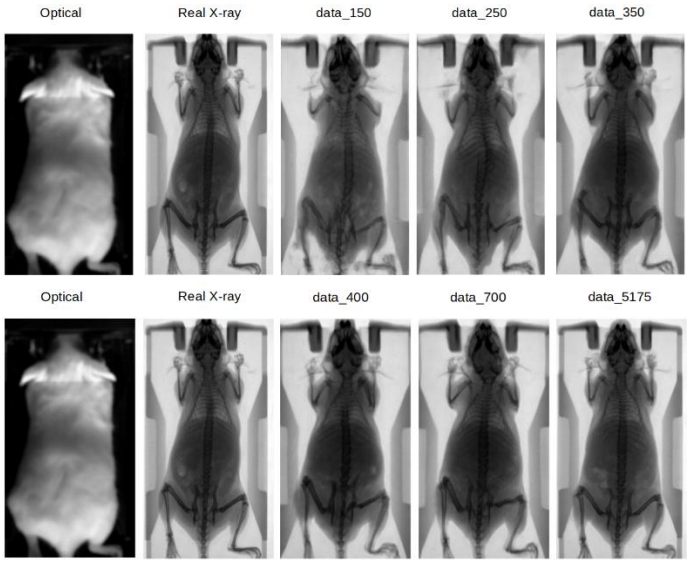


Figure 3: Indicative optical to X-ray translations of the same optical image in the test dataset for different training set sizes.

5 Conclusion

In the present work we evaluate the performance of pix2pix for image-to-image translation in our case study, in terms of dataset size, data augmentation, transfer learning techniques and the performance metrics themselves, that proved to be the most challenging issue. Our proposed weighted average ensemble performance evaluation metric (WAEM), which combines the five metrics used in this study (and are the most frequently used metrics in this field) could help researchers evaluate the performance of image-to-image DL models according to their case/dataset too.

References

[1] E. Fysikopoulos et al. "Optical to Planar X-ray Mouse Image Mapping in Preclinical Nuclear Medicine Using Conditional Adversarial Networks". J Imaging 7(12):262 (2021). DOI: 10.3390/jimaging7120262.

[2] E. Fysikopoulos et al. "Photograph to X-ray Image Translation for Anatomical Mouse Mapping in Preclinical Nuclear Molecular Imaging". In: Su, R., Zhang, YD., Liu, H. (eds) Proceedings of MICAD 2021. Lecture Notes in Electrical Engineering, vol 784. Springer, Singapore. DOI: 10.1007/978-981-16-3880-0_31.

[3] P. Isola et al. "Image-to-image translation with conditional adversarial networks". Proceedings of the IEEE CVPR Proceedings; Honolulu, HI, USA. 21–26 July 2017. DOI: 10.1109/CVPR.2017.632.

[4] M. Rouchota et al. "A prototype PET/SPET/X-rays scanner dedicated for whole body small animal studies". Hell. J. Nucl. Med. 2017;20:146–153. DOI: 10.3390/jimaging7120262.

[5] A. Buslaev et al. "Albumentations: Fast and Flexible Image Augmentations". Information 2020, 11, 125. DOI: 10.3390/info11020125.

Fast and automated dose verification procedure for reinforced online adaptive MRI-guided RT

Oleksii Semeniuk[1,4], A. Shessel[2], M. Velec[2], [3]A. Hosni[1,2], L. Dawson[1,2], J. Kim[1,2], A. Hope[1,2] and Teo Stanescu[1-3]

[1]Medical Physics, Princess Margaret Cancer Centre, UHN, Toronto, Canada
[2]Department of Radiation Oncology, University of Toronto, Toronto, Canada
[3]Mechanical and Industrial Engineering, University of Toronto, Toronto, Canada
[4]The Warren Alpert Medical School of Brown University, Department of Radiation Oncology, Rhode Island, USA

Abstract This work discusses a dose verification procedure tailored for online adaptive MRI-guided RT workflows specific to the Unity MR-Linac system. The methodology was implemented as a fully automated pipeline using Python-based scripting available in the RayStation treatment planning system. The pipeline input consisted of adapt-to-shape (ATS) plan, including target and organs-at-risk (OARs) contours, and dose maps, as well as 3D image data acquired for treatment verification (VER) step as part of the online workflow and data collected during beam-on delivery (BON). The automated analysis algorithm relies on deformable image registration (DIR) between ATS and VER/BON image series to propagate contours (no need for extra contours generation) and facilitate dosimetric analysis. Contours derived on VER/BON are representative for intra-fraction motion and can be used to quantify dosimetric changes. 100 treatment fractions were investigated for liver, pancreas, oligometastases and lung SBRT cancer patients. 620 contours and their associated DIRs were evaluated by expert reviewers. 98% of the DIR-generated contours were found to be clinically acceptable. The results showed a significant correlation between DICE values and intra-fraction motion. The 3D dose verification tool on both VER and BON datasets detected significant dosimetric changes which were not apparent during the online treatment sessions. This analysis holds potential as a valuable tool for assessing plan quality and can streamline the decision-making process during the verification stage of the online adaptive workflow.

1 Introduction

Online adaptive MRI-guided RT procedures are rapidly advancing, integrating both anatomical and functional imaging information. They extend capabilities such as gating using real-time multi-planar imaging, the ability to rapidly adjust treatment plans for sudden target baseline shifts, and synchronous contouring by multiple users. Opportunities for improvement exist, especially regarding plan dose verification before and during treatment delivery. This aspect is particularly crucial in sites affected by organ motion, such as the upper gastrointestinal (UGI) area (including the liver, pancreas, and esophagus) and the lung. Previous studies have investigated dose reconstruction using delivery log files and 2D cine image series in prostate treatment [1]. Offline, post-treatment procedures for recomputing dose on additional datasets have also been reported [2]. In this study, we propose a fast and automated procedure to assess plan quality given various degrees of intra-fraction motion on any verification 3D scan at any point in the online adaptive workflow. The methodology also facilitates (a) the implementation of reinforced online adaptive strategies, including the generation of bespoke clinical goals between subsequent treatment fractions and/or (b) dose accumulation.

2 Materials and Methods

The Unity MR-Linac (Elekta) workflow is streamlined to include the acquisition of 3D imaging for planning, DIR-based contours mapping, and/or manual re-contouring of targets and organs-at-risk (OARs), generation of a fully-optimized treatment plan, real-time motion monitoring via 2D cine imaging in multiple planes, 3D verification scans, and treatment delivery [3]. For UGI sites as well as for lung patients, the treatment sessions are on average 60 minutes long, and can extend up to 90 min for the single fraction cases of 24-30 Gy due to prolong beam-on time [3].

The automated dose verification pipeline was developed using RayStation's Python-based scripting environment and validated for 21 patients, including liver treated with 1/3/5 fractions, as well as pancreas, abdominal oligometastases, and lung patients who received 5-fraction SBRT treatments.

The pipeline input comprised the adapt-to-shape (ATS) plan, including 3D planning images (i.e., time-averaged or navigator-triggered in exhale phase of the breathing cycle), contours, and 3D dose maps, as well as 3D image series from the verification (VER) step of the online workflow and similar data acquired during treatment beam-on delivery (BON). The automated analysis involved the sequential execution of the following steps: (a) data sorting and import from the RayStation data repository (data received from Online Monaco and Unity-MR console), (b) image and plan nomenclature update to facilitate robust data processing, (c) deformable image registration (DIR) based on image intensity between the ATS image data and VER/BON data, (d) contours mapping via DIR from ATS to VER/BON, (e) addition of site and prescription specific clinical goals , (f) transfer of contours rigidly from VER/BON to ATS to reflect any potential changes due to setup or intra-fraction motion, (g) dosimetric and similarity analysis for all relevant ROIs, and (h) generation of a summary report including dosimetric evaluation of clinical goals and assessment of DIR performance.

The DIR and contours mapping processes were visually validated by independent reviewers. Dose invariance was assumed between the ATS and VER/BON datasets, thus avoiding the need for dose recomputation. This assumption was validated by performing dose recomputations in

Offline Monaco for select cases covering the full range of clinical scenarios investigated.

3 Results and Discussion

The retrospective analysis involved 100 treatment fractions, encompassing the evaluation of 620 DIR-generated ROIs. Over 98% of the contours were accurately propagated within the region relevant to the treatment volume, with minor deviations having a negligible impact on the accuracy of reconstructed dose. Additionally, 30 plans were recomputed offline to confirm that the agreement between the ATS and BON/VER data was within 1%, which falls within the tolerance of Monaco's dose in magnetic field computations based on Monte Carlo.

The DICE similarity metric, readily available via RayStation scripting, was utilized to detect intra-fraction motion relevant to plan dosimetry. DICE was evaluated for both full organs and fractional volumes within the treatment region, defined as a 4 cm extension from PTV. The DICE values for fractional volumes proved to be more sensitive. ROI similarity analysis between ATS and VER/BON within the treatment region strongly correlated with intra-fraction changes in OARs, flagging dosimetric discrepancies. It is possible to define site-specific DICE thresholds to trigger user investigation into the DIR quality. The dose verification analysis detected outliers not initially identified during the online sessions (Fig. 1). These discrepancies were not necessarily specific to a given site or prescription.

In the ATS to BON analysis, relevant to the evaluation of dose-of-the-day, the large bowel emerged as the OAR with the highest level of variation, with roughly 33% and 40% of set clinical goal value for liver and pancreas 5-fraction cases, respectively. Duodenum, stomach, and small bowel ranged from 6% to 25% variations in the clinical goal values for liver and pancreas 5-fraction treatments. The largest change from all cases investigated was for a single liver case – duodenum dose varied by roughly 60%.

For the 2 lung cases investigated, the maximum variations between ATS and BON clinical goal values for OARs (heart, esophagus, bronchus) across all treatment fractions were found to be within 10%.

The maximum dose variation between the VER and BON analysis for UGI sites was found to be within ±100 cGy for the target and all OARs. The change for the mean dose for healthy liver (liver minus GTVs) was more modest, within ±20 cGy. The above findings underscore the need for a dose verification analysis on the VER data set to support the clinical decision-making process regarding whether to pass or fail an ATS plan immediately prior to treatment delivery. In contrast, the analysis on the BON data acts as a post-delivery data generation process to assess how the plan was delivered and generate data that can be used to link multiple treatment fractions. We are currently investigating the automated generation of updated clinical goals based on a given online plan, which can be propagated as the reference for subsequent treatment fractions. This type of data can reinforce the adaptive process from the online sessions.

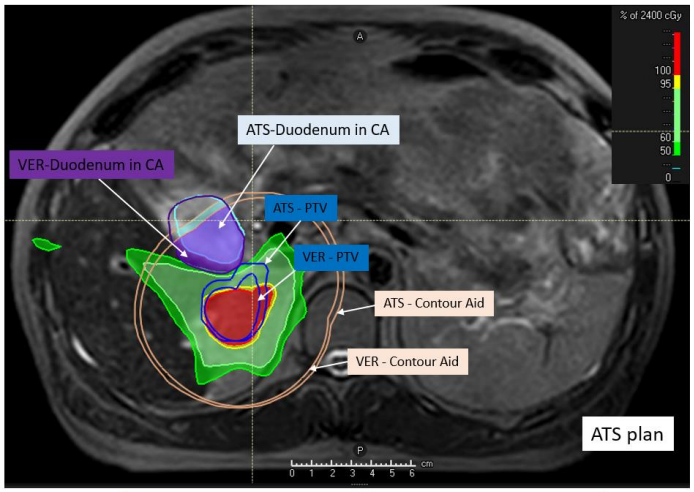


Figure 1. Example of the ATS-VER analysis highlighting intra-fractional changes which led to an increase in dose delivered to the duodenum.

4 Conclusion

We developed and validated a dose verification automated analysis pipeline for UGI and lung cancer patients treated with SBRT on the Unity MR-Linac system. The study demonstrated the importance of such an analysis in aiding the online adaptive process and facilitating future advancements in methodology for further adapting online treatments (reinforced adaptation). If efficiently implemented within the MR-Linac workflow, it can serve as a robust on-demand tool for quickly assessing patient setup and intra-fraction motion. Currently, we are initiating a pilot study to deploy and evaluate the dose verification pipeline as an online clinical tool.

References

[1] Menten M. J. et al. "Automatic reconstruction of the delivered dose of the day using MR-linac treatment log files and online MR imaging". Radiother Oncol. 2020; 145: 88–94. DOI: 10.1016/j.radonc.2019.12.010

[2] Tyagi N. et al. "Feasibility of ablative stereotactic body radiation therapy of pancreas cancer patients on a 1.5 Tesla magnetic resonance-linac system using abdominal compression". Phys. Imaging Radiat. Oncol. 2021, 19, 53–59. DOI: 10.1016/j.phro.2021.07.006.

[3] Stanescu T. et al. "MRI-Guided Online Adaptive Stereotactic Body Radiation Therapy of Liver and Pancreas Tumors on an MR-Linac System." Cancers. 2022; 14: 716. DOI: 10.3390/cancers14030716.

Optimizing the contrast to noise ratio for tumor tracking in the lung using dual energy imaging strategies for soft tissue enhancement

Nawal Alqethami[1], Wentao Xie[1], Prasannakumar Palaniappan[1], Felix Ginzinger [2], Philipp Steininger [2], Marco Riboldi[1]
1. Department of Medical Physics, Ludwig-Maximilians-Universität, Munich, Germany
2. medPhoton GmbH, Salzburg, Austria

Abstract: Lung tumor tracking using x-ray encounters challenges, primarily due to shadowing effects from bony anatomy. This study addresses the experimental optimization of the soft tissue contrast and noise reduction in Dual Energy (DE) imaging-based tumor tracking relying on fast energy switching. DE Cone-Beam CT(CBCT) projections were acquired on a mobile CBCT scanner with a modified CIRS phantom in static mode at two energy settings (80-120 kV) and (60-120 kV) to study the effect of various energy gaps. Tumor inserts with different materials (adipose tissue, breast, muscle, and liver) were tested, by selecting projections with maximal overlap with the bony anatomy. Deformable registration was utilized to correct the misalignment between frames. We specifically investigated (i) the optimal weight in the weighted logarithmic subtraction (WLS) to enhance the soft tissue target and (ii) the optimal image filter to reduce noise. Median filters (MF) of different kernel size and anti-correlated noise reduction (ACNR) were considered.
The impact of (60-120 kV) setting demonstrated higher CNR, emphasizing its superior bone suppression and soft tissue enhancement in comparison with (80-120KV).
The result indicates optimal weighting factors of (0.65-0.7) for (60-120 kV) and (0.75-0.8) for (80-120 kV), aiming at maximal contrast to noise ratio (CNR) in the target. For a smaller energy range, smaller kernel size was deemed appropriate, with an optimal MF size of 11 pixels for the (80-120 kV) and a size of 9 pixels for (60-120 kV). The best noise suppression was achieved with ACNR+MF.
In conclusion, the study showed the effectiveness of using DE imaging and WLS to optimize the CNR to advance lung tumor tracking capabilities.

1 Introduction

X-ray based tumor tracking in the lung encounters a significant hurdle, primarily caused by the interference of high-density structures, notably bony anatomy. The presence of superimposed shadows cast by these bony structures poses a limitation on the contrast of lung lesions in images, frequently leading to missed tracking opportunities (1).
The visual attributes of bones take on a distinctive pattern, showcasing reduced contrast in comparison to the surrounding soft tissues when high kilovoltage peak (kVp) settings are employed. Conversely, a shift to lower kVp settings alters the appearance of bones, resulting in an increased contrast relative to the surrounding soft tissues (2).
In response to these challenges, the application of Dual Energy (DE) imaging comes into play. This advanced technique involves the acquisition of two images at different x-ray energies, followed by a weighted logarithmic subtraction (WLS). The outcome is a third image intentionally designed to cancel either bone or soft tissue selectively (3). Despite its effectiveness in mitigating bone interference, the DE method often yields images with a perceptible increase in noise.
This study aims to evaluate the efficacy of the DE and WLS approach. The primary objective is to gauge its success in suppressing the visibility of bones while concurrently enhancing the visual representation of soft tissue. Moreover, the study will incorporate various noise reduction methods, contributing to the continual improvement of tumor tracking capabilities within the lung.

2 Materials and Methods

DE-Cone-Beam CT (CBCT) projections were acquired using the mobile ImagingRing (IRm) developed by medPhoton GmbH (Salzburg, Austria) (Figure 1A). Images were acquired using the dynamic CIRS MRI-Linac Phantom (Sun Nuclear, Melbourne, Florida) in static mode.
The phantom was adjusted to incorporate cylindrical inserts of different target materials with about 30 mm in diameter and 50 mm in length, taken from the CIRS Model 062 Electron Density Phantom. These materials include adipose, breast, muscle, and liver tissues (Figure 1C).
In addition, pork ribs were attached on the phantom to simulate a more intricate representation of the human anatomy as illustrated in Figure 1B.
The acquisition process resulted in a set of eight DE-CBCT scans with the X-ray generator operating in pulsed mode and the detector set to 0.3 mm pixel spacing.
Each target material underwent two scans, with the first scan employing settings of 80 kV (15 mA, 17 ms) and 120 kV (15 mA, 12 ms), while the second scan using 60 kV (15 mA, 17 ms) and 120 kV (15 mA, 12 ms).

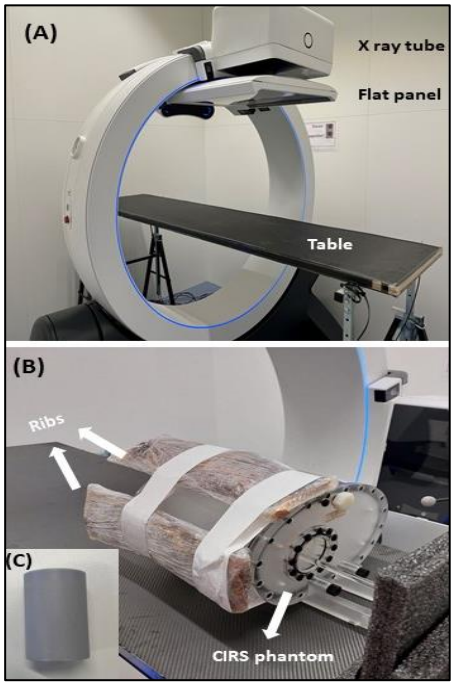


Figure 1: Experimental setup. Panel (A) shows the mobile ImagingRing (IRm). Panel (B) depicts the experimental setup of the phantom and the ribs with panel (C) showing the CIRS liver insert.

Each scan consisted of X-ray frames with alternating energies, capturing a complete 360° rotation within 120 seconds, with a 165 ms and 84 ms interval between consecutive frames respectively.
From each experimental study, two frames were manually selected which contained the maximum overlap between the ribs and the target.
To rectify the misalignment between frames, a deformable registration was employed, followed by images WLS calculated as in equation (1), using a range of weighting factors (w_{st}) within the interval [0.5, 0.95] with 0.05 spacing:

$$I_{DEst} = ln\ I_H - w_{st}\ ln\ I_L \quad (1)$$

In this equation, I_{DEst} refers to the DE soft tissue image, with I_H representing the input from high energy (HE) image, and I_L representing the input from low energy (LE) image.
Separate regions of Interest (ROIs) for bone and soft tissue were segmented, while avoiding sharp intensity transitions (e.g., bone edges). Subsequently, the CNR between the two ROIs was computed, and the optimal weighting factor was identified as the one yielding the lowest CNR to eliminate the bone. Figure 2 illustrates the segmentation employed for CNR calculation in the evaluation of the optimal weighting factor.
It was observed that the WLS process resulted in a compromise in image quality. In response to this trade-off, two noise reduction methodologies were applied to enhance image quality and increase the CNR for the soft tissue target.

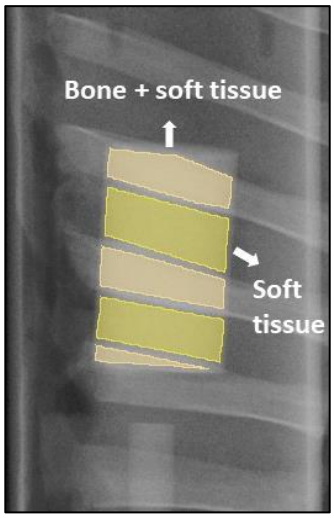


Figure 2: The segmentation used to the calculate the CNR

The CNR for the soft tissue insert was calculated between the segmented insert and the background, obtained by expanding the segmentation as illustrated in Figure 3.

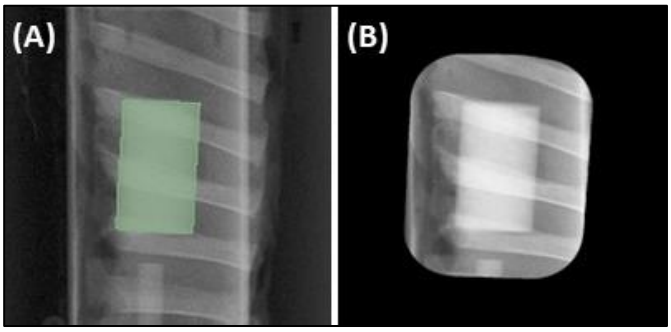


Figure 3: Panel (A) shows the tumor insert segmentation. Panel (B) represents the expanded segmentation region to include the background.

Different competitive approaches were applied for filtering the resulting DE image. The first approach was based on Median Filtering (MF) (4).A range of filter sizes, spanning from 9 to 11 pixels, was tested, with the final selection being guided by the achievement of a higher soft tissue CNR value combined with minimal image blurriness.
The second approach incorporated the anti-correlated noise reduction (ACNR), a technique to reduce noise in material specific images generated by WLS. (5) (6). This method relies on the fundamental observation that there exists a high correlation between the noise present in soft tissue and bone images.
To address noise reduction in the soft-tissue image, the ACNR algorithm applies a high-pass filter to the complementary image, precisely the bone-only image (I_{DEB}).

This filtering operation eliminated the bone component from the complementary image, leaving only the residual noise. Subsequently, the original DE soft-tissue image (I_{DEst}) and the high-pass filtered bone image (I_{DEB}^{HPF}) were combined using a weighted subtraction, incorporating a parameter denoted as w_n in equation (2):

$$I_{DEst}^{ACNR} = I_{DEst} - w_n \, I_{DEB}^{HPF} \quad (2)$$

A Wilcoxon matched-pair test was used to investigate potential variations in the CNR between the specified energy ranges of 80-120 kV and 60-120 kV. Additionally, both the Friedman test and the post-hoc Nemenyi test were utilized to systematically evaluate the statistical significance of differences observed among various noise reduction techniques in comparison with the DE-enhanced input images. These tests were essential for investigating the significance in CNR performance associated with different energy ranges and filter type.

3 Results

Figure 4 depicts an example of the original frame before and after the WLS using a combination of different weighting factors for both (80-120 kV) and (60-120 kV).
For the (80-120 kV) energy range, the optimal weighting factor, determined through quantitative CNR values and visual assessments, ranged between 0.75 and 0.8. On the other hand, for the (60-120 kV) energy range, the results indicated that the optimal weighting factor is within the range of 0.65 to 0.7, across all material tissues tested in the study.

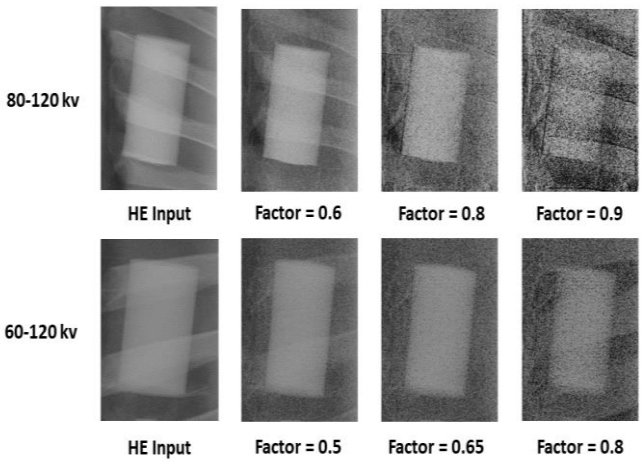


Figure 4: Weighted logarithmic subtraction for the breast target

The noise reduction techniques including MF and ACNR were carried out, with the ACNR parameter (w_n) in equation (2) optimized to the optimal weight of 0.95. This aimed at achieving a balance in the noise reduction process, and the contribution to the enhancement of the image. Furthermore, an investigation into a composite approach, including MF + ACNR, was conducted, and the outcomes are visually presented in Figure 5.

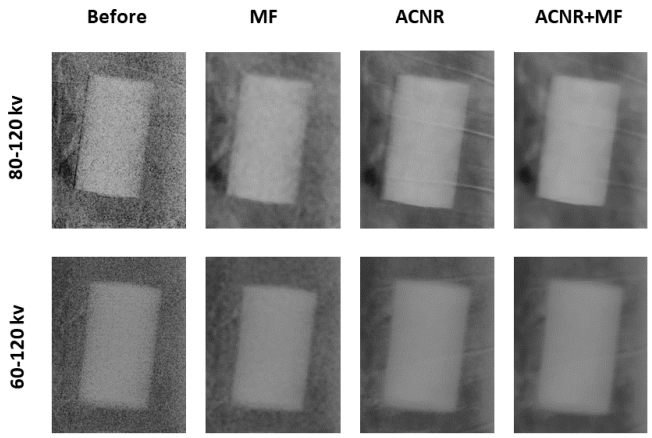


Figure 5: Result of MF, ACNR, and MF+ACNR applied on (80-120 kV) using w_{st}= 0.8 and (60-120 kV) using w_{st}= 0.65

The CNR was computed for both HE input frame and the LE input frame. The HE input frame yielded a superior CNR value compared to the LE input frame. Consequently, the HE frames CNR were chosen for subsequent comparison with the CNR values obtained using different noise reduction techniques.
The obtained p-value in the Friedman test, which stands at 1.81 x 10^{-5}, indicates a statistically significant difference across the various techniques.
The mean ranks resulting from the post-hoc Nemenyi test are plotted in Figure 6.

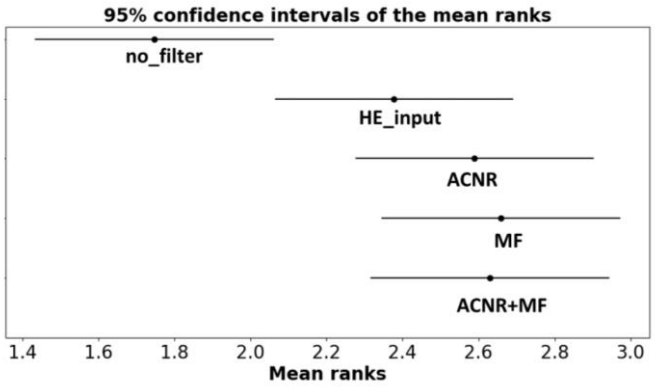


Figure 6: A plot of the mean ranks from smallest to largest for all noise reduction techniques compared to the HE input

Table 1 shows the p-values of the one to one comparison, with the asterisk marks indicating significant result at 95% confidence.

	HE Input	No filter	MF	ACNR	ACNR+MF
HE Input	1.000	0.687	0.084	0.319	0.014*
No filter		1	0.001*	0.014*	0.001*
MF			1	0.900	0.900
ACNR				1	0.687
ACNR+MF					1

Table 1: p-values of the Nemenyi test, with * indicating a statistically significant difference at 95%

The median CNR value for the HE input is presented in Table 2 for comparative analysis before and after the implementation of various noise reduction techniques.

	HE Input	No filter	MF	ACNR	ACNR+MF
80-120kv	2.19	1.52	2.33	2.44	2.49
60-120kv	2.77	2.04	3.25	2.97	3.01

Table 2: Median CNR values before and after introducing noise reduction methods

The difference in CNR between the energy ranges of 80-120 kV and 60-120 kV was statistically significant (Wilcoxon matched pair test), with a p-value of 0.017.

4 Discussion

The quantitative CNR values incorporated with visual assessments revealed that the (80-120 kV) energy range exhibits optimal weighting factors ranging between 0.75 and 0.8. Conversely, for the (60-120 kV) energy range, our results demonstrate that the optimal weighting factors fall within the range of 0.65 to 0.7 with the result being consistent across all material tissues used in this experiment.

The assessment of various noise reduction techniques within this study revealed distinct performances in CNR. Without the use of a filter, the outcome obtained the lowest CNR value when compared to the HE input, indicating potential limitations in terms of reduced image quality and increase in the noise. ACNR+MF exhibited best results, outperforming other techniques and displaying a higher CNR value compared to the HE input, suggesting its effectiveness in elevating image quality.

It is notable that for MF enlarging the filter size leads to an increase in image blurriness. A correlation was observed, indicating that a smaller energy range requires a correspondingly smaller filter size. Specifically, for the (80-120 kV) range, an optimal MF size of 11 pixels was deemed appropriate. In contrast, for the (60-120 kV) range, a size of 9 pixels was considered suitable for optimal image quality.

5 Conclusion

The study tested the effectiveness of utilizing DE imaging and WLS, highlighting their efficacy in optimizing the CNR for the advancement of lung tumor tracking capabilities. The observed variations in optimal weighting factors for the two energy ranges attributed to differences in the attenuation characteristics of tissues at various energy levels. A smaller weighting factor will be optimal as the gap between the two energy ranges is greater. (7)

A comprehensive assessment of various noise reduction techniques to reduce the image noise resulting from WLS process was carried out. Among the techniques examined, ACNR+MF emerged as the top performer, possessing a higher CNR value than the HE input.

The outcomes of the study conducted between the energy ranges of 80-120 kV and 60-120 kV, emphasizing the impact of energy range variation on the observed CNR values. Notably, the 60-120 kV range, characterized by a greater energy separation, exhibited higher CNR, showcasing greater bone suppression and soft tissue enhancement. (7)

Future work will focus on optimizing imaging protocols to achieve superior bone suppression and enhanced soft tissue visibility while minimizing patient dose on full rotation DE-CBCT scans. Non-linear methods to achieve WLS are currently being tested, aiming at soft-tissue enhanced images (3). The exploration will also include motion studies in tumor inserts with different size and shape.

References

1. Patel R, Panfil J, Campana M, Block AM, Harkenrider MM, Surucu M, Roeske JC. Markerless motion tracking of lung tmors using dual-energy fluoroscopy. Medical physics. 2015 Jan;42(1):254-62. DOI:10.1118/1.4903892.

2. Jerrold T. Bushberg, J. Anthony Seibert, Edwin M. Leidholdt, Boone JM. The essential physics of medical imaging. Philadelphia, PA: Lippincott Williams and Wilkins; 2011.

3. Brody WR, Butt G, Hall A, Macovski A. A method for selective tissue and bone visualization using dual energy scanned projection radiography. Medical physics. 1981 May;8(3):353-7. DOI:10.1118/1.594957

4. Ohki M, Zervakis ME, Venetsanopoulos AN. 3-D Digital Filters. Multidimensional Systems: Signal Processing and Modeling Techniques. Control and Dynamic Systems1995. p. 49-88.

5. Kalender WA, Klotz E, Kostaridou L. An algorithm for noise suppression in dual energy CT material density images. IEEE transactions on medical imaging. 1988 Sep;7(3):218-24. DOI:10.1109/42.7785

6. Kaur M, Wagstaff P, Mostafavi H, Lehmann M, Morf D, Zhu L, Kang H, Walczak M, Harkenrider MM, Roeske JC. Effect of different noise reduction techniques and template matching parameters on markerless tumor tracking using dual-energy imaging. Journal of Applied Clinical Medical Physics. 2022 Dec;23(12):e13821. DOI:10.1002/acm2.13821

7. Haytmyradov M, Patel R, Mostafavi H, Surucu M, Wang A, Harkenrider MM, Roeske JC. A novel phantom for characterization of dual energy imaging using an on-board imaging system. Physics in Medicine & Biology. 2019 Jan 21;64(3):03NT01. DOI:10.1088/1361-6560/aaf9

Performance Evaluation of Virtual Laser-Based Surface-Guided Radiation Therapy Setup System Using Humanoid Head Phantom

Jeong-Woo Lee[1,2], Byoung-Moon Park[1], Yong-Ki Bae[1], Min-Young Kang[1] and Semie Hong[1,2]

[1]Department of Radiation Oncology, Konkuk University Medical Center, Seoul, Korea
[2]Department of Medical Physics Convergence, Graduate School of Engineering, Konkuk University, Seoul, Korea

Abstract : We aimed to verify patient setup accuracy using a combination of virtual laser-based SGRT systems (LUNA 3D, LAP, Germany). The experiments were divided into (1) skin-surface guidance with virtual laser and (2) mold-surface guidance with optical laser. The humanoid head phantom setups were analyzed into vertical (mm), lateral (mm), longitudinal (mm), and rotational (yaw in degree). In the virtual laser-based SGRT results, the displacement in the lateral direction is less than 0.5 mm, and the displacement in the longitudinal direction is more pronounced. In comparison, the vertical direction shows a difference of about 1 mm. The mold-based optical laser system showed vertical and longitudinal direction errors up to 2 mm. The SGRT system with the virtual laser can provide accurate alignment information of a large area and the area of interest set before treatment, providing accuracy of treatment and stable treatment setup.
Keywords: SGRT, virtual laser, optical laser, head phantom

1 Introduction

Radiotherapy technology has generally been developed along the two axes of efficient intensity modulation and image-guidance technologies.

In recent years, surface-guided radiation therapy (SGRT) has begun to replace traditional laser-based setups. It is even used as an alternative to patient motion management in image-guided radiation therapy (IGRT). Also, the installation of non-invasive and non-radiographic devices for managing patient setup posture before and during treatment using optical surface imaging technology has been expanding. Patients move against their will during treatment, which can lead to significant errors in high-precision treatments such as intensity-modulated radiation therapy (IMRT) or stereotactic body radiation therapy (SBRT).

Since the setup using an optical laser-based system uses ink or pen directly on the patient's skin to draw lines according to the laser projected from 3 directions (frontal & bilateral), the systemic error of the laser system itself and the random error of the radiotherapist still affect the quality of the patient's treatment. In addition, if a tattoo is applied to the patient's skin, it cannot be erased, and if the line drawn with ink is smeared or erased by sweat or water, CT simulation may have to be repeated.

One of the most popular treatment areas for SGRT is the breast, and studies have shown that tattoo-based treatments can reduce cosmetic and psychological burdens, as well as improve reproducibility and accuracy of patient positioning. These results have been confirmed by studies that have expanded the application to the chest, abdomen, pelvis, and head and neck.

In this study, we aimed to verify patient setup accuracy using a combination of virtual laser and optical camera using dicom information from the CT-based treatment plans.

2 Materials and Methods

The equipment used in the experiment was a C-arm type linear accelerator (iX, Varian, USA) and an SGRT system (LUNA 3D, LAP, Germany) with a virtual laser function. To realize accurate equipment performance, once the equipment was installed, calibration was performed using a cubic phantom and plate for calibration (Figure 1). The calibration of the virtual laser is crucial in the calibration process of SGRT equipment, especially in checking the agreement with the precisely calibrated CBCT.

To evaluate the performance of the LUNA 3D system, CT simulation was performed using a head phantom (Edward model 605, CIRS, USA) with and without a thermoplastic mask (Figure 2). CT scans (LB, Canon, Japan) were performed at 1 mm thickness to reconstruct 1 mm CT slices to minimize the effect of the resolution of the fine reference surface.

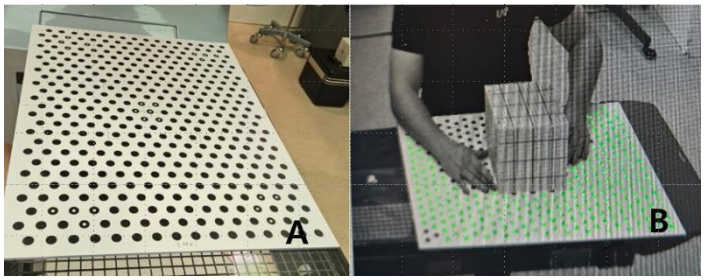


Figure 1: LUNA 3D system calibration (A) SGRT calibration plate (B) cubic phantom for SGRT calibration

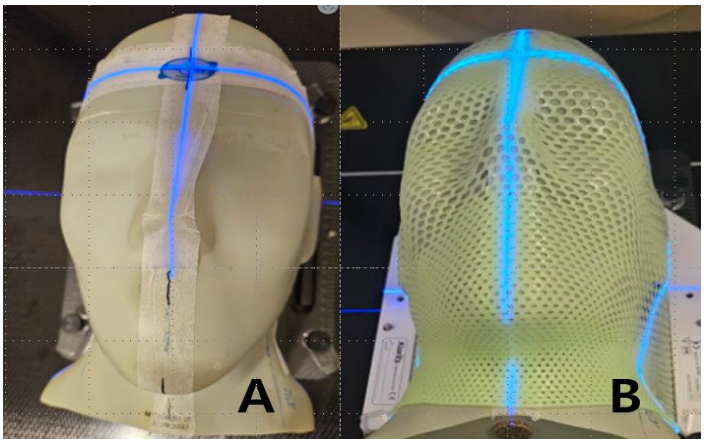


Figure 2: CT simulation using a head phantom (Edward model 605, CIRS, USA) without (A) and with (B) a thermoplastic mask

A volumetric modulated arc therapy (VMAT) treatment plan was established for the phantom plan, and the dicom

RT plan and dicom RS files were generated and imported into the LUNA 3D server.

The experiments were divided into (1) skin-surface guidance with virtual laser and (2) mold-surface guidance with optical laser. The phantom setups were analyzed into vertical (mm), lateral (mm), longitudinal (mm), and rotational (yaw in degree). The calibrated cone beam CT (CBCT) was used as a baseline to analyze the errors. The CBCTs were acquired in half fan mode with a scan condition of 512 x 512 pixels and a slice distance of 2 mm (Figures 3 and 4). Each setup was performed three times; the average and deviation were calculated, and the magnitude was analyzed by calculating the composite value of the error. The region of interest (ROI) of the skin surface was set to the whole frontal surface of the phantom. The virtual laser was referenced to the setup by matching the coordinate axis of the isocenter set during calibration with the value provided by the SGRT system based on the dicom information of the CT image (Figure 5).

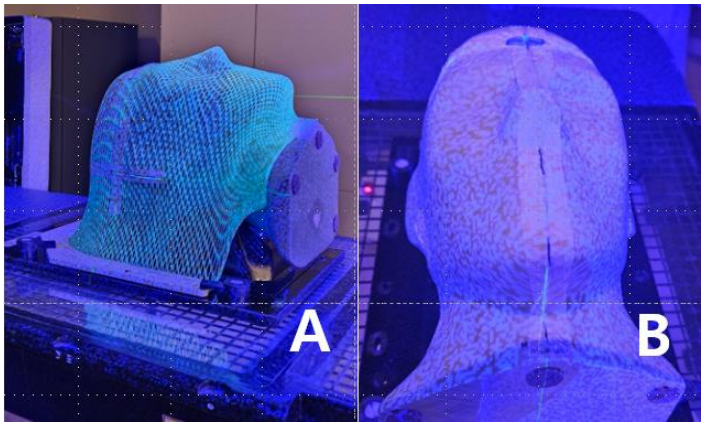


Figure 3: SGRT setup with the optical and the virtual lasers (A) mold-surface guidance with the optical laser vs. (B) skin-surface guidance with the virtual laser using humanoid phantom

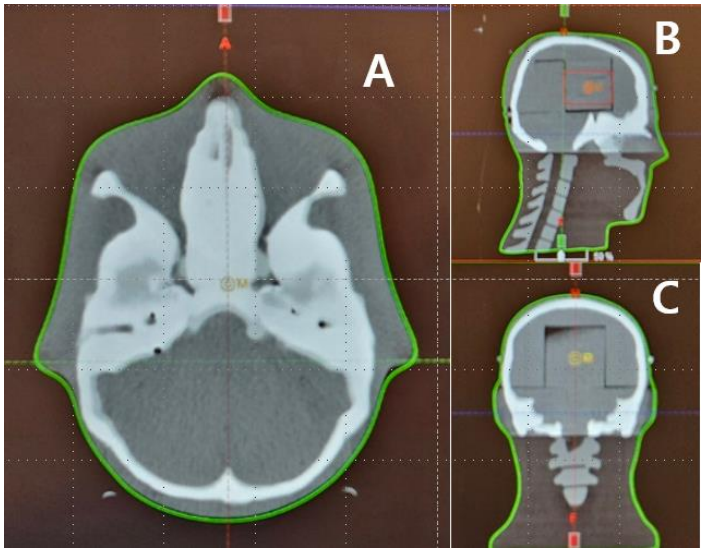


Figure 4: CBCT imaging to analyze the setup error. (A) axial CT image, (B) sagittal CT image, and (C) the coronal image

3 Results

Table 1 shows the results of calculating the displacement by taking a CBCT after setting up the humanoid phantom based on the virtual laser-based skin-surface guidance. As you can see from the results, the displacement in the lateral direction is less than 0.5 mm, and the displacement in the longitudinal direction is more pronounced. In comparison, the vertical direction shows a difference of about 1 mm. The magnitude of the three directions showed results in the range of 1 to 1.5 mm. In all three measurements, the displacement in the yaw direction was less than 0.1 degrees.

SG+VL*	Vt[1] (mm)	Lt[2] (mm)	Lg[3] (mm)	Mg[4] (mm)	Yw[5] (°)
1st	0.3	0.5	1.4	1.5	-0.1
2nd	-0.6	-0.3	-1.0	1.2	0
3rd	-1.0	0	-0.2	1.0	-0.1

Table 1: Skin-surface guided setup using the virtual laser (SG+VL). The acquired CBCT images were used to analyze the deviations: Vt[1] (vertical), Lt[2] (lateral), Lg[3] (longitudinal), Mg[4] (magnitude), and Yw[5] (yaw). Table 1 shows the results of calculating the displacement using an SGRT system after CBCT imaging correction.*

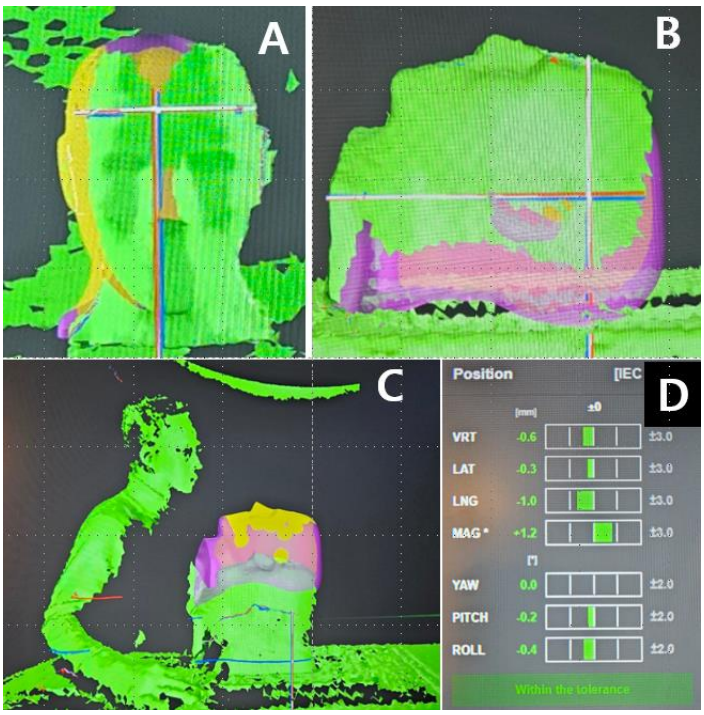


Figure 5: SGRT setup screen with the virtual lasers(blue lines) and patient dicom center (red lines) (A) frontal view, (B) left lateral view, (C) red crosslines to blue crosslines matching operation, and (D) mismatching results based on the virtual lasers and the phantom's dicom center

MG+OL*	Vt[1] (mm)	Lt[2] (mm)	Lg[3] (mm)	Mg[4] (mm)	Yw[5] (°)
1st	0.0	0.0	2.0	2.0	-0.1
2nd	2.0	0.0	1.0	2.2	-0.6
3rd	1.0	0.0	2.0	2.2	0.0

Table 2: Mold-surface guided setup using the optical laser (MG+OL). The acquired CBCT images were used to analyze the deviations: Vt[1] (vertical), Lt[2] (lateral), Lg[3] (longitudinal), Mg[4] (magnitude), and Yw[5] (yaw). Table 2 shows the results of calculating the displacement using the CBCT system after a mold-based surface-guided setup.*

On the other hand, the results of the optical laser-based mold-surface guided setup error are shown in Table 2. As shown in Table 2, the error in the lateral direction was zero in all three cases, but the errors in the vertical and longitudinal directions were up to 2 mm. The magnitude was about 2 mm, and the vertical and longitudinal directions were complementary, just like the virtual laser-based setup.

4 Discussion

In radiotherapy, stable patient positioning is more fundamental than any other technique, but it is common to all modalities and is essential for successful treatment.
In recent years, this has been the subject of active research, particularly in the breast deep-inspiration breath hold (DIBH) technique, which is synchronized with breathing, and patient motion management interfaces for high-precision radiotherapy such as frameless stereotactic radiosurgery. Traditional vendors in this SGRT market include VisionRT's AlignRT, C-RAD's Catalyst, and Varian's IDENTIFY, and their use is expected to expand as insurance coverage becomes available. Technically, many variables are utilized to evaluate the quality of an instrument, including data compatibility with the instrument, surface information acquisition and processing speed (CPU, GPU, etc.), treatment area, accuracy and reproducibility, and image resolution. Therefore, when introducing SGRT equipment, the equipment's characterization and quality evaluation are essential, and validation of new functions is necessary before use.

In this study, the functions of LAP's virtual laser with new functions were analyzed compared to the general optical laser system, and the significance of the research process was described.
1) After hardware installation of the SGRT solution and virtual laser calibration, cross-validation with a well-commissioned CBCT is necessary.
2) Comparison and verification of surface generation according to the resolution of dicom CT images is necessary.
3) Verify the correlation between the response speed of the patient's interfractional motion and the drive of the treatment device.
In the initial process of this study, we found a systematic error in the calibration of the virtual laser, which showed a difference of 1.7 mm in the same direction as CBCT, which was corrected. In addition, we found a slight difference in the 3D surface generated depending on the resolution of the reconstructed CT image, which may be important to note when using dicom images as a direct reference surface, and we believe that further research is needed. Finally, we found a delay in positional detection during the setup process based on the virtual laser, and sufficient research is needed for treatments based on real-time monitoring, such as motion tracking. This study aimed to compare the effectiveness of surface guidance based on a virtual laser to the mask molding surface guidance of an optical laser. We confirmed its usefulness with CBCT as a baseline.

5 Conclusion

In this study, the usefulness of the virtual laser-based SGRT system was compared to the conventional SGRT with the optical laser system to check the accuracy and reproducibility. By utilizing the virtual laser system, which is not adopted in the conventional SGRT system, it was possible to improve the accuracy of the setup through the information of the alignment during the patient setup, which suggested the possibility of reducing the setup time and additional imaging dose. We believe that the SGRT system, with the function of the virtual laser, can provide accurate alignment information of a large area and the area of interest set before treatment, which can provide accuracy of treatment and stable treatment setup.

References

[1] H.A. AI-Hallaq, L. Cervino, A. N. Gutierrez, et al. “AAPM task group 302: Surface-guided radiotherapy”. Med Phys 49 (2022), e82-e112. DOI: 10.1002/mp.15532.

[2] J. Rigley, P. Robertson, L. Scattergood. “Radiotherapy without tattoos: Could this work?”. Radiography 26 (2020), pp. 288-293. DOI: 10.1016/j.radi.2020.02.008

[3] P. Freislederer, M. Kügele, M. Ö llers, et al. “Recent advances in Surface Guided Radiation Therapy”. Rad Oncol 15:187 (2020),. DOI: 10.1186/s13014-020-01629-w.

[4] Z. Li, Q. Xiao, G. Li, et al. “Performance assessment of surface-guided radiation therapy and patient setup in head-and-neck and breast cancer patients based on statistical process control”. Phys Med 89 (2021), pp. 243-249. DOI: 10.1016/j.ejmp.2021.08.007

Preliminary evaluation of a software application for target tracking without fiducials during Cyberknife treatment of pancreas

James Bedford, Simeon Nill and Uwe Oelfke

Joint Department of Physics, The Institute of Cancer Research and The Royal Marsden NHS Foundation Trust, London, United Kingdom

Abstract Cyberknife is an effective treatment for pancreatic cancers but visualization of the planning target volume (PTV) on the orthogonal kV imaging system is usually accomplished by means of fiducial markers. Insertion of the markers can lead to complications, so this study aims to visualize and track the pancreatic PTV without fiducials, using in-house software CKImage v1.1. The PTV delineated on the planning computed tomography scan is projected onto digitally reconstructed radiographs (DRRs) in the image planes to define the region of interest. The acquired kV images are filtered and the average image is subtracted from the individual images to remove stationary features such as bony anatomy. Rigid registration is then used to determine the offset of the actual pancreas from that defined on the DRR. The method is applied to a high-contrast test object to verify the performance of the software, and also retrospectively applied to four patients with pancreatic cancer. In each case, the results are compared with the actual position of the PTV as determined using Cyberknife itself, using fiducials for the patient cases. The median accuracy of tracking with CKImage for the 177 image pairs acquired of the high-contrast test object is 0.8 mm (range 0.0 – 21.8 mm). The median accuracy of tracking for the 1249 image pairs acquired of the pancreas patients is 4.1 mm (range 0.2 – 16.1 mm). There is much scope for improvement in the method, but the novel imaging approach shows promise for minimally invasive treatment of pancreas cancer.

1 Introduction

The Cyberknife system (Accuray Inc., Sunnyvale, CA) consists of a short waveguide 6 MV linear accelerator mounted on a robotic arm, and with a choice of fixed cone collimator, variable iris collimator or high resolution multileaf collimator, is ideal for intracranial or extracranial stereotactic radiotherapy. The planning target volume (PTV) is tracked accurately using two orthogonal kV imaging systems, mounted at 45° to the horizontal plane [1].

The Cyberknife system has great potential for treatment of pancreatic cancer, where stereotactic ablative body radiotherapy (SABR) is currently of considerable interest. The imaging system of the Cyberknife can be used to track the respiratory motion of the target volume in real time, thereby minimizing margins and dose to surrounding critical structures. However, the pancreas is currently not explicitly visible on the kV images, and fiducial markers are therefore inserted to enable tracking. Although this is a successful approach, insertion of the markers can lead to clinical complications, so tracking without fiducial markers is a desirable research goal.

Other authors have reported on tracking pancreas using kV images without fiducial markers. For example, Zhou et al. [2] train a deep learning prediction model for each patient to predict the contour location based on the digitally reconstructed radiograph (DRR). They then test the method on further orthogonal DRRs and are able to predict the location of the clinical target volume (CTV) to within 1 mm accuracy. The scenario described is well-defined, so high accuracy is to be expected, but the results are promising for markerless tracking. A related work studies inter-patient organ-at-risk deformation in the organs surrounding the pancreas [3].

These studies form the basis of a more general and challenging study of mapping 3D organ models onto 2D kV images in the presence of organ motion [4]. For the pancreas gross tumor volume (GTV), the surface of the volume is predicted with a median accuracy of 1 mm, maximum error of 5 mm and dice similarity coefficient of around 90%. Similar results are obtained for the other abdominal organs, but with a maximum error of around 10 mm. Accuracy is approximately 1 mm worse when setup errors are included. This is therefore a more general model which is capable of transforming 3D organ structures according to 2D images, which is very relevant to the Cyberknife scenario.

The increasing experience of magnetic resonance (MR)-guided treatment is also a possibility for radiotherapy of pancreatic adenocarcinoma [5]. Stemkens et al. [6] report a study of 4DMR pancreas imaging in which they produce 10 k-spaces at the 10 phases of the breathing cycle. They use deformable image registration to quantify the motion of the pancreas and duodenum. Grimbergen et al. [7] treat 13 patients using stereotactic ablative body radiotherapy (SABR) to 5 x 8 Gy in the MOMENTUM study. Motion is minimized using an abdominal corset. Interfraction motion is found to be stable, and breathing amplitudes are in the order of 1-4 mm, these amplitudes being quantified in each of the three cardinal directions.

These studies are the motivation for the present work, which aims to develop a software system capable of tracking pancreas without fiducial markers with sufficient speed and accuracy that it can be used in real time to track pancreatic target volumes during treatment on the Cyberknife. In this paper, the software is described, the methods outlined and benchmarking tests reported in which the current speed and accuracy of the system are evaluated for a small retrospective cohort of pancreatic cancer patients.

2 Materials and Methods

The image processing for this project, was performed using in-house software CKImage v1.1. This was written in multithreaded Java (Oracle Corporation, Reading, UK) for maximum flexibility and platform-independence. The graphical user interface (GUI) for this software is shown in figure 1.

Although eventually, the goal is to use the software in a real-time context, at present it is operated in an offline retrospective approach. When a patient is loaded, the software reads the patient data file structure that is normally used for providing patient information to Accuray. The planning computed tomography (CT) scan and outline set are read and the PTV delineated on the planning CT scan is projected onto the DRRs in the image planes to define the region of interest. The DRRs themselves are created by the Precision software (Accuray Inc.) using bone suppression, where necessary, so as to remove confounding bony anatomy from the DRRs, leaving the soft tissue.

The acquired kV images are filtered using a band-pass cosine-shaped filter to increase the image sharpness without excessively accentuating the high-frequency noise, and the average image is subtracted from the individual images to remove stationary features such as bony anatomy. This is most effective when there are many images in the sequence being analyzed. A rectangular bounding box is defined which encloses the PTV projection with a margin of 30 mm, and the image intensities of the DRR and treatment images are equalized in this region. Rigid registration is then used to determine the offset of the actual pancreatic PTV projection from that defined on the DRR using a quadratic penalty function in the bounding box. An additional term in the penalty function increases with increasing mis-registration of the DRR and treatment images, thereby discouraging large, implausible shifts. At present, only shifts are used in this process, without rotations. Geometric considerations are then used to convert 2D shifts to 3D.

Computationally, the GUI class responds to user requests in the GUI and SwingWorker threads are used to accomplish more demanding tasks. For the tracking process, the application thread filters the images, using one thread per pair of images in the set of images to be processed. The registration process is carried out sequentially with respect to the set of images, but multiple threads are used to register each pair of images. The superoinferior position of each image pair is synchronized as a shift in anatomy in this direction is reflected equally in both images. These shifts are dealt with by multiple threads, and the independent orthogonal shifts, one for each image, are dealt with sequentially within those threads (figure 2).

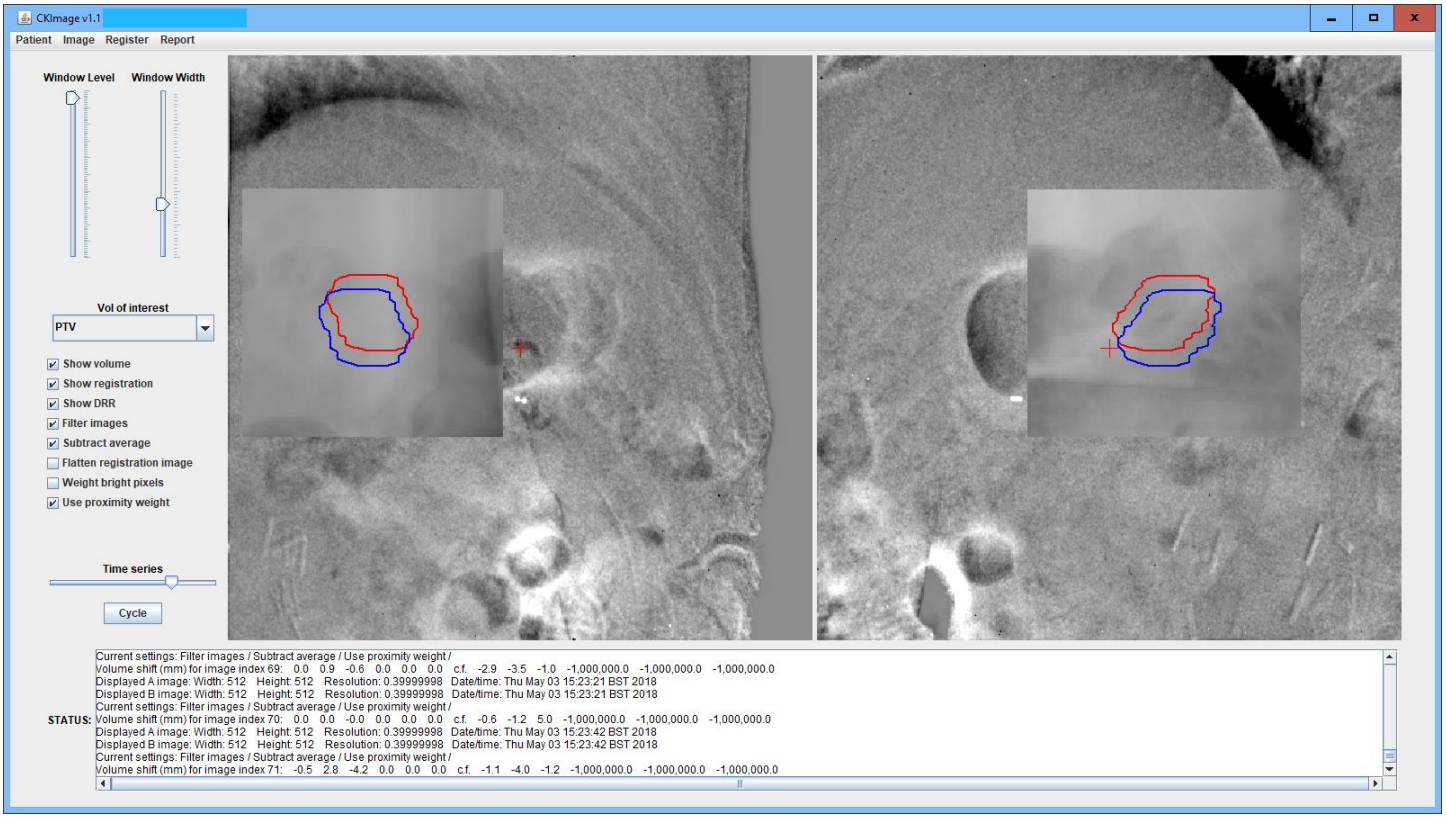


Figure 1. Graphical user interface for the CKImage software. The outlines of the planning target volume from the planning scan are shown in red and the corresponding matched regions in the treatment images are shown in blue. The digitally reconstructed radiographs are shown around the planning target volume. The results are reported in the status panel for each of the images in the time series as lateral, anteroposterior and superoinferior offsets of the imaged volume of interest with respect to the volume of interest projected onto the digitally reconstructed radiograph.

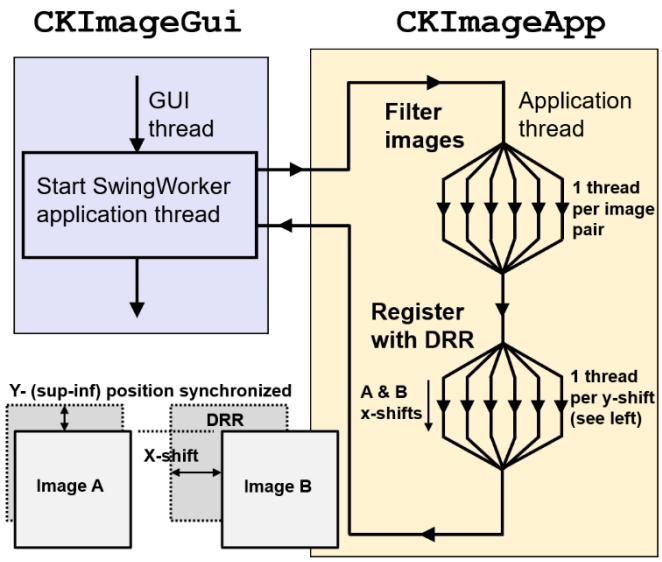


Figure 2. Thread structure of the CKImage software.

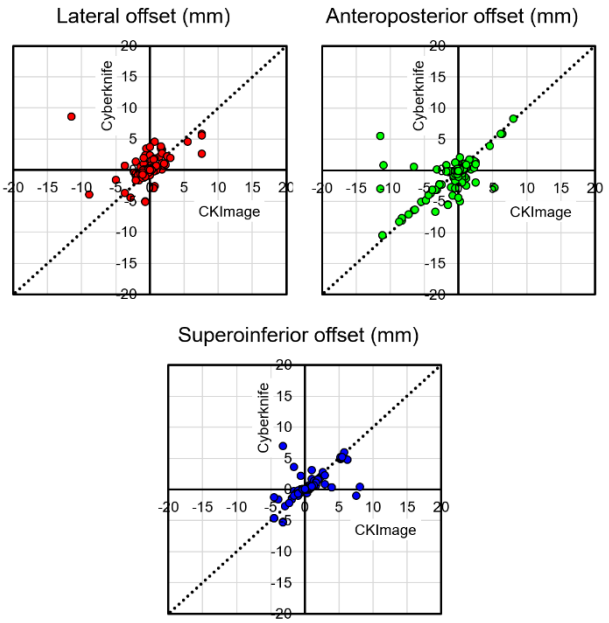


Figure 3. Comparison of positions recorded by CKImage and Cyberknife for the high-contrast test object. The 1:1 diagonal line is to guide the eye.

In this benchmarking study, the method was applied to 177 pairs of images acquired of a dose checker device, which formed a high-contrast test object. The results were compared with the rigid registration method ('6D skull') used by Cyberknife during positioning of the device. The method was also applied retrospectively to 1249 pairs of images acquired from four patients with pancreatic PTVs. As the patients were treated with fiducials, there were fiducials present in the treatment images, but the fiducials were removed from the DRRs by the bone suppression algorithm used, so this was not considered to influence the registration algorithm. The results using CKImage were compared with those produced at the time of treatment using a point-based fiducial registration method.

3 Results

The software takes 1-5 s to register each image in the time sequence using 16 processor threads, depending on the size of the planning target volume and the margin around it. Test results using the high-contrast test object are given in table 1 and the distribution of differences between CKImage and Cyberknife is shown in figure 3.

Table 1. Absolute differences between CKImage and Cyberknife for the high-contrast test object.

	Median (mm)	Min (mm)	Max (mm)
Lateral	0.4	0.0	20.1
Anteroposterior	0.3	0.0	17.0
Superoinferior	0.3	0.0	10.2
3D difference	0.8	0.0	21.8

The agreement for the four clinical patients between fiducial and fiducial-free tracking is shown in table 2 for the three cardinal directions. A histogram of the 3D differences is shown in figure 4.

Table 2. Absolute differences between CKImage (markerless tracking) and Cyberknife (fiducial tracking) for the four pancreatic target volumes.

	Median (mm)	Min (mm)	Max (mm)
Lateral	1.2	0.0	10.1
Anteroposterior	1.7	0.0	9.5
Superoinferior	2.4	0.0	15.3
3D difference	4.1	0.2	16.1

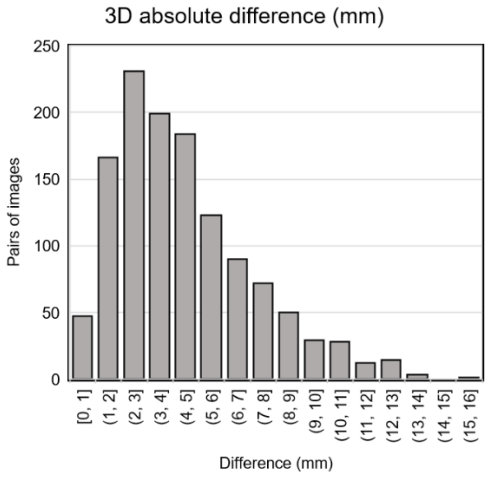


Figure 4. Histogram of absolute 3D differences between markerless (CKImage) and fiducial (Cyberknife) tracking for the four patients with pancreatic targets.

4 Discussion

This study uses image processing and careful image region matching to provide optimal conditions for accurate correlation of the DRRs and the treatment images. The results demonstrate that for a well-defined geometric target and for a small cohort of pancreas targets, it is possible to identify the position of the PTV in patient space from the treatment images relative to the DRRs with a median accuracy of several millimeters. For the well-defined target, the accuracy is better than in the pancreatic targets, where the target volume and surrounding anatomy is considerably less well defined on both the DRRs and treatment images.

This accuracy is promising but not as high as that shown by Nakao et al. [4]. In their study, an image generative network is used to map from 2D images, in the form of DRRs with randomly added setup shifts, to 3D deformation vectors. These vectors, together with vertices from 3D organ models are then input to a graph convolutional network to provide deformed 3D organ shapes. This approach allows multiple vertex points in the organ model, all projecting to the same 2D image point, to have different deformations. This is a much more sophisticated method than that of the present study, but it also uses DRRs to represent kV images, thereby permitting greater accuracy.

The present work compares markerless tracking with fiducial tracking at the time of treatment, but further work is needed to compare with a static method without tracking to ensure that the method is superior to an approach using an internal target volume. Further work is also desirable to define a more customized bounding region for the registration process. Use of rotations is thought to be a useful addition to the process, although rotations are not very prominent in this tumor site.

Deformable image registration may also be valuable for dealing with multiple organs in the pancreatic region. As an intermediate step, it may be useful to perform separate rigid registrations for several organs in the vicinity and then combine these into an overall map of organ displacement. Some measure of the dosimetric impact of the loss in accuracy of the markerless tracking method compared to tracking with fiducials is also desirable. The methods described in this study may also be beneficial for other abdominal tumor sites.

5 Conclusion

This study demonstrates the performance of new software for the tracking of pancreatic targets. The software currently operates retrospectively, but the multithreaded design is expected to give sufficiently high speed that it can be applied in the real time context. This study is limited to some preliminary comparisons, but using a small cohort of pancreas patients shows that the software is able to perform tracking without fiducials. The median accuracy appears to be in the order of 4 mm for pancreas. Some further developments are also proposed.

6 Acknowledgements

The authors acknowledge funding from the National Institute for Health Research (NIHR) Biomedical Research Centre at the Royal Marsden NHS Foundation Trust and the Institute of Cancer Research. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health. This research is also supported by Cancer Research UK under Program No. A28284, in collaboration with Accuray Inc. We are grateful to Richard Holloway, Philip Corrado and Calvin Maurer of Accuray for their assistance with this work. We are also grateful to Honorata Chajecka-Szczygielska and Dr Michael Thomas of The Royal Marsden NHS Foundation Trust for their support of this work.

References

[1] Asmerom G, Bourne D, Chappelow J, et al. “The design and physical characterization of a multileaf collimator for robotic radiosurgery”. Biomed Phys Eng Express 2.1 (2016), p. 017003. DOI: 10.1088/2057-1976/2/1/017003.

[2] D. Zhou, M. Nakamura, N. Mukumoto, et al. “Development of a deep learning-based patient-specific target contour prediction model for markerless tumor positioning”. Med Phys 49.3 (2022), pp. 1382-1390. DOI: 10.1002/mp.15456.

[3] M. Nakamura, M. Nakao, N. Mukumoto, et al. “Statistical shape model-based planning organ-at-risk volume: application to pancreatic cancer patients”. Phys Med Biol 66.1 (2021), p. 014001. DOI: 10.1088/1361-6560/abcd1b.

[4] M. Nakao, M. Nakamura, T. Matsuda. “Image-to-graph convolutional network for 2D/3D deformable model registration of low-contrast organs”. IEEE Trans Med Imag 41.12 (2022), pp. 3747-3761. DOI: 10.1109/TMI.2022.3194517.

[5] W. Hall, C. Small, E. Paulson, et al. “Magnetic resonance guided radiation therapy for pancreatic adenocarcinoma, advantages, challenges, current approaches, and future directions”. Front Oncol 11 (2021), p. 628155. DOI: 10.3389/fonc.2021.628155.

[6] B. Stemkens, R. Tijssen, B. de Senneville, et al. “Optimizing 4-dimensional magnetic resonance imaging data sampling for respiratory motion analysis of pancreatic tumors”. Int J Radiat Oncol Biol Phys 91.3 (2015), pp. 571-578. DOI: 10.1016/j.ijrobp.2014.10.050.

[7] G. Grimbergen, H. Eijkelenkamp, H. Heerkens, et al. “Intrafraction pancreatic tumor motion patterns during ungated magnetic resonance guided radiotherapy with an abdominal corset”. Phys Imaging Radiat Oncol 21.12 (2022), pp. 1-5. DOI: 10.1016/j.phro.2021.12.001.

The impact of training dataset size and model size on the accuracy of deep learning dose prediction

Joep van Genderingen[1], Dan Nguyen[2], Franziska Knuth[1], Hazem Nomer[1], Luca Incrocci[1], Abdul Sharfo[1], Uwe Oelfke[3], Steve Jiang[2], Linda Rossi[1], Ben Heijmen[1], and Sebastiaan Breedveld[1]

[1]Erasmus MC Cancer Institute, University Medical Center Rotterdam, Department of Radiotherapy, Rotterdam, The Netherlands.
[2]UT Southwestern Medical Center, Medical Artificial Intelligence and Automation (MAIA) Laboratory, Department of Radiation Oncology, Dallas, USA.
[3]The Institute of Cancer Research and The Royal Marsden NHS Foundation Trust, Joint Department of Physics, London, United Kingdom.

Abstract Deep learning has demonstrated to be capable of predicting realistic radiotherapy dose distributions. Collecting and curating clinical data is a labour-intensive and time-consuming task, making large, curated datasets scarce for training deep learning dose prediction models. Partly due to the lack of training data, model size has never been fully explored in previous studies. In this study, we investigated how training dataset size and model size impact 3D dose prediction performance for prostate cancer. For a curated dataset consisting of 1250 contoured planning CT-scans, ground truth dose distributions were fully automatically generated for each patient using Erasmus-iCycle, guaranteeing consistent trade-offs. Three HD U-Nets with two, four and six pooling layers were trained from scratch with different sized datasets (50/250/1000 plans) for prediction of dose distributions. Mixed precision was applied to optimize GPU VRAM usage. Prediction accuracy was lowest for the 2-level model, while differences between 4-levels and 6-levels were small. For all model sizes, agreement improved with increasing training dataset size, suggesting that more training patients could further increase prediction accuracy.

1 Introduction

Treatment planning for RT is a fragmented, labor-intensive and time-consuming collaboration between physicians, medical physicist and dosimetrists. There is a growing need for shorter treatment planning times, as cancer incidences and staff shortages are increasing [1], while the RT field is also moving towards (online) adaptive treatment strategies and same-day treatment [2]. Although multi-criterial optimization based automated treatment planning as applied in our center has shown to drastically reduce hands-on treatment planning workload in the clinic while producing high quality plans [3, 4], substantial computational time is required. Deep learning has been proposed to predict the patient-specific 3D dose in the order of seconds [5, 6, 7, 8]. The predicted dose can either be used as a directive for the physician [9] or converted to machine parameters to obtain a deliverable treatment plan [10, 11]. From previous studies it is expected that the dosimetric performance of deep learning dose prediction depends on the size and quality of the dataset used for training [12, 13]. However, no consistent dataset size study has been performed to assess whether larger datasets could improve prediction accuracy. In addition, model size has never been fully explored in previous works.

In this study, we systematically investigated the impact of training dataset size and model size on 3D dose prediction performance for prostate cancer, using a curated dataset of 1250 contoured planning CT-scans. All ground truth (GT) plans used for the analyses were generated with Erasmus-iCycle [14], a system for fully automated multi-criterial plan generation, guaranteeing consistent trade-offs between all treatment objectives. Many studies have demonstrated superiority of Erasmus-iCycle plans compared to manual planning [3, 4, 14, 15, 16, 17].

2 Materials and Methods

2.1 Patient Cohort

Patients with low, intermediate or high-risk prostate cancer without metal hip prostheses were included in this study. The data was combined from two sources: 719 patients were included from the multicenter phase 3 randomized trial on hypofractionation for prostate cancer (HYPRO) [15, 18], which was complemented with 531 recent patients from the Erasmus University Medical Center Rotterdam.
Erasmus-iCycle was used to automatically generate for each patient a 10 MV 23-beam GT IMRT plan, mimicking VMAT dose distributions, using original CT resolutions. The clinically established and applied delineations of the target volumes, rectum, anus, femoral heads, bladder and body were also used in this study. The patient cohort consisted of three groups, based on the risk of involvement of the seminal vesicles. For Group 1 patients, a dose of 78 Gy (PTV high) was prescribed to the prostate. For Group 2, a dose of 78 Gy was prescribed to the prostate and 72.2 Gy (PTV low) to the seminal vesicles. In Group 3 patients, prescribed dose for both prostate and seminal vesicles was 78 Gy.

2.2 Experimental setup for Deep Learning

The CT scan, binary masks of the structure delineations and GT dose distributions were downsampled to a grid size of 256 x 256 x 128, reflecting a patient resolution 1.95 x 1.52 x (1.5-3) mm^3, where the original slice distance was conserved. The data was divided into 1000 training, 100 validation, and 150 test patients using stratified sampling. Stratification was based on the patient Groups (1-3), size of the patient in the isocenter slice, and the volume of the high

dose target. For training, target volumes were combined into one target map where each voxel value was set to the corresponding prescription dose. Generated models had seven input channels: the CT scan, target map, and binary masks of the anus, bladder, rectum, combined femoral heads and patient contour. The Hounsfield units in the CT were clipped to a range of [-1000, 1500] to improve training convergence, following Liu et al.[19]. All model inputs were left unscaled for training.

The dose prediction models for the investigated dataset sizes and model sizes were based on the Hierarchically Densely Connected U-Net (HD U-Net) architecture proposed by Nguyen et al. [20]. We extended the HD U-Net with a mixed precision learning strategy to optimize VRAM usage. Three HD U-Nets with two, four and six pooling layers were separately trained on randomly stratified subsets of the training data, N=50/250/1000 (5/25/100%), resulting in nine trained models. All model training was performed with the Adam optimizer with a learning rate of 10^{-4}, batch size of 1 and mean squared error (MSE) as the loss function. We trained each experiment with four different patience parameters (5, 10, 15 and 20) for the reduce on plateau learning rate scheduler, and early stopping with a threshold of 25 epochs was used to reduce overfitting. The model with the lowest validation loss was used to assess the test data. Training of all models was performed on a cluster with 15 NVIDIA A40 GPU cards with 48 GiB of VRAM.

Predicted dose distributions were interpolated back to their original CT resolutions. All models were evaluated on the same test set without normalization of doses. The performance of the models was evaluated using dose metrics, spatial dose distributions, mean absolute error (MAE) and MSE. To evaluate target dose, we computed $V_{95\%}$ for PTV high and PTV low, and $V_{107\%}$ for PTV high for the predicted dose distributions and GT. For the rectum, we computed V_{75Gy}, V_{60Gy} and Dmean. The bladder was evaluated for V_{65Gy} and Dmean, and anus for Dmean. The near maximum dose, $D_{0.03cc}$, was used to evaluate dose delivered to the femoral heads, and V_{20Gy} was computed for the patient contour to evaluate the prediction quality in the low to intermediate dose regions. Statistical significance of differences between predicted and GT parameters was assessed with two-sided Wilcoxon signed rank tests, using $p<0.05$ for significance.

3 Results

Figure 1 shows predicted and GT dose distributions for an axial slice of a random patient from the test set. Where the 2-level model with 50 training patients only gives an impression of the dose distribution, details start showing up with 4-levels and 250 training patients, with highest correspondence to GT for the 6-level model and 1000 training patients. For the 4-level and 6-level models, the accuracy of the predicted dose distributions clearly increased with more training data, especially regarding the distinct spikes in the intermediate dose regions (20-30 Gy). In figure 2, the population MAE and MSE show the same trend. For all model sizes, a reduction in both metrics was found with increasing training dataset size, indicating that more plans are needed for training to reach convergence. For each training dataset size, the 4- and 6-level models clearly outperformed the 2-level model in both MAE and MSE, while differences between the 4-level and 6-level models were minor for all dataset sizes.

Figure 3 shows box-and-whisker plots for the selected dose metrics. Overall, all models showed good performance for targets and OARs. For PTV high $V_{95\%}$, PTV low $V_{95\%}$, rectum V_{75Gy}, rectum V_{60Gy}, bladder V_{65Gy}, bladder Dmean, anus Dmean and patient V20, increases in agreement between predictions and GT were observed with increasing training dataset size. For fixed dataset sizes, no noticeable differences were observed between model sizes, except for bladder Dmean and V_{20}, where the 2-level models had larger errors. The 6-level model trained on 1000 patients had overall the smallest dose metrics differences between predictions and GT.

4 Discussion

The purpose of this study was to systematically investigate the effect of training dataset size and model size on deep learning 3D dose prediction accuracy for prostate cancer, based on a large curated dataset of contoured planning CT-scans and corresponding highly consistent, automatically generated GT plans.

The 3D dose distributions showed a consistent trend in dosimetric detail when using more data and/or deeper models (Fig. 1). The MAE and MSE (Fig. 2) were clearly highest for the 2-level model, while differences between 4-levels and 6-levels were small. For all model sizes, MSE and MAE went down with increasing training dataset size, suggesting that more patients could further increase prediction accuracy. Differences in the 20-30 Gy areas (spikes in Fig. 1) contributed substantially to the found MSE and MAE differences. A study by Kandalan et al. [21] also showed issues with prediction accuracy in these dose regions. Probably, there is more dosimetric variation between patients for intermediate and low doses.

Contrary, for the evaluated (high dose) plan parameters, there was overall no clear dependency of prediction accuracy on model size, while predictions based on 1000 training patients were slightly more accurate than those for 250 or 50 training patients (Fig. 3).

Given the observations, evaluation of deep learning dose prediction accuracy based only on target and OAR plan parameters appears insufficient for a full assessment.

This study was based on a large dataset with highly consistent, automatically generated GT plans. It remains to be investigated how the conclusions from this work will translate to datasets with clinically/manually generated plans. Due to well-known variability in quality and trade-offs in manual plans, we do expect that convergence will already be reached with smaller training datasets, and differences between model sizes will become smaller.
This study has some limitations. We have only investigated model size in terms of model depth. U-Net based architectures are capable of capturing global information. That capability is influenced by the number of trainable parameters. The number of trainable parameters of a model can be increased in more ways than only in depth. Further research is needed to investigate if model width can compensate for a lack of depth. This study investigated prostate dose prediction. In a next study, we will investigate dataset size and model size for head-and-neck cancer, with more anatomical variability and more complex treatment plans than the prostate site. Currently, data is gathered to create a large size head-and-neck database.

5 Conclusions

Using a large set of fully-automatically generated prostate cancer plans with consistent trade-offs, increasing the training dataset size from 50 to 250 to 1000 patients resulted for (high dose) PTV and OAR plan parameters in relatively small reductions in differences between deep learning predictions and GT. More substantial differences were found for dose spikes in the 20-30 Gy dose range around the PTV. For all model sizes, MAE and MSE decreased with dataset size, possibly indicating that further enhancing the training dataset size could further increase prediction accuracy.

References

[1] ECIS, E. C. I. S.-. 2022. Estimated incidence by year - summary [Online]. Available: https://ecis.jrc.ec.europa.eu/explorer.php?$0-4$1-AE28E$4-1,2$3-0$6-0,85$5-All$77$21-0$2-All$CLongtermChart3_1$X0_14- $CLongtermChart3_2$X1_14- [Accessed Jan 26th 2024].

[2] N. Hoffmans-Holtzer, L. Smolenaers, R. Peeters, et al. “Robust scheduling for a One Stop Shop palliative radiotherapy clinic using genetic algorithms”. Radiotherapy and Oncology 170 (2022), S874–S876

[3] L. Rossi, S. Breedveld, S. Aluwini, et al. “Noncoplanar beam angle class solutions to replace time-consuming patient-specific beam angle optimization in robotic prostate stereotactic body radiation therapy”. 92.4 (2015), pp. 762–770.

[4] P. W. Voet, M. L. Dirkx, S. Breedveld, et al. “Toward fully automated multicriterial plan generation: a prospective clinical study”. International Journal of Radiation Oncology* Biology* Physics 85.3 (2013), pp. 866–872.

[5] V. Kearney, J. W. Chan, S. Haaf, et al. “DoseNet: a volumetric dose prediction algorithm using 3D fully-convolutional neural networks”. Physics in Medicine & Biology 63.23 (2018), p. 235022.

[6] A. Babier, R. Mahmood, A. L. McNiven, et al. “Knowledge-based automated planning with three-dimensional generative adversarial networks”. Medical Physics 47.2 (2020), pp. 297–306.

[7] V. Kearney, J. W. Chan, T. Wang, et al. “DoseGAN: a generative adversarial network for synthetic dose prediction using attention-gated discrimination and generation”. Scientific reports 10.1 (2020), p. 11073.

[8] B. Zhan, J. Xiao, C. Cao, et al. “Multi-constraint generative adversarial network for dose prediction in radiotherapy”. Medical Image Analysis 77 (2022), p. 102339.

[9] M. Mashayekhi, R. McBeth, D. Nguyen, et al. “Artificial intelligence guided physician directive improves head and neck planning quality and practice Uniformity: A prospective study”. Clinical and Translational Radiation Oncology 40 (2023).

[10] A. Babier, B. Zhang, R. Mahmood, et al. “OpenKBP: the open-access knowledge-based planning grand challenge and dataset”. Medical Physics 48.9 (2021), pp. 5549–5561.

[11] A. Babier, R. Mahmood, A. L. McNiven, et al. “The importance of evaluating the complete automated knowledge-based planning pipeline”. Physica Medica 72 (2020), pp. 73–79.

[12] A. M. Barragán-Montero, M. Thomas, G. Defraene, et al. “Deep learning dose prediction for IMRT of esophageal cancer: the effect of data quality and quantity on model performance”. Physica Medica 83 (2021), pp. 52–63.

[13] D. Nguyen, M.- H. Lin, D. Sher, et al. “Advances in Automated Treatment Planning”. Seminars in Radiation Oncology. Vol. 32. 4. Elsevier. 2022, pp. 343–350.

[14] S. Breedveld, P. R. Storchi, P. W. Voet, et al. “iCycle: Integrated, multicriterial beam angle, and profile optimization for generation of coplanar and noncoplanar IMRT plans”. Medical physics 39.2 (2012), pp. 951–963.

[15] A. W. M. Sharfo, M. L. Dirkx, R. G. Bijman, et al. “Late toxicity in the randomized multicenter HYPRO trial for prostate cancer analyzed with automated treatment planning”. Radiotherapy and Oncology 128.2 (2018), pp. 349–356.

[16] S. Breedveld, P. R. Storchi, and B. J. Heijmen. “The equivalence of multi-criteria methods for radiotherapy plan optimization”. Physics in Medicine & Biology 54.23 (2009), p. 7199.

[17] B. Heijmen, P. Voet, D. Fransen, et al. “Fully automated, multi-criterial planning for Volumetric Modulated Arc Therapy–An international multi-center validation for prostate cancer”. Radiotherapy and Oncology 128.2 (2018), pp. 343–348.

[18] S. Aluwini, F. Pos, E. Schimmel, et al. “Hypofractionated versus conventionally fractionated radiotherapy for patients with prostate cancer (HYPRO): acute toxicity results from a randomised non-inferiority phase 3 trial”. The Lancet Oncology 16.3 (2015), pp. 274–283.

[19] S. Liu, J. Zhang, T. Li, et al. “A cascade 3D U-Net for dose prediction in radiotherapy”. Medical Physics 48.9 (2021), pp. 5574–5582.

[20] D. Nguyen, X. Jia, D. Sher, et al. “3D radiotherapy dose prediction on head and neck cancer patients with a hierarchically densely connected U-net deep learning architecture”. Physics in medicine & Biology 64.6 (2019), p. 065020.

[21] R. N. Kandalan, D. Nguyen, N. H. Rezaeian, et al. “Dose prediction with deep learning for prostate cancer radiation therapy: model adaptation to different treatment planning practices”. Radiotherapy and Oncology 153 (2020), pp. 228–235.

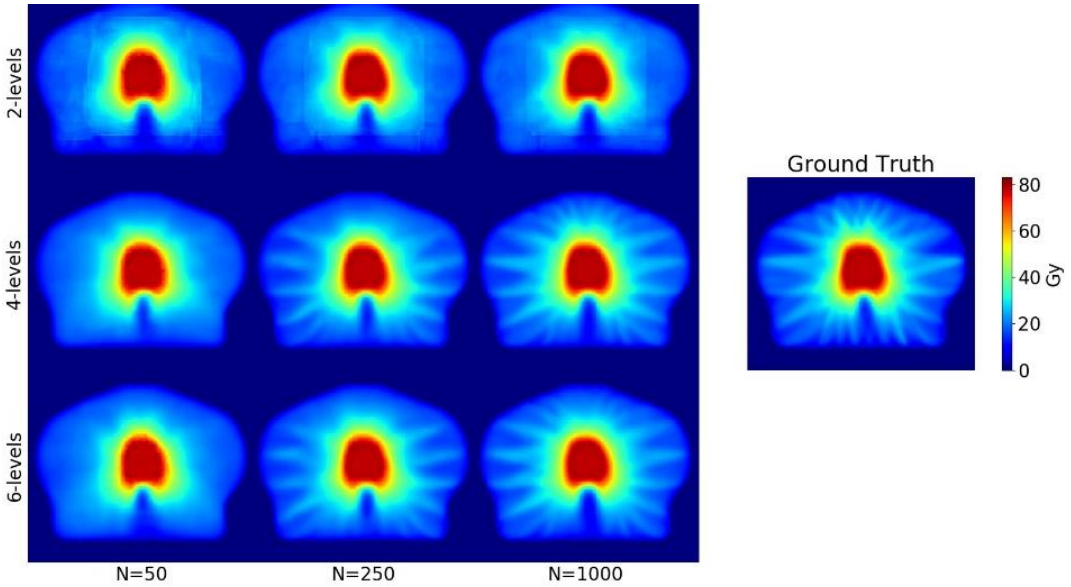


Figure 1: Axial slice of the predicted 3D dose distributions of a randomly selected patient from the test set (left) compared to ground truth (right).

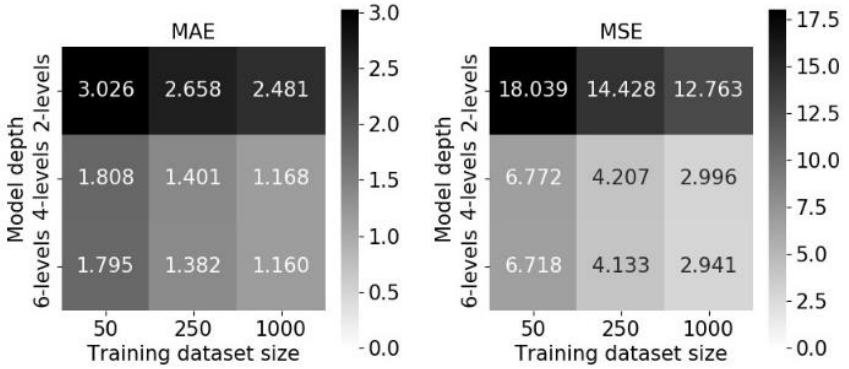


Figure 2: MAE and MSE for the test set. MAE and MSE were computed over voxels with a dose higher than 1 Gy.

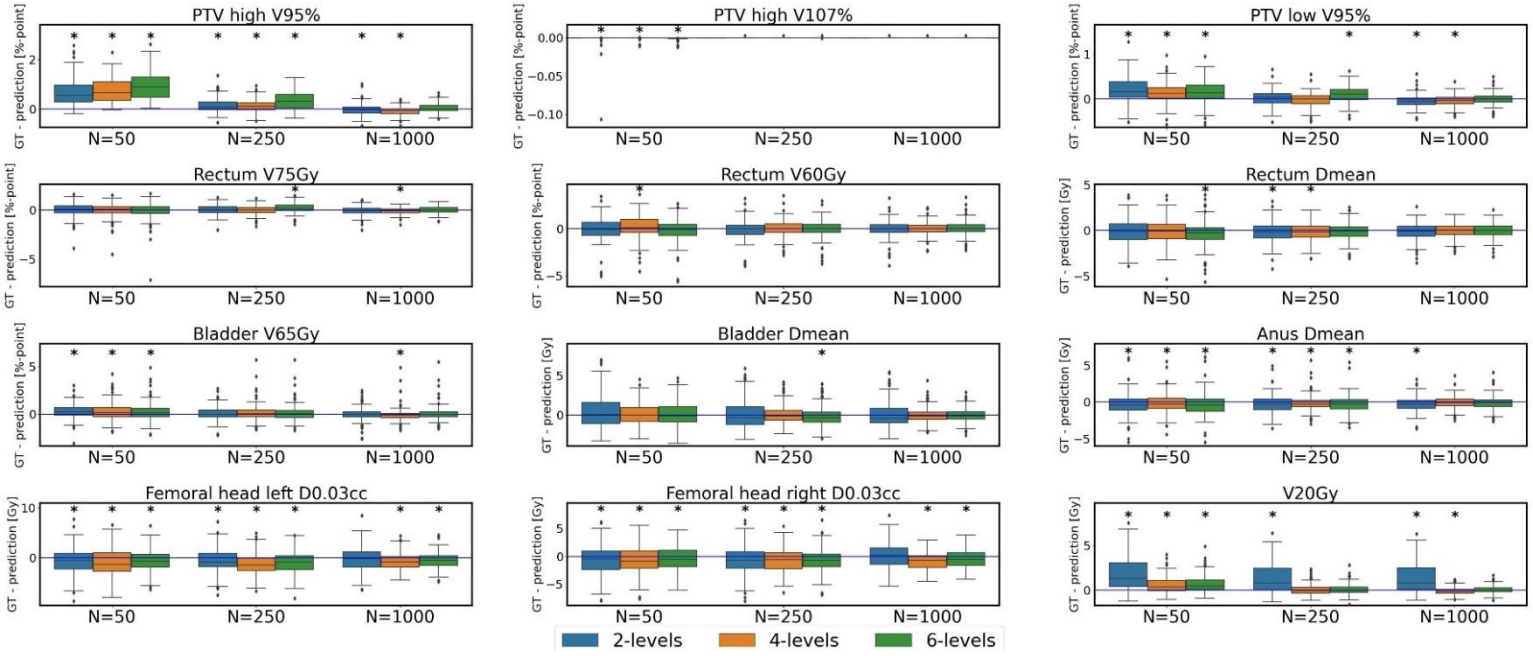


*Figure 3: Box-and-whisker plots of the differences between the predicted and ground truth (GT) dose metrics. * denotes statistical significance difference to GT as a result of Wilcoxon signed rank test ($p<0.05$).*

Deep learning for instantaneous estimation of the out-of-field dose in external beam photon radiotherapy – A proof of concept

Nathan Benzazon[1, 2], Mohammed El Aichi[1, 2], Alexandre Carré[1, 2], François de Kermenguy[1, 2], Stéphane Niyoteka[1, 2], Pauline Maury[1, 2], Julie Colnot[1, 2, 3], Meissane M'hamdi[1, 2], Cristina Veres[1, 2], Rodrigue Allodji[4], Florent de Vathaire[4], David Sarrut[5], Neige Journy[4], Claire Alapetite[6], Vincent Grégoire[7], Eric Deutsch[1, 2], Ibrahima Diallo[1, 2] and Charlotte Robert[1, 2]

[1]Unité Mixte de Recherche 1030, ImmunoRadAI, Université Paris-Saclay, Institut Gustave Roussy, Inserm, Villejuif, France
[2]Department of Radiation Oncology, Gustave Roussy, Villejuif, France
[3]THERYQ, PMB-Alcen, Peynier, France
[4]Unité Mixte de Recherche 1018, Radiation Epidemiology Team, Université Paris-Saclay, Institut Gustave Roussy, Inserm, Villejuif, France
[5]Université de Lyon; CREATIS; CNRS UMR5220; Inserm U1294; INSA-Lyon; Université Lyon 1, Léon Bérard cancer center, Lyon, France.
[6]Department of Radiotherapy, Institut Curie, Paris, France.
[7]Department of Radiation Oncology, Centre Léon-Bérard, Lyon, France.

Abstract

Doses deposited outside the treatment field during external photon-beam radiotherapy, known as out-of-field doses, appear to favor the appearance of radiation-induced cancer and hematological toxicities. These low doses are not currently estimated on a routine clinical basis, due to the lack of a suitable solution.
With this proof of concept, we demonstrate the value of deep learning for the development of a tool enabling rapid, routine-compatible out-of-field dose estimation. For this purpose, a 3D U-Net, considering as inputs the in-field dose, as computed by the treatment planning system, and the patient's anatomy, was trained to predict out-of-field dose maps, using a dataset of 3151 pediatric whole-body dose maps estimated with analytical methods as ground truth. Consistent results in line with the literature were obtained (RMSD of 0.28, 0.41 and 0.32 $cGy.Gy^{-1}$ for the training, validation and testing), arguing in favor of the use of neural networks for out-of-field dose estimation.

1 Introduction

In external beam radiotherapy (EBRT) with photons, non-zero doses are inevitably delivered outside the treatment field; often referred to as out-of-field or peripheral dose. The out-of-field doses are mostly low dose values (< 4 Gy), but the question of their potential impact on the probability of radiation-induced adverse events, such second cancers (1) and immunological dysfunctions like radiation-induced lymphopenia (2) is of the utmost importance. In particular, it has recently been shown that even very low doses resulting from computed tomography (CT) scans contribute to the development of radiation-induced cancers (3). In radiotherapy, this topic is currently experiencing renewed interest, particularly as modulated treatments (such as volumetric modulated arc therapy (VMAT) or intensity-modulated radiation therapy (IMRT)) are now routinely used in clinical care and are known to be associated with higher peripheral doses due to longer beam on times and larger irradiated volumes when compared to three-dimensional conformal radiotherapy (3D-CRT) (3,4)

Treatment planning systems (TPS) are used in clinical routine to estimate in-field dose distribution, but have been shown to systematically underestimate out-of-field dose for 3D conformational radiotherapy treatments, for intensity modulated treatments and for CyberKnife devices (Accuray, Sunnyvale, USA) (5–8). Thus, despite clear clinical potential, the out-of-field dose computation is currently not available in clinical practice. Two methods are currently used in the literature for out-of-field dose estimation for research purposes: Monte Carlo simulations and analytical approaches. These methods can provide accurate out-of-field dose estimation (7) but are inappropriate for clinical routine implementation, because of their lack of versatility and their cumbersome application (9).

Artificial intelligence and in particular technologies based on deep learning (DL) have drastically changed clinical practice in radiotherapy in recent years, allowing notably the automation of several time-consuming tasks, including segmentation and treatment planning, or the generation of virtual images (synthetic images), demonstrating their ability to identify complex hierarchical features from spatially structured data (10). Three main components contribute to the out-of-field dose: patient scatter, head scatter and head leakage. Apart from leakage, which is a signature specific to each accelerator and whose signal is minimal in the irradiation field, the other components exhibit a valuable in-field signal for assessing the out-of-field dose. Based on the assumption that the in-field dose map associated with the patient's anatomy contains most of the information needed to calculate the out-of-field dose, our objective was to evaluate the ability of a unique deep learning-based out-of-field dose estimation algorithm to adapt to a wide variety of configurations, bringing an answer to the problems of computation time, information extraction, and versatility, making it compatible with clinical use.

2 Materials and Methods

Three thousand one hundred fifty-one whole-body dose maps, from the French childhood cancer survivor study (FCCSS) cohort were used as ground truth for the development of our neural network. With the goal to develop a deep learning model for out-of-field dose estimation suitable for megavoltage photon irradiations, we only selected patients treated with a photon beam and using linear accelerators with a high voltage (> 1 MV) or ^{60}Co irradiators. Whole-body dose maps were estimated using an analytical method originally developed for epidemiological purposes (11,12). Twenty-five irradiation devices were represented in our dataset and were grouped into 3 categories: standard linacs, cobalt units and betatron units. Our dataset included 38 different pathologies, of which the most represented were nephroblastoma and other nonepithelial renal tumors (695 patients), Hodgkin lymphomas (449 patients) and astrocytomas (235 patients).

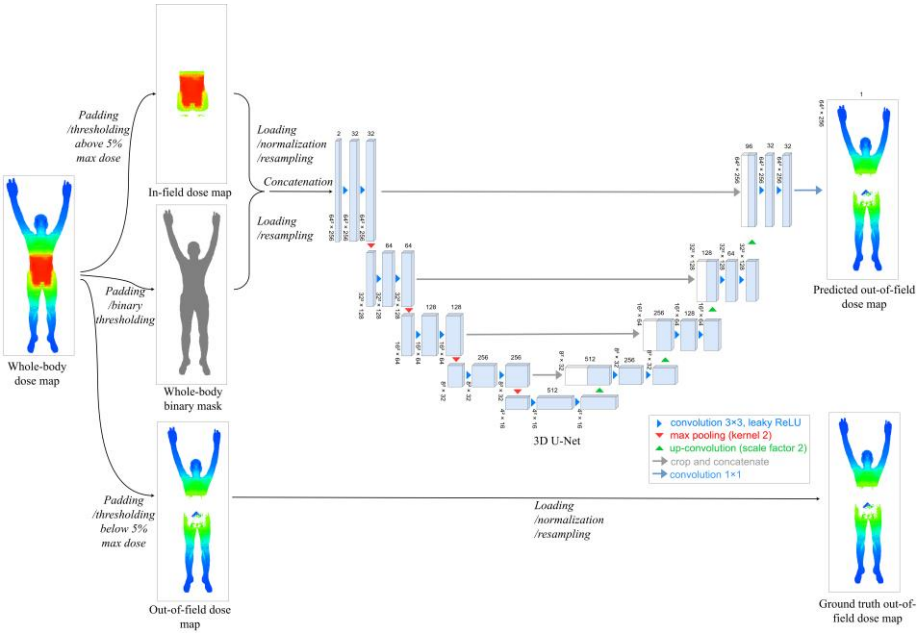


Figure 1 Dose map preprocessing pipeline and comprehensive design of the implemented 3D U-Net. Dose maps are displayed using logarithmic scale

Several preprocessing steps were applied to the analytical dose maps to make them deep-learning compliant. These included: 1) padding into [370, 242, 1131] matrix sizes, 2) resampling of padded files (originally in Statistical Analysis System (SAS) format) to [128, 128, 512] sizes, 3) extraction of in-field and out-of-field dose maps from whole-body dose maps (a 5% isodose threshold was chosen in this goal considering the maximum dose per patient as the reference dose), 4) creation of binary masks from the whole-body dose maps by a thresholding method separating the background from the foreground. An on-the-fly preprocessing pipeline was then applied using Medical Open Network for Artificial Intelligence (MONAI 0.8.0) (13), including in the following order: loading, normalization, resampling, and concatenation. During the normalization step, 3D dose maps intensities were normalized by dividing by 100 Gy, to provide the neural network with values within [0,1]. The on-the-fly resampling step was implemented to test the impact of batch size as a function of available VRAM (video random-access memory). A nearest neighbor interpolation strategy was used for the resampling step. Finally, we transformed the data to fit a multi-dimensional structure denoted as B×2×H×W×D. In this expression, B refers to the batch size used during the training phase of the process. C indicates the number of channels present, while H, W, and D correspond to the height, width, and depth of the matrix, respectively. Figure 1 summarizes the preprocessing pipeline.

A conventional 3D U-Net (14,15), composed of four down-sampling blocks followed by four up-sampling blocks, was implemented (Figure 1). The Mean Square Error (MSE) evaluated only on the foreground voxels outside the 5% isodose, i.e. only in the region considered as the out-of-field dose, was selected as loss function. Learning rate and weight decay were considered in the ranges $[1e-7, 1e-3]$ and $[1e-8, 1e-4]$, respectively, with 1e-4 and 1e-6 providing the optimal results. The Adam optimizer was used to update the network parameters. Instance normalization was preferred. Batch size of 20 corresponding to a resampled size of $64 \times 64 \times 256$ was selected. The 3D U-Net was trained for 600 epochs (~50 hours) on a Nvidia RTX A6000. No early stopping was used. Weight based on best performance on validation loss were saved.

The dataset of 3151 patients was conventionally split into training (N = 2213), validation (N = 505), and test cohorts (N = 433). The test set was divided into 5 sub-cohorts, including two classic linear accelerators (names of accelerator model unavailable) operating at 6 MV and 16 MV, two cobalt 60 units (Alcyon and Mobiletron), and finally a betatron operating at 1.25 MV. All patients treated with betatron devices were voluntarily kept into the test set, because of its very specific design compared with a conventional linear accelerator or a cobalt unit. None of the devices in the test set were included in the training and validation sets. This was made in order to test the generalizability of the trained network on unknown devices.

We finally performed 2 additional experiments. The first aimed at evaluating if the inclusion of betatron-treated patients in the training set could improve performance for a subgroup of 46 patients here considered as a new sub-test set. The second additional experiment aimed to study the

benefits of a custom model, compared with the use of a generalized model. In this aim, we selected 100 of the 128 patients treated with a ^{60}Co-Alcyon, and trained the 3D U-Net from scratch considering only these data. Finally, as a last experiment, we fine-tuned the original model using the 100 ^{60}Co-Alcyon training patients for 2000 epochs and tested the performance of this fine-tuned model on the 28 ^{60}Co-Alcyon patients included in the test set. We used the root mean square deviation (RMSD) (corresponding to the root of the MSE) as performance evaluation metric.

3 Results

Best performances were achieved at epoch 467/600. For visual purposes, Figure 2 shows three examples of out-of-field dose maps generated by our trained network from test sets for patients with good (RMSD of 0.16 cGy.Gy^{-1}), median (RMSD of 0.29 cGy.Gy^{-1}) and poor performance (RMSD of 1.00 cGy.Gy^{-1}). RMSD of 0.28 ± 0.08 and 0.41 ± 0.26 cGy.Gy^{-1} were obtained for the training and validation datasets, respectively. Values of 0.27 ± 0.06, 0.26 ± 0.07, 0.28 ± 0.06, 0.30 ± 0.12 and 0.45 ± 0.25 cGy.Gy^{-1} were achieved for the 6 MV linac, 16 MV linac, ^{60}Co-Alcyon, ^{60}Co-Mobiletron, and betatron devices test sets, respectively, demonstrating overall performance similar to or better than that of the validation set, except for the fifth test set, corresponding to the betatron device. A RMSD threshold value of 0.6 cGy.Gy^{-1} was considered to separate good from poor out-of-field dose reconstructions. 87 out of 505 patients showed weaker performances in the validation set; 85 of whom being treated with a single device: a Sagittaire linear accelerator operating at 25 MV. This value was equal to 24 (out of 433 patients) in the test set, with 21 of the 24 patients identified having been treated with the betatron accelerator.

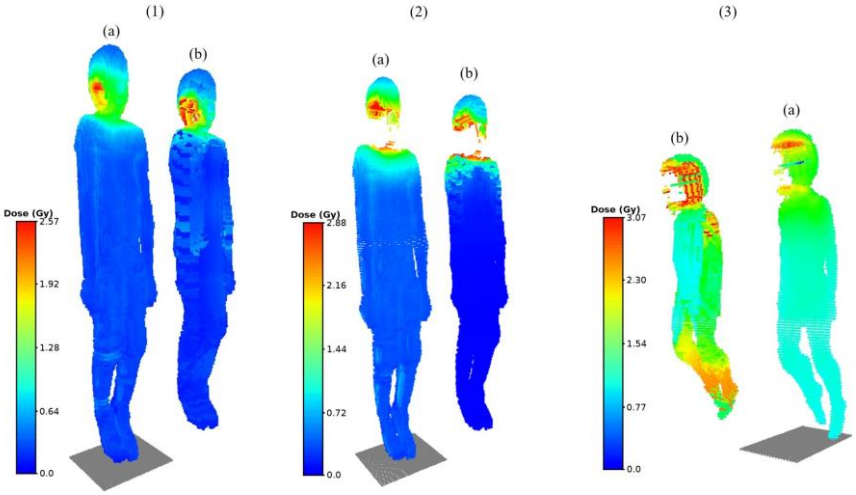


Figure 2 Comparison of out-of-field dose maps predicted by the 3D U-Net (a) and computed by the analytical model, the ground truth (b), for 3 patients. Patient 1 (RMSD=0.16 cGy.Gy^{-1}, into the 5% percentile) corresponds to a 15-year-old girl diagnosed for a primary pathology of retinoblastoma in 1982 and treated with a Neptune 6 MV device. Patient 2 (RMSD=0.29 cGy.Gy^{-1}, in ± 1% of median) is also a 15-year-old girl who was diagnosed and treated in 1982 for an intracranial neoplasm using an Alcyon 60Co device. Patient 3 (RMSD=1.00 cGy.Gy^{-1}, below 95% percentile) is a 5-year-old girl diagnosed in 1982 treated the same year with a 1.25 MV betatron device for an

Retraining the neural network using the original patient training set plus the 46 betatron patients kept aside resulted in a mean RMSD of 0.64 ± 0.41 cGy.Gy^{-1}, to be compared with 0.26 ± 0.11 cGy.Gy^{-1} for the same 46 patients as a benchmark value (no betatron patients included at the training stage). For the second additional experiment, mean RMSDs of 0.16 ± 0.01 cGy.Gy^{-1}, 0.53 ± 0.22 cGy.Gy^{-1} and 0.13 ± 0.05 cGy.Gy^{-1} were obtained for the benchmark configuration, configuration considering training with only the 100 ^{60}Co-Alcyon training set patients, and configuration considering the original model fine-tuned with the 100 ^{60}Co-Alcyon training set patients, respectively.

4 Discussion

The results are rather encouraging, with RMSD values of the same order of magnitude in the test sets (mean value of 0.32 ± 0.15 cGy.Gy^{-1} in the test set) as in the validation set (0.41 ± 0.26 cGy.Gy^{-1}) or the training set (0.28 ± 0.08 cGy.Gy^{-1}), suggesting that the neural network has acquired a strong degree of robustness and generalization. However, we notice that significantly lower results were observed for test set corresponding to the betatron accelerator. In addition, most of the poorest results in the validation set concerned patients treated with the Sagittaire accelerator operating at 25 MV (the highest voltage of the linear accelerators considered in the training set was equal to 20 MV). These results logically highlight the fact that the generalization capabilities of the neural network cannot be extended to out-of-distribution cases including non-conventional or highly atypical linear accelerators, i.e. with different spectral characteristics, shielding properties, or head geometries. Another limitation of our proof of concept is the tendency by the neural network to systematically overestimated doses at locations farthest from the field. This discrepancy appears to stem from our choice of cost function (MSE), which minimizes absolute dose errors. Consequently, the greatest relative deviations are observed in these distant regions, where the lowest dose are expected.

The first additional experiment shows that adding betatron patients to the training set resulted in an unexpected decline in performance (0.26 ± 0.11 cGy.Gy^{-1} for the benchmark compared to 0.63 ± 0.41 cGy.Gy^{-1} when considering betatron patients in the training set). We hypothesize that the number of betatron cases in the training set is too small to allow the neural network to converge towards a solution tailored to this accelerator, and that this accelerator may be too unusual and confuses the network. The results obtained from the second additional experiment demonstrated that the model developed from 100 ^{60}Co-Alcyon patients showed poorer averaged performance on the 28 ^{60}Co-Alcyon test patients than the original model, trained

without any ^{60}Co-Alcyon data, but on more than 2000 patients (0.53 ± 0.22 cGy.Gy^{-1} for the custom model versus 0.16 ± 0.01 cGy.Gy^{-1} for the benchmark). This suggests that the inclusion of a wider variety of patient anatomies and irradiation configurations allows the network to better predict out-of-field dose distributions than focusing on a model that has been trained on data from a single accelerator, but with a smaller number of patients and diversity of treatment conditions. The results of the fine-tuned model nevertheless demonstrate the benefits of specializing the model to a certain extent. Indeed, the fine-tuned model makes it possible to achieve hitherto unequalled performance (0.13 ± 0.05 cGy.Gy^{-1}). This approach appears to be the most promising since it enables both optimizing performance on a particular machine of interest and using the vast diversity of treatment conditions found in the original database.

To address the identified limitations (overestimation of the lowest dose values far from the field and limited application for the most atypical linear accelerators), a strategy involving the incorporation of a relative or local component in the network's cost function, along with the inclusion of additional data in the training dataset, could offer remedies. Furthermore, a separate research team is collaboratively working on developing whole-body dose maps for contemporary linear accelerators and treatments, such as intensity modulation, using Monte Carlo simulation. Retraining the network with these new data will enhance its applicability to modern irradiation conditions. Finally, we carried out experimental measurements (and are still continuing this campaign) on a whole-body anthropomorphic phantom using radiophotoluminescent dosimeters with the aim of experimentally validating the whole-body dose maps generated by deep learning.

5 Conclusion

Based on this proof of concept, we have shown that deep learning is a relevant tool for addressing the limitations of analytical methods or MC simulations for out of field dose estimation. Thanks to its generalization capabilities and short inference times of just a few seconds, this tool should make it possible to move forward for routine clinical application and mass application in retrospective studies. We are convinced that a tool for estimating out-of-field dose in clinical routine would be a powerful aid for optimizing modern radiotherapy treatments.

References

1. Goy E, Tomezak M, Facchin C, Martin N, Bouchaert E, Benoit J, et al. The out-of-field dose in radiation therapy induces delayed tumorigenesis by senescence evasion. Elife. 2022 Mar 18;11:e67190.
2. de Kermenguy F, Meziani L, Mondini M, Clémenson C, Morel D, Deutsch E, et al. Radio-induced lymphopenia in the era of anti-cancer immunotherapy. In: International Review of Cell and Molecular Biology [Internet]. Academic Press; 2023 [cited 2023 Jun 30].
3. Colnot J, Zefkili S, Gschwind R, Huet C. Out-of-field doses from radiotherapy using photon beams: A comparative study for a pediatric renal treatment. J Appl Clin Med Phys. 2021 Feb 5;
4. Yoon J, Heins D, Zhao X, Sanders M, Zhang R. Measurement and modeling of out-of-field doses from various advanced post-mastectomy radiotherapy techniques. Phys Med Biol. 2017 Nov 13;62(23):9039–53.
5. Colnot J, Barraux V, Loiseau C, Berejny P, Batalla A, Gschwind R, et al. A new Monte Carlo model of a Cyberknife® system for the precise determination of out-of-field doses. Phys Med Biol. 2019 Oct 4;64(19):195008.
6. Howell RM, Scarboro SB, Kry SF, Yaldo DZ. Accuracy of out-of-field dose calculations by a commercial treatment planning system. Phys Med Biol. 2010 Dec 7;55(23):6999–7008.
7. Sánchez-Nieto B, Medina-Ascanio KN, Rodríguez-Mongua JL, Doerner E, Espinoza I. Study of out-of-field dose in photon radiotherapy: A commercial treatment planning system versus measurements and Monte Carlo simulations. Med Phys. 2020 Sep;47(9):4616–25.
8. Schneider U, Hälg RA, Hartmann M, Mack A, Storelli F, Joosten A, et al. Accuracy of out-of-field dose calculation of tomotherapy and cyberknife treatment planning systems: a dosimetric study. Z Med Phys. 2014 Sep;24(3):211–5.
9. Benzazon N, Colnot J, de Kermenguy F, Achkar S, de Vathaire F, Deutsch E, et al. Analytical models for external photon beam radiotherapy out-of-field dose calculation: a scoping review. Frontiers in Oncology [Internet]. 2023 [cited 2023 May 29];13.
10. Huynh E, Hosny A, Guthier C, Bitterman DS, Petit SF, Haas-Kogan DA, et al. Artificial intelligence in radiation oncology. Nat Rev Clin Oncol. 2020 Dec;17(12):771–81.
11. Diallo I, Lamon A, Shamsaldin A, Grimaud E, de Vathaire F, Chavaudra J. Estimation of the radiation dose delivered to any point outside the target volume per patient treated with external beam radiotherapy. Radiother Oncol. 1996 Mar;38(3):269–71.
12. Veres C, Allodji RS, Llanas D, Vu Bezin J, Chavaudra J, Mège JP, et al. Retrospective reconstructions of active bone marrow dose-volume histograms. Int J Radiat Oncol Biol Phys. 2014 Dec 1;90(5):1216–24.
13. Cardoso MJ, Li W, Brown R, Ma N, Kerfoot E, Wang Y, et al. MONAI: An open-source framework for deep learning in healthcare [Internet]. arXiv; 2022 [cited 2023 Apr 18]. Available from: http://arxiv.org/abs/2211.02701
14. Çiçek Ö, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O. 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation [Internet]. arXiv; 2016 [cited 2023 Jan 5]. Available from: http://arxiv.org/abs/1606.06650
15. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. 2015 May 18 [cited 2020 Aug 11]; Available from: https://arxiv.org/abs/1505.04597v1

Accurate Selection of Calculation Grid Size for Radiation Plans with Small Brain Tumors in Intracranial Stereotactic Radiotherapy

T.A-ALAWI, M. AL-GAOUD, E. RASHAIDI, N. SHORBAJI

Department of Radiotherapy, King Abdullah Medical City, Makkah, Saudi Arabia

1 Introduction

Stereotactic Radiosurgery or Radiotherapy SRS/SRT and Stereotactic Body Radiation Therapy SBRT technique facilitates delivery of high doses of radiation to the target volume in one to five fractions with acceptable toxicities to OARs.

Faster treatment time, higher accuracy in the delivery of high total dose to the target with acceptable toxicities to OARs, and patient comfort all contribute to a better clinical outcome [1]. SRS/SRT treatment for brain tumors requires conformity in both high and intermediate dose regions for better sparing normal brain tissue [2].

Intracranially, there are many indications that are currently being treated with SRS including, for example, AVMs, acoustic neuroma, meningiomas, pituitary adenoma, vagal schwannoma / paragangliomas, brain metastasis, anaplastic astrocytoma/ glioblastoma multiforme (GBM), trigeminal neuralgia, cochlea, thalamotomy, and pallidotomy [1]. Volumetric modulated arc therapy (VMAT) is a common treatment method for brain SRS/SRT which leads to a greater improvement in the radiation therapy treatments, where the radiation is delivered by the rotation of the linear accelerator gantry along one or more arcs, with the beam continuously on.

Typically, VMAT brain treatment planning is done with a beam energy from 6MV to 10 MV; these beams are often “hardened” with a flattening filter to produce a flat beam profile. However, there is also the option of a flattening filter free (FFF) beam, i.e. removing the flattening filter [3]. Removal of the flattening filter from a linear accelerator permits a considerable increase in the dose rate delivery during volumetric-modulated arc therapy (VMAT) [4].

Calculation grid size of 2.5mm spacing is sufficient to reduce dose calculation error to less than 1% in VMAT [5]. In radiosurgery, high doses of radiation are delivered safely due to more rigid immobilization and localization techniques, Current localization and positional techniques such as cone beam computed tomography (CBCT) and 6 degree of freedom couches have improved the localization accuracy to 0.2-0.6mm [5].

Thus with the ability to achieve tighter positional accuracy as well as sharper dose gradients, the smallest grid size should then be used in for accurate dose calculation. However, dose calculations using a fine grid size increases computation time significantly and may not be practical [5]

2 Materials and Methods

The study design is a retrospective study using the CT planning image acquired with radiosurgery treatment approach for 16 patients having small brain lesions treated between Jan 2021 to Jul 2023 and prescribed dose is 22.5Gy in five fractions. Eight treatment plans are to be generated for each patient using Varian Eclipse AAA analytical anisotropic algorithm for dose calculation and PO (photon optimizer) for dose optimization. First: four plans with the same MUs (monitor units) and with different grid sizes from 1mm to 4mm will be created. Second: another set of four plans will be planned with normalization at target mean and different grid sizes from 1mm to 4mm. Each plan contains two complete arcs with two different collimator angle sets. 6FFF energy (flattening filter free) beam will utilized for all plans to avail the advantage of high dose rate 1400MU/min, that has the benefit of using limited number of radiation beams to deliver such highly dose per fraction and to minimize the total treatment time.

- Arc1 start angle 181 end 179 with collimator 45 deg.
- Arc2 start angle 179 end 181 with collimator 315 deg

The ability to meet DVH objectives and constraints can be used along with dose conformity, homogeneity, and mean dose to the OARs to assess the plan quality [7].

The conformity index (CI) and homogeneity index (HI) of the resultant treatment plans will be computed using Eq. 1 and 2, respectively:

$$CI = V95\% / VPTV \quad (1)$$

The CI describes the volume within the PTV that receives the prescription dose or higher (V95%) divided by the

volume within the external body contour that receives the prescription dose or higher (VPTV) and provides a metric to determine how close the prescription iso-dose volume is spread within the PTV and normal tissue[7].

HI= D99.5% / D0.5% (2)

The HI describes the uniformity of the dose within the planning target volume and is a ratio of the minimum dose (D99.5%) to the maximum dose (D0.5%) [7].

The Gradient Index defined as the ratio of the 50% isodose (V50) volume to the prescription isodose volume [8]

GI = V50% /V100% (3)

3 Results

In terms of CI, it was clearly observed that for the first four plans where NN (no normalization) is selected, as the GS increases, the CI decrease Fig (1), whereas for the second group where TM (target mean) is selected , CI is not affected by changing the grid size.

However, the homogeneity index is not affected by changing GS for all types of plan normalization selected Fig(2) . On the other side, GI increases significantly with deceasing the GS, Fig (3).

Also there were no significant dosimetric differences in terms of OAR maximum doses as the grid size changes with both normalization types Fig (4, 5).

Finally the target volume PTV covered by 95 % of the prescribed doses is impacted significantly with changing the grid size , as the grid size increases the coverage decreases in the plans with NN is selected but there were no change in the target coverage as we keep plan normalized to target mean even with different GS Fig(6).

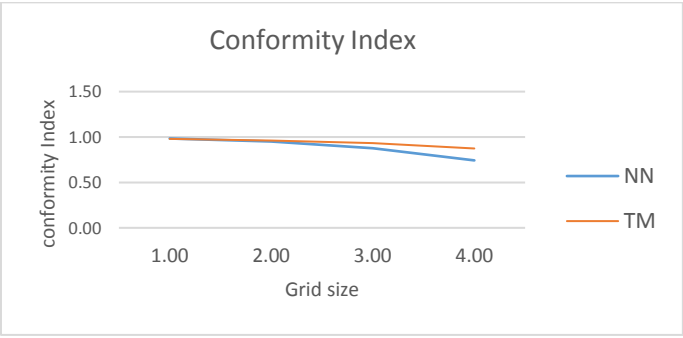


Fig 1 the effect of increasing the grid size on conformity Index

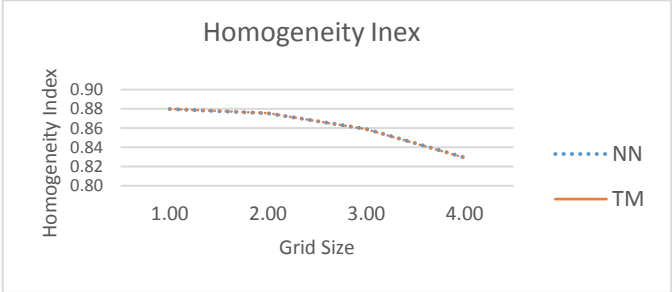


Fig 2 The Effect of increasing the grid size on homogeneity index HI and gradient index GI

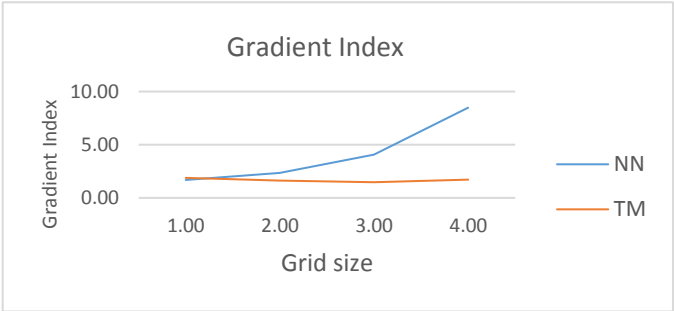


Fig 3 The Effect of increasing the grid size on homogeneity index HI and gradient index GI

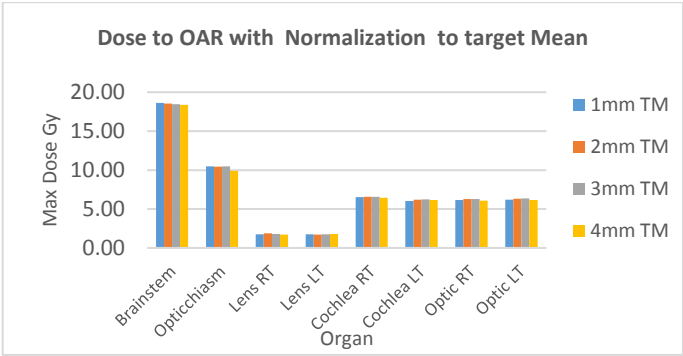


Fig 4 Dose to OAR with normalization to target mean plans

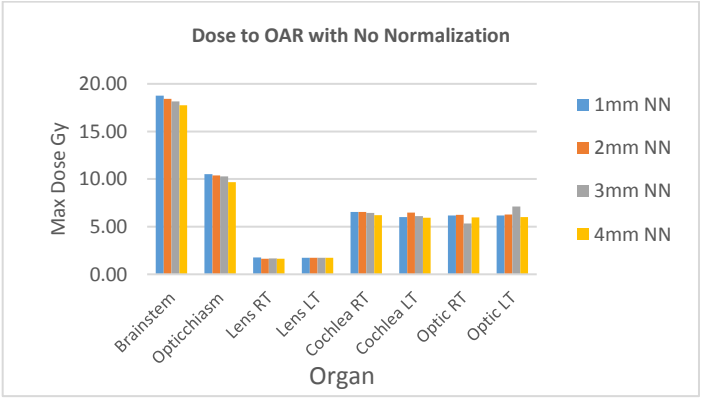


Fig 5 Dose to OAR with no normalization plans

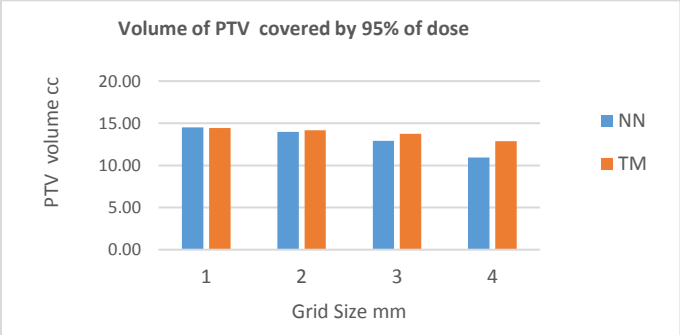


Fig 6 Target volume PTV covered by 95 % of the prescribed doses

4 Discussion

The main purpose of this study is to evaluate the effect of different calculation grid size on the accuracy of calculated dose for small brain lesions in intracranial stereotactic radiosurgery (SRS).

It was confirmed that by changing grid size during planning of small brain lesions, PTV covered by 95 % of the dose increased when the GS was decreased from 4mm to 1mm.

5 Conclusion

Selection of the calculation grid size is significantly influences the quality of the treatment plans in both conformity, gradient indexes and target coverage. For an accurate calculation of the treatment plans, it is recommended that the use of smallest possible grid size GS.

References

1) Stereotactic Radiosurgery and Stereotactic Body Radiation Therapy (SBRT), Dwight E. Heron, M. SaifulHuq, Joseph M. Herman, (Demos Medical Publishing is an imprint of Springer Publishing Company, LLC,2019)
2) Strategies to optimize stereotactic radiosurgery plans for brain tumors with volumetric-modulated arc therapy David Wang J Appl Clin Med Phys 2020; 21:3:45–51
3) A Dosimetric Study on Photon Beam Energies with and Without a Flattening Filter in SBRT Treatments for Prostate Cancer. Alexis F. Puruleski (Grand Valley State University, 2021)
4) Effects Of Flattening Filter-Free And Volumetric-Modulated Arc Therapy Delivery On Treatment Efficiency, Evan M. Thomas, (Journal Of Applied Clinical Medical Physics, Volume 14, Number 6, 2013)
5) Investigating the dosimetric effects of grid size on dose calculation accuracy using volumetric modulated arc therapy in spine stereotactic radiosurgery Karen Chin Snyder, MS, Department of Radiation Oncology, Henry Ford Health System, Jour. of Radiosurgery and SBRT, Vol. 4, pp. 303-313
6) Dose–Volume Constraints for organs at risk In Radiotherapy (CORSAIR): An “All-in-One” Multicenter– Multidisciplinary Practical Summary Curr. Oncol. 2022, 29, 7021–7050. https://doi.org/10.3390/curroncol29100552
7) Comparing planning time, delivery time and plan quality for IMRT, RapidArc and tomotherapy, Mike Oliver,a Will Ansbacher, Wayne A. B (Journal Of Applied clinical Medical Physics, Volume 10, Number 4, Fall 2009).
8) Plan Quality Assessment of Fractionated Stereotactic Radiotherapy Treatment Plans in Patients With Brain Metastases Frontiers in Oncology | www.frontiersin.org , March 2022 | Volume 12 | Article 846609

To calibrate or not to calibrate: a decision support approach using microcontroller-based tiny machine learning

Nicolas Boussion[1,2], Lou Bosc[1], Gaëlle Goasduff[1], Yoann Jegat[1], Laetitia Lann[1], Anne-Sophie Lucia[1], Emmanuelle Martin[1] and Dimitris Visvikis[1,2]

[1]Radiation Oncology Department, CHU de Brest, Brest, France
[2]LaTIM, INSERM, UMR 1101, Univ Brest, Brest, France

Abstract A decision support tool is proposed to determine whether it is reasonable to calibrate a linear accelerator with an ionization chamber according to recent changes in atmospheric pressure. The response of an ionization chamber is indeed sensitive to variations in atmospheric pressure, which can be very unstable in certain regions at some periods of the year. For this purpose, a neural network is trained on a two-year local weather station database classified using tools from statistical process control. The originality of the method relies on the use of machine learning code explicitly implemented for deployment on a microcontroller. More specifically, the training was done using a TensorFlow neural network implemented on a standard desktop computer. The obtained model was then deployed on an IoT Arduino board, by means of the TinyML library. The resulting device is low-cost, low-energy and has an accuracy of 0.834 for classifying atmospheric pressure.

1 Introduction

Dose calibration of a linear accelerator (linac) involves quantifying the link between measured dose in water under reference conditions and the fluence of emitted particles [1,2]. The dose measuring tool used for this purpose is an ionisation chamber whose response is highly dependent on atmospheric pressure. The measurement given by the chamber must therefore be corrected according to the current atmospheric pressure, which is generally read on an analogical barometer that is itself calibrated and regularly checked.

Many factors can lead to a correction in the calibration of the linac, which is checked daily before the first treatment. If the observed deviations exceed the permitted tolerances, recalibration is necessary. This is a long and tedious process, requiring careful experimental set-up and rigorous measurement. Calibration determines the accuracy of the dose that will actually be delivered to future patients, both to the target and to organs at risk and healthy tissue.

In many parts of the world, atmospheric pressure can change abruptly and/or reach exceptionally low or high values. Such unstable conditions are thus not ideal for ionisation chambers [3,4]. As a consequence, it becomes counterproductive to calibrate the accelerator, with the risk of having to start again a few days later when the pressure returns to a standard value in a more stable context.

In this context, we present a decision support tool based on an electronic board containing a pressure sensor and a microcontroller. A specific machine learning model is trained and then deployed inside the microcontroller.

2 Materials and Methods

The whole process has three main stages. The first step consists of data gathering and preparation. As a second step, a machine learning algorithm is trained to help deciding whether or not to calibrate according to the current atmospheric pressure. Finally as a last step, the trained model is modified so that it can be used by a microcontroller with ultra-light operating system.

Data collection and preparation

Atmospheric pressure values from the last two years, measured every three hours, were collected from the local weather station. These values were separated into three-month periods to reflect potential seasonal variations.

Exceptionally high or low pressure values, or values that drifted slowly or abruptly, were detected using statistical process control (SPC) tools [5]. More specifically, individual values chart (IV), moving range charts (MR) and exponentially weighted moving average charts (EWMA) were used to detect unusually low or high values, abruptly changing values and slowly drifting values, respectively [6]. Upper and lower control limits were established using 1.96 standard deviations for each type of chart.

Pressure values were then classified as unacceptable for linac dose calibration as soon as they exceeded the control limits of at least one of the three SPC charts.

Finally, all classified data were separated into 12 successive pressure values. The objective of the machine learning step was then to correctly classify the acceptable/unacceptable status of a given pressure value given the last eleven measures recorded every three hours.

Machine learning process

The proposed approach is fairly straightforward once the data has been collected and prepared as described above.

A simple neural network was implemented on a standard desktop computer using the TensorFlow [7] platform v2.4.1 and Python 3.8 in the Anaconda interface. The network consisted of two consecutive fully connected layers with 48 neurons followed by a ReLU activation function, one dropout layer with a 0.2 rate to prevent overfitting, and one fully connected layer with one output neuron followed by a sigmoid activation function. Dataset was split into 80% train-

ing data, 10% validation data and 10% test data. Three hundred epochs were used during the learning computation, using the Adam optimizer.

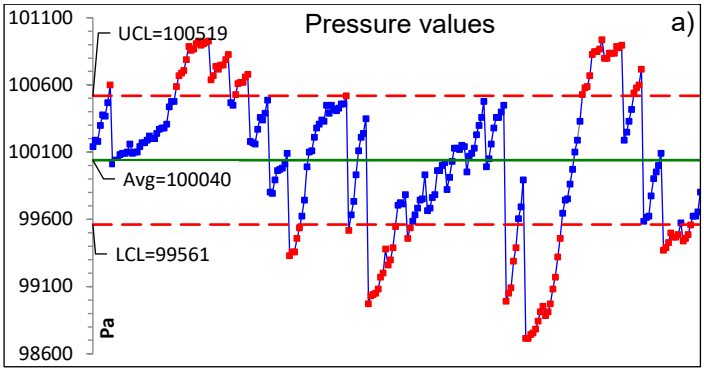
Microcontroller incorporation

Once built, the model was transferred into a Nano 33 BLE Sense board (Arduino, Torino, Italia) with embedded nRF52840 microcontroller (Nordic Semiconductor, Trondheim, Norway) and LPS22HB piezoresistive absolute pressure sensor (STMicroelectronic, USA). The process requires several steps in order to make the model work inside such a device, with only 1MB of CPU flash memory and 256 kB static RAM.
The objective of the first step was to convert the model to the TensorFlow Lite format using the dedicated TFLiteConverter function from TinyML [8]. No quantization was used because of the relatively small size of the model.
The aim of the next step was the effective deployment of the generated *.tflite* model inside the microcontroller. This was made possible by a simple conversion of this file into a hexadecimal *.h* format directly usable by the microcontroller.
As a last step, a C++ program was generated inside the Arduino IDE in order the microcontroller to read and record the pressure value every three hours during 33 hours, and to classify the current value according to the last eleven values and the deployed tiny machine learning model.

3 Results

A total of 5834 atmospheric pressure values were retrospectively collected during 2021 and 2022, from the local weather station, at a frequency of 1 measurement every 3 hours. A representative SPC analysis for the month of May 2022 is shown in Figure 1.
Atmospheric pressure was classified as unacceptable for calibration (Fig1-d) as soon as it exceeded control limits in at least one of the three analysis charts (Fig1-a,b,c).

For the sake of clarity, values are presented for a one month length, but in the study data were tracked during three months periods to account for potential seasonal effects.



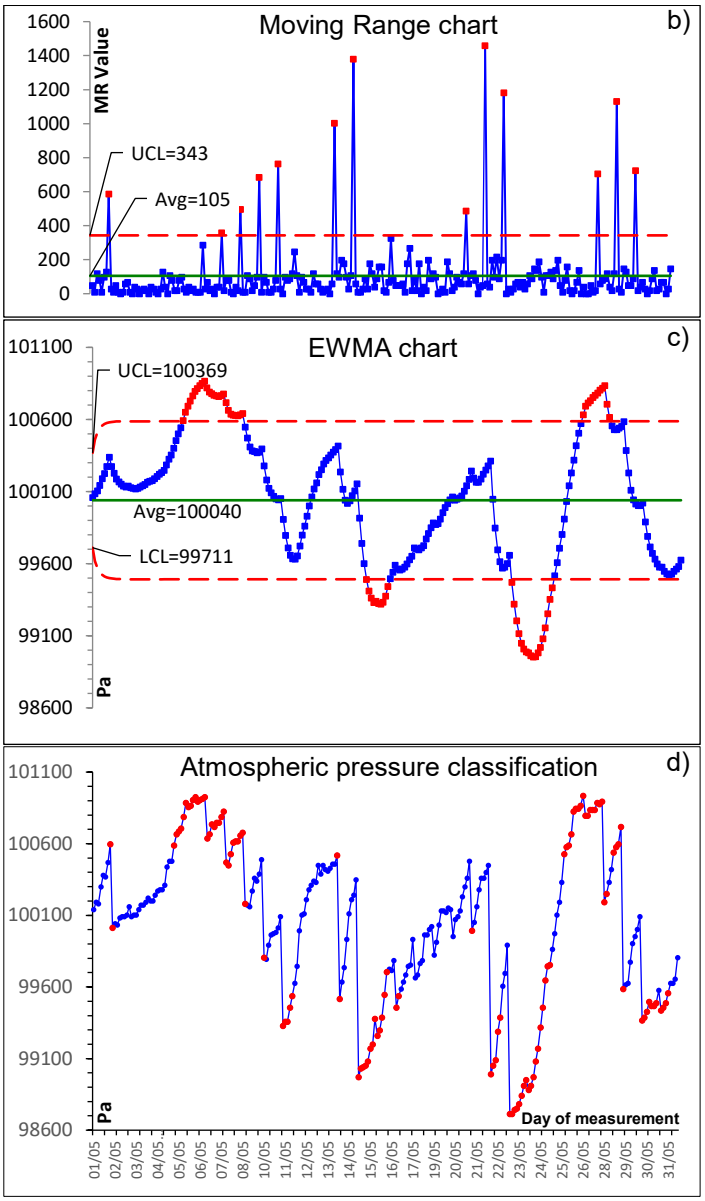


Figure 1: Atmospheric pressure values from the local weather station, recorded during May 2022, and SPC analysis. From top to bottom : a) individual values (for detection of very high or very low pressure values), b) moving range values (for detection of abrupt changes), c) EWMA values (for detection of smooth drifts), and d) final classification. Red dots correspond to pressure values above control limits. Values on the x-axis are all the same and correspond to the date and time of measurement during the month.

This type of individual classification was used to build the actual learning dataset, which consisted of independent series of twelve consecutive atmospheric values. The objective for the model was to predict the classification of the current pressure (acceptable or not for dose calibration) from the last eleven measurements (recorded every three hours, for a total of 33 hours prior to the current value).
Using this principle, a total of 486 independent series of 12 consecutive pressure measurements were generated (see Figure 2 for an example). Three hundred and eighty-eight were used for training, 49 for validation and a further 49 for model testing.

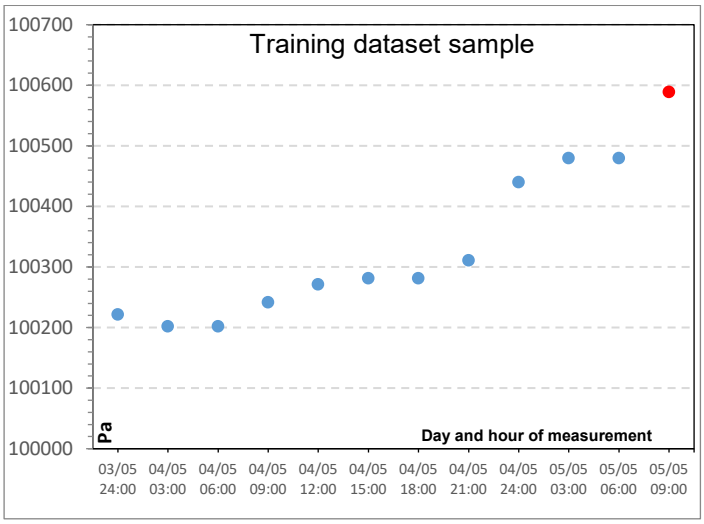


Figure 2: Sample from the training dataset. A sample from the training data, representing 12 consecutive measurements every three hours, over a period of 33 hours. The tested point is the last (05/05 at 09am), classified as unacceptable (red dot) by the SPC analysis. The training, validation and test datasets include 388, 49 and 49 independent series of this type, respectively.

The function loss for training and validation with respect to the 300 epochs is shown on Figure 3. Final accuracy for validation dataset was 0.874. Results for the test dataset were 0.834 for accuracy, 0.783 for recall, 0.841 for precision and 0.811 for f-score.

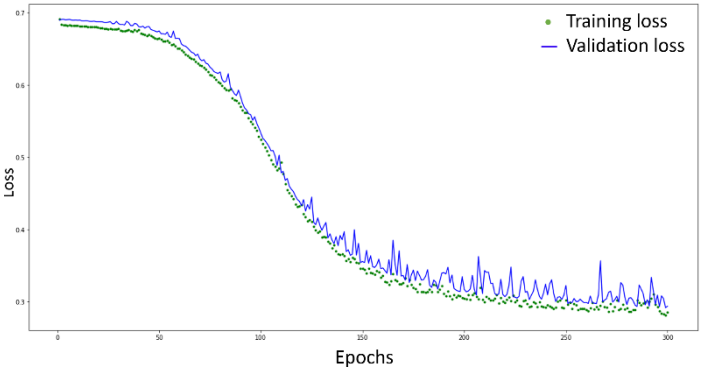


Figure 3. Training and validation loss during neural network optimization.

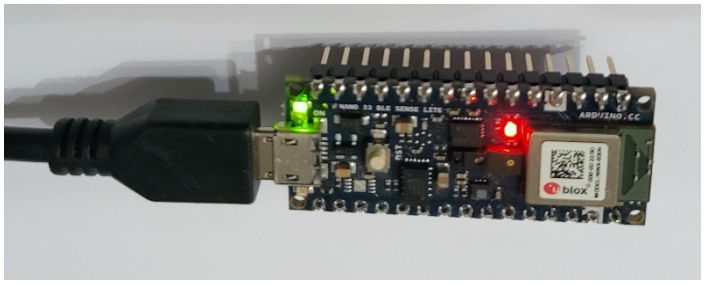


Figure 4. An example of minimalist output. In this configuration, the Arduino board was programmed to put the output LED on red in case of an unacceptable atmospheric pressure, and on green in the case of an acceptable pressure for linac calibration (LED on the left is on/off LED).

An example of possible output is given on Figure 4, where the classification of the current atmospheric pressure is provided by a simple output LED embedded on the board. If the LED is green, the atmospheric pressure is considered by the device to be sufficiently consistent and typical to proceed with calibration of the linac, by measuring the dose with an ionization chamber. Otherwise, the LED lights up red, as shown on the Figure.

4 Discussion

In this work a decision-support tool was presented for deciding whether to calibrate the delivered dose of a linac according to the stability of atmospheric pressure, which can drastically alter the reliability of dose measurement by ionization chamber.

In this exploratory and feasibility study, we have demonstrated the simplicity of the procedure, even if substantial room for improvement remains. In particular, the database can considerably be expanded, since atmospheric pressure measurements have been available since 2010 from the local weather station. Also, pressure measurements were separated into independent samples of 12 consecutive values, leading to only 486 samples. This number could be increased by sampling every 6 hours instead of every 12 hours, while maintaining a number of 12 values in each sample. It is not certain whether the resulting overlap would cancel out the gain in database size. Furthermore, the learning model proposed in this work is a very simple neural network that would benefit from further optimization.

5 Conclusion

The decision-support tool proposed in this study is relatively easy to implement for people with standard python and machine learning knowledge. The presented work provides the user with a low energy and low cost device able to classify the current atmospheric pressure as acceptable or not for linac dose calibration. Trends in atmospheric pressure, measured every three hours over the past 33 hours, are taken into account when making the decision. Accuracy of the machine learning model is 0.834 on test values, a score which could be easily improved by extending the training database and by refining the neural network structure.

References

[1] TRS 398. "Absorbed dose determination in external beam radiotherapy" (2000) IAEA, Vienna.

[2] PR. Almond, PJ. Biggs, BM. Coursey *et al*. AAPM's TG 51 protocol for clinical reference dosimetry of high-energy photon and electron beams. Med. Phys. (1999), pp 1847-1870. DOI: https://doi.org/10.1118/1.598691

[3] CB. Saw, LJ. Korb, KV. Krisna. "Local temperature and pressure

variations and its effects on daily ionometric measurements of radiations beam outputs from linear accelerators". Medical dosimetry (1997), pp 1-4. DOI: 10.1016/s0958-3947(96)00131-8

[4] G. Christ, OS. Dohm, E Schüle, S. Gaupp, M. Martin. "Air density correction in ionization dosimetry". Phys Med Biol (2004) pp 2029-2039. DOI: 10.1088/0031-9155/49/10/013

[5] T. Pawlicki, M. Whitaker, AL. Boyer. "Statistical process control for radiotherapy quality assurance". Med. Phys. (2005), pp.2777-2786. DOI: 10.1118/1.2001209

[6] K. Gérard, JP. Grandhaye, V. Marchesi, H. Kafrouni, F. Husson, P. Aletti. "A comprehensive analysis of the IMRT dose delivery process using statistical process control (SPC)". Med. Phys. (2009), pp.1275-1285. DOI: 10.1118/1.3089793

[7] M. Abadi, A. Agarwal, P. Barham, *et al*. "TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems". arXiv:1605.08695 (2015). Software available from tensorflow.org. https://doi.org/10.48550/arXiv.1605.08695

[8] R. Kallimani, K. Pai, P. Raghuwanshi, S. Iyer, O. López. "TinyML: Tools, Applications, Challenges, and Future Research Directions". arXiv:2303.13569 (2023), https://doi.org/10.48550/arXiv.2303.13569

Experiment of proof-of-principle on C-arm CT/SPECT system for in-vivo dose verification in online image-guided adaptive brachytherapy

Saerom Sung[1], Hyemi Kim[2], Chul Hee Min[1], Sei Hwan You[2], Hyun Joon Choi[2*]

1 Department of Radiation Convergence Engineering, Yonsei University
2 Department of Radiation Oncology, Wonju Severance Christian Hospital, Yonsei University Wonju College of Medicine

Abstract Traditional methodologies often fail to accommodate anatomical shifts or source movement, compromising treatment efficacy. In addressing the limitations of current brachytherapy, particularly the challenge of precise dose delivery and minimizing radiation exposure to normal tissues, this study experimentally validates the use of an integrated C-arm CT/SPECT system for real-time adaptive, image-guided brachytherapy. Utilizing a DRTECH detector, our prototype demonstrated significant advancements by accurately distinguishing between materials within a PMMA-Agar phantom and tracking Co-60 source movements with high precision. Experimental findings revealed the system's capacity for real-time modifications, with the in-house developed image intensity enhancement algorithm reducing the full width at half maximum of the Co-60 source to 3.80 mm and 2.91 mm along the x/y direction, respectively. This result shows the system's potential to surmount the prevailing challenges in brachytherapy by ensuring precise radiation delivery and enhancing patient safety and treatment outcomes through adaptive imaging and dose monitoring capabilities.

1 Introduction

Brachytherapy, especially for cervical cancer, generally prescribes doses ranging from 300 to 700 cGy across 5 to 10 sessions, raising the risk of side effects like bleeding and tissue necrosis if sensitive organs such as the rectum receive excessive radiation. Shigematsu et al. reported that 36% of 143 patients who underwent cervical cancer treatment experienced adverse effects due to rectal bleeding [1], while Cook et al. reported that 48% of adverse effects occurring after treatment in 348 patients with cervical cancer were related to the rectum [2].

The consistency of device placement and tissue positioning is vital to reduce these risks, yet variability in the attending physician for a single patient, patient movement, and anatomical changes pose challenges to maintaining treatment accuracy. Regular analysis of daily CT images during treatment sessions is crucial to adjust and verify treatment plans based on changes in the patient's condition, aiming to minimize side effects and complications. Prospective studies, like those by Georg et al., affirm the benefits of image-guided adaptive brachytherapy in lowering late side effects [3].

Nonetheless, the global standard procedure in image-guided brachytherapy, requiring patient transport between imaging and treatment rooms, risks device displacement and consequent dose distribution errors, as noted by Andrew et al. [4]. This highlights the urgent need for advanced, in-room 3-D imaging techniques that can obviate the need for patient transfer, offering precise verification of internal applicator positioning.

Despite technological advancements, many centers still continue to use 2D X-ray imaging for brachytherapy due to various internal constraints. Developing affordable, efficient, and room-integrated 3-D imaging solutions is crucial for advancing brachytherapy's effectiveness. Furthermore, real-time 3-D monitoring of the trajectory of radiation source movement and activity is essential during treatment sessions. Also, real-time in vivo patient dose monitoring based on this information is required.

Therefore, we previously proposed an integrated C-arm CT/SPECT imaging system to overcome the current limitation discussed above and realize an online adaptive image-guided brachytherapy (Fig. 1). In this study, we

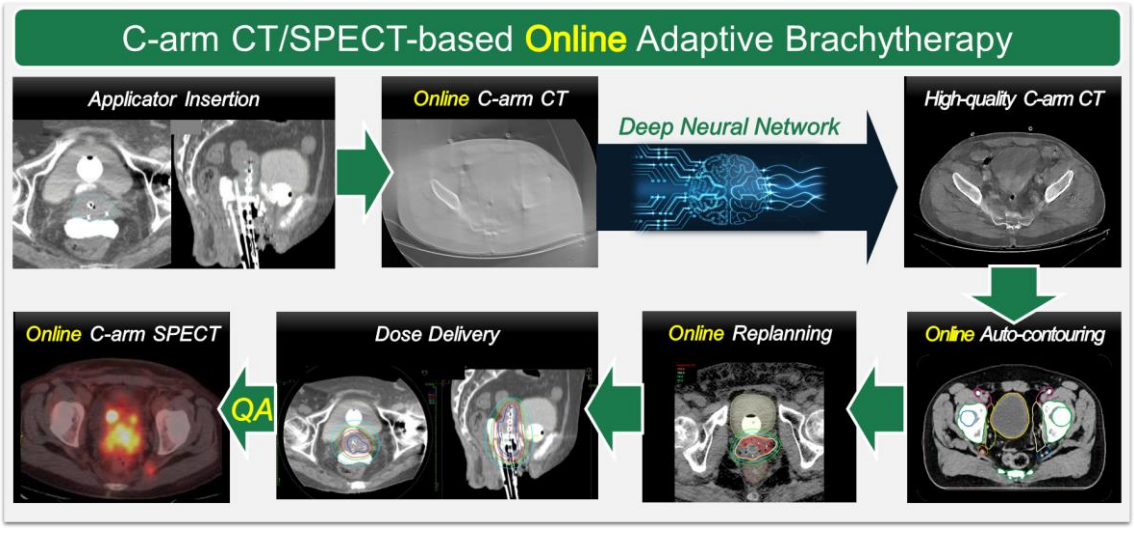


Fig. 1. C-arm CT/SPECT system-implemented clinical workflow for online adaptive brachytherapy.

experimentally validate the proof-of-principle of C-arm CT/SPECT system using commercial digital detector and in-house designed experimental solution.

2 Materials and Methods

A. Construction of a Prototype C-arm CT/SPECT System for Proof-of-Principle

For the acquisition of C-arm CT/SPECT images, an industrial DRTECH detector, EXT1616R, composed of a Csl scintillator and an IGZO TFT with a pixel pitch of 98 µm , was employed in this study. The detector, weighing ~3 kg with an imaging area of 161x161 mm² and external dimensions of 195x190x50 mm², was optimized for capturing both C-arm CT and C-arm SPECT images under various operational conditions. For C-arm SPECT imaging, a parallel-hole tungsten collimator, determined through prior Monte Carlo simulations, featuring 0.5 mm hole thickness, 45 mm length, and 0.2 mm septal thickness, was utilized. Additionally, to facilitate the acquisition of CT or SPECT images in three dimensions according to the rotation of the C-arm device, a rotation stage capable of simultaneously rotating both a PMMA phantom and a Two-ovoid applicator was constructed (Fig. 2). This rotation stage was equipped with an automatic rotation scheduler that allows rotation to a limited angle, considering the dwell time of each 58.12 GBq Co-60 source position according to the brachytherapy plan.

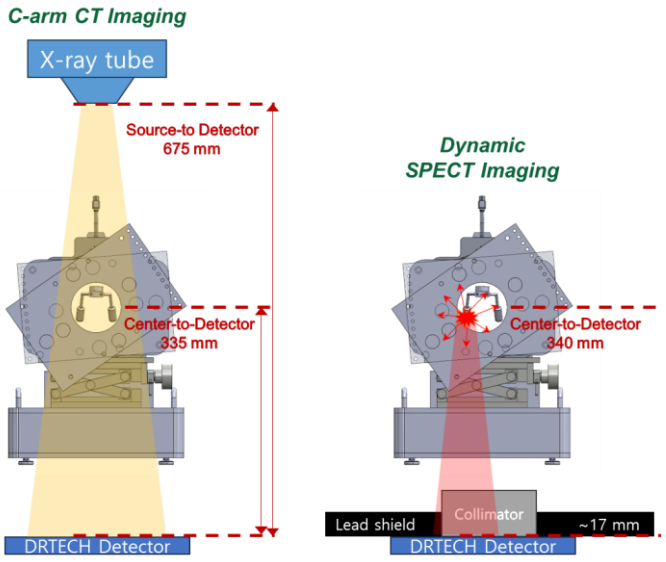


Fig. 2. Experimental setup for C-arm CT/SPECT imaging with an in-house designed phantom and applicator rotation equipment

B. C-arm CT/SPECT Image Acquisition based on Brachytherapy Plan using PMMA Phantom

To rotate a Two-ovoid applicator fixed at the center of a PMMA ($\rho \approx 1.18$ g/cm³) phantom, a phantom composed of an agar material ($\rho \approx 1.01$ g/cm³) in a cylindrical shape with a diameter and length of 10 cm each was fabricated. CT images of the PMMA-Agar phantom with the Two-ovoid applicator inserted were acquired using a TOSHIBA Aquillion CT scanner, and these images were utilized for planning a vaginal-cuff brachytherapy for a virtual target volume. The treatment plan was designed such that a 58.12 GBq Co-60 source would move 2 cm at 5 mm intervals from the endpoint, ensuring 1.83 Gy dose coverage over 90% of the target volume. The dwell times for the source at four positions were approximately 24 seconds, 51 seconds, 67 seconds, and 70 seconds, respectively, and the phantom rotation stage was scheduled to rotate repetitively by 110 degrees for each source position during their respective dwell times. C-arm CT images were acquired by capturing projection images at 1-degree intervals as the phantom rotated, under conditions of 90 kV and 10 mAs, and reconstructing these into a three-dimensional space. C-arm SPECT images were obtained by recording video in real-time at a rate of 1 frame per second while the phantom rotated. To enhance the quality of the acquired SPECT images, an in-house developed image intensity enhancement algorithm was applied.

3 Results and Discussion

When we obtained X-ray images at various phantom rotation angles using a commercially available industrial DRTECH detector, it was confirmed that the geometric structures of the Two-ovoid applicator, PMMA material, agar gel, and air cavity within the phantom were successfully distinguished. Similarly, CBCT images acquired following a 200-degree rotation of an Elekta VersaHD linear accelerator-mounted CBCT device, using the same phantom, also successfully differentiated boundaries between materials with similar densities inside the phantom (Fig. 3). This indicates that an online immediate brachytherapy plan modification is feasible in the treatment room during the patient setup using limited-angle C-arm CT images.

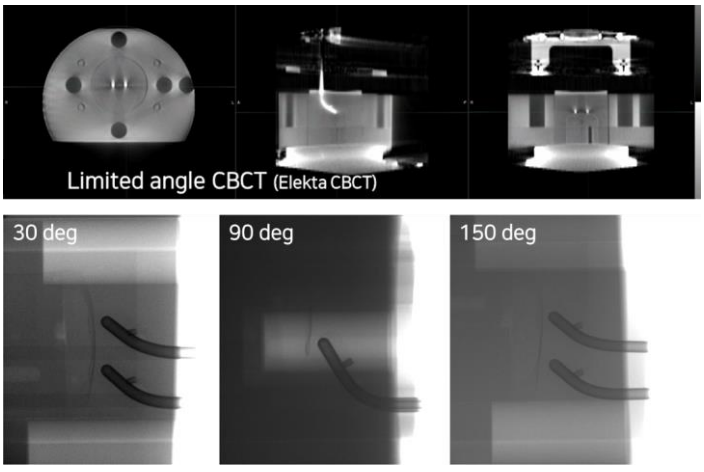


Fig. 3. X-ray Imaging at Various Rotation Angles Using a Commercial Detector (90 kV, 10 mAs)

In this experimental study for C-arm SPECT image acquisition, we verified the capability to track the position of Co-60 sources in real-time with a parallel-hole collimator structure optimized for Ir-192 sources, as predicted by our prior Monte Carlo simulation studies conducted. Analysis of a single frame of SPECT image demonstrated that, upon applying the in-house developed image intensity enhancement algorithm, the full width at half maximum (FWHM) of the peak for Co-60 source were evaluated to be 3.80 mm and 2.91 mm in the x and y directions, respectively (Fig. 4). Comparing this to the FWHM of 2.66 mm for a 275 GBq Ir-192 source assessed without the application of the image intensity enhancement algorithm, as found in our preceding Monte Carlo simulation study, it was confirmed that the movement characteristics of Co-60 sources within patients can be tracked with sufficiently high accuracy in actual clinical settings [5].

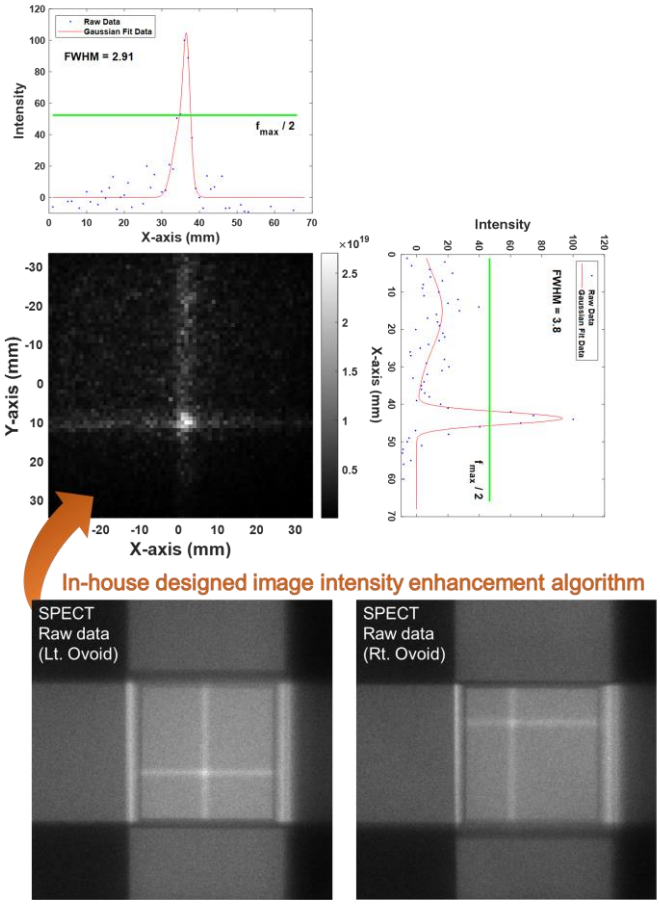


Fig. 4. Result of SPECT Imaging and FWHM Evaluation for Co-60 Source with In-house Designed Image Intensity Enhancement Algorithm

4 Conclusion

In this study, we experimentally verified the feasibility of online adaptive, image-guided brachytherapy using a C-arm CT/SPECT system. Given that the experiments were conducted using commercially available digital detectors, we believe these experimental results showed significant credibility for potential clinical application. The C-arm CT/SPECT device, designed to be more compact than O-arm devices and minimally obstructive to the conventional environment utilizing the C-arm device commonly used in various institutions, offers the advantage of online C-arm CT imaging for verifying a patient's internal condition. Additionally, it enables real-time in vivo patient dose monitoring during brachytherapy with C-arm SPECT imaging. In the future, we plan to further develop and optimize the operational and structural characteristics of the C-arm CT/SPECT device for practical use in clinical settings, leveraging these advantages.

References

[1] Shigematsu, Y., Nishiyama, K., Masaki, N., Inoue, T., Miyata, Y., Ikeda, H., ... & Kurachi, K. (1983). Treatment of carcinoma of the uterine cervix by remotely controlled afterloading intracavitary radiotherapy with high-dose rate: a comparative study with a low-dose rate system. International Journal of Radiation Oncology* Biology* Physics, 9(3), 351-356. DOI: 10.1016/0360-3016(83)90295-X

[2] Crook, J. M., Esche, B. A., Chaplain, G., Isturiz, J., Sentenac, I., & Horiot, J. C. (1987). Dose-volume analysis and the prevention of radiation sequelae in cervical cancer. Radiotherapy and Oncology, 8(4), 321-332. DOI: 10.1016/S0167-8140(87)80182-2

[3] Georg P, Lang S, Dimopoulos JC, et al. Dose-volume histogram parameters and late side effects in magnetic resonance imageguided adaptive cervical cancer brachytherapy. Int J Radiat Oncol Biol Phys 2011;79:356e362. DOI: 10.1016/j.ijrobp.2009.11.002

[4] Andrew M, et al. Reduction of applicator displacement in MR/CT-guided cervical cancer HDR brachytherapy by the use of patient hover transport system. Journal of contemporary brachytherapy, 2018; 10.1: 85. DOI: 10.5114/jcb.2018.73755

[5] S. Sung, M. Lee, H. J. Choi, H. Park, B. W. Cheon, C. H. Min, Y. S. Yeom, H. Kim, S. H. You, and H. J. Choi*, (2023). “Feasibility of internal-source tracking with C-arm CT/SPECT imaging with limited-angle projection data for online in vivo dose verification in brachytherapy: a Monte Carlo simulation study”. Brachytherapy, 22(5), 673-685. DOI: 10.1016/j.brachy.2023.05.003

Multiphysics modeling and simulations for Y-90 liver radioembolization

Emilie Roncali[1,2], Carlos A. Ruvalcaba[1], and Gustavo C.A. Costa[1]

[1]Department of Biomedical Engineering, University of California Davis, Davis , USA
[2]Department of Radiology, University of California Davis, Davis , USA

Abstract. Yttrium-90 (^{90}Y) radioembolization is a form of liver cancer therapy where radioactive microspheres are injected into the liver arterial bloodstream via a catheter, delivering localized radiation to the tumor cells. Studies have shown a strong correlation between the administered dose and patient outcomes, highlighting the importance of accurate dosimetry in treatment planning.

However, current standard-of-care dosimetry practices in liver radioembolization rely on generic models that do not accurately consider individual patient anatomy, vasculature, and tumor burden. To address this limitation, we present an approach involving computational fluid dynamics (CFD) and Monte Carlo simulation developed at UC Davis. Patient cone-beam CT (CBCT) images are collected to extract the liver vasculature, which is then used to simulate the blood and microsphere transport through CFD in OpenFOAM. The volumetric distribution of microspheres and ^{90}Y absorbed dose distribution in the liver is then computed with the Geant4-based application GATE. The pre-treatment absorbed dose calculations aim at determining the optimal dose distribution for personalized treatment planning.

This abstract outlines ongoing investigations to develop a personalized dosimetry approach using CFDose improving the model fidelity and validating it against clinical ^{90}Y PET.

1 Introduction

Yttrium-90 (^{90}Y) radioembolization is an internal radiation therapy to treat liver cancer (primary or metastatic), using microspheres 20-60 µm in diameter loaded with ^{90}Y. The ^{90}Y microspheres are injected selectively into the liver arterial bloodstream through a catheter to deliver localized radiation. ^{90}Y represented more than 10,000 interventions in the US in 2022 [1]; multiple studies have shown a clear correlation between the tumor dose and patient outcome, which indicates that dosimetry should be considered in treatment planning. However, standard-of-care dosimetry remains simplistic, based on generic models in which a simple scaling factor is used to compute the absorbed dose distribution from ^{90}Y activity [2]. This model does not account for the patient's unique anatomy, vasculature, and tumor burden. Due to the intricacy between blood flow, microsphere transport in the liver vasculature, and radiation physics, multi-physics computational methods are well-suited to achieve accurate radioembolization dosimetry [3]. This abstract presents our current efforts in combining computational fluid dynamics (CFD) and Monte Carlo simulation to develop personalized dosimetry informed with patient imaging. We collect patient cone-beam CT (CBCT) images to extract the liver vasculature, in which we simulate the blood and microsphere transport using CFD in OpenFOAM [4]. This provides the microsphere volumetric distribution and ^{90}Y activity distribution which in turn is used to compute the absorbed dose with Geant4 Application for Tomographic Emission (GATE) [5]. Our goal is to create a liver digital twin for radioembolization treatment planning, in which the digital twin is used to compute the optimal dose distribution. We called this liver model CFDose [3].

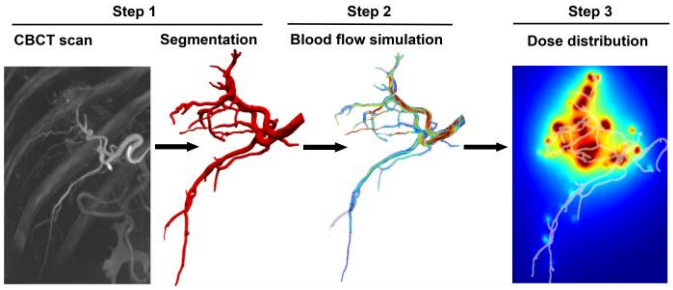


Figure 1. CFDose. 1) Hepatic artery segmentation from CBCT. 2) CFD simulation. 3) Simulated radiation dose distribution.

2 Materials and Methods

2.1 Patient data and image processing

In CFDose, personalization is achieved by using each patient's angiogram to create the vascular mesh to carry out the CFD simulation. The CBCT acquired during treatment planning was segmented with a piece-wise approach based on level sets in 3D Slicer and meshed with gmsh [6]. In the example shown in Figure 1, segmented vessels were as small as 0.45 mm in diameter, and 46 outlets were identified.

2.2 CFD simulation and multiscale model

To improve the precision of the microsphere simulation in small vessels, we modeled three phenomena that strongly direct the local microsphere deposition and were not previously considered in CFDose: blood is modeled as a non-Newtonian fluid, the catheter tip geometry and positioning extracted from patient scans (Figure 2), and discrete particle transport simulating microsphere dynamics. This new model included a 1-way coupling between the microspheres and the blood through discrete phase modeling to account for particle forces experienced by the microspheres during transport [7]. The model adequately captures the microsphere transport dynamically throughout the vasculature. Boundary conditions at the vessel wall are considered non-slip and non-deforming, while resistor-capacitor circuit analogs (known as 2- and 3-element

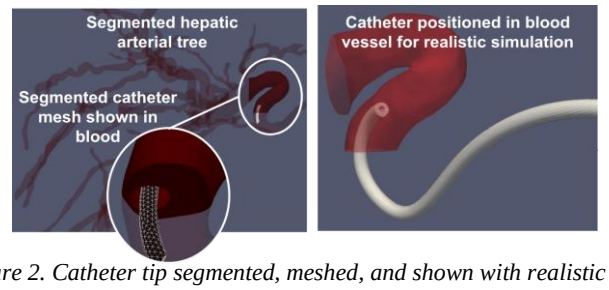


Figure 2. Catheter tip segmented, meshed, and shown with realistic positioning in blood flow.

windkessel-type) are prescribed at the vasculature downstream from the simulation domain which is not resolved in the image-based CFD domain. The resistors were assumed to be equally distributed amongst outlets, for each liver segment.

2.3 Microsphere dose distribution with GATE

^{90}Y is a β- emitter with a maximum energy of 2.28 MeV and a mean energy of 930 keV, corresponding to a maximum and mean electron range of 11 mm and 2.5 mm in water [8].

An isotropic dose point kernel with a 0.25 mm x 0.25 mm x 0.25 mm size was computed in GATE, using the ^{90}Y energy spectrum calculated from the Fermi theory for beta decay (Figure 3).

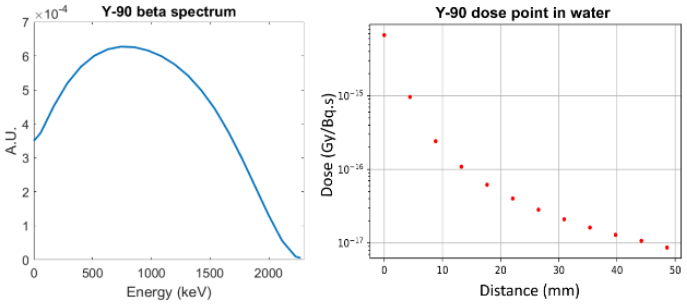


Figure 3. Left: ^{90}Y beta spectrum (mean energy 930 keV). Right: Dose kernel profile showing a rapid decrease of deposition with distance (log scale).

The dose distribution was obtained by convolving the dose kernel with the microsphere distribution obtained from the CFD simulation or using a full Monte Carlo simulation [9].

2.4 Comparison with clinical dosimetry

Scans up to 60 minutes were acquired within 48 hours of treatment to image the ^{90}Y activity with the total-body uEXPLORER scanner (United Imaging Health) for high quantitative accuracy at high resolution in a single bed position, using a protocol we previously validated [9]. The actual dose distribution was computed in the liver using the quantitative ^{90}Y images, the patient CT, providing the closest measurement to ground truth [9]. All PET acquisitions were accompanied by a low-dose CT scan for attenuation correction and anatomical delineation. Partial volume correction was applied using a previously determined recovery coefficient [10].

3 Results

3.1 CFD simulations

CFD simulations including particle transport showed the asymmetry of the flow and distribution between vessels and the importance of considering the vessel tortuosity when developing the flow upstream of the inlet (Figure 4). [11]

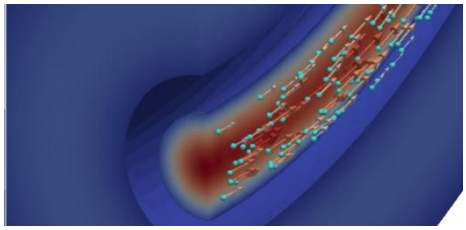


Figure 4. Illustration of skewed microsphere flow toward the vessel wall.

Our most recent work has focused on modeling the catheter geometry and flow within the vessel, to fully account for its perturbation of the blood flow in which the injected microspheres blend.

3.2 Comparison between simulated dose distribution and clinical PET

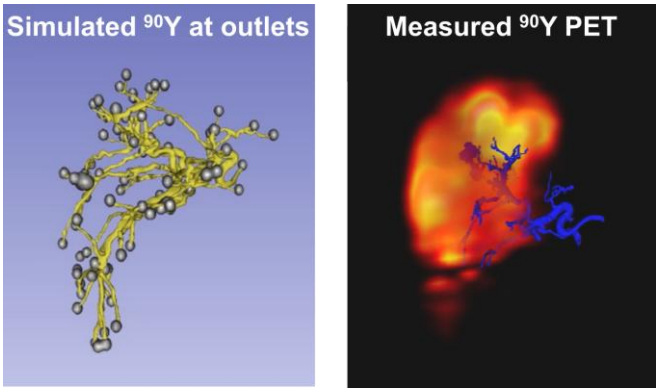


Figure 5. Left: Arterial tree segmented from CECT with predicted microsphere clusters at the outlets. Right: Clinical ^{90}Y PET overlaid with the arterial tree.

We compared the liver dose distribution with a clinical ^{90}Y PET scan acquired 1 h post treatment (Coronal view, Figure 5) and found that the predictive dosimetry was consistent with the activity mostly located in segments 7 and 8, as determined by a board-certified radiologist [3]. Figure 5 shows the 3D vascular mesh with clusters of microspheres simulated at each outlet. The cluster size is indicative of the number of microspheres. This encouraging first comparison indicates the accuracy of CFDose and shows the feasibility of validating CFD-based dosimetry with ^{90}Y PET.

Next steps include voxelwise comparison of the simulated and measured distributions, adding motion correction to the clinical PET data, and developing the liver model.

4 Discussion and conclusion

We present our newest results on the development of CFDose, a computational vascular model of the liver that we plan to integrate into an optimization program to find the best activity and injection spot in ^{90}Y radioembolization. Our next steps include personalization of the boundary conditions, which is an important part of transforming this model into a liver digital twin. While this model utilizes computationally intense CFD simulations, we plan to apply our expertise on generative adversarial networks (GANs) to speed up the computation and rapidly provide an estimated dose distribution.

5 References

[1] BTG. "Package Insert TheraSphere® Yttrium-90 Glass Microspheres. Rev 14," 2016; http://www.therasphere.com/.

[2] S. A. Gulec, G. Mesoloras, and M. Stabin, "Dosimetric Techniques in 90Y-Microsphere Therapy of Liver Cancer: The MIRD Equations for Dose Calculations," *J. Nucl. Med.*, vol. 47, no. 7, pp. 1209-1211, 2006.

[3] E. Roncali, A. Taebi, C. Foster, and C. T. Vu, "Personalized Dosimetry for Liver Cancer Y-90 Radioembolization Using Computational Fluid Dynamics and Monte Carlo Simulation," *Ann. Biomed. Eng.*, vol. 48, pp. 1499–1510 2020.

[4] C. Greenshields, *OpenFOAM v10 User Guide*, London, UK: The OpenFOAM Foundation, 2022.

[5] D. Sarrut *et al.*, "A review of the use and potential of the GATE Monte Carlo simulation code for radiation therapy and dosimetry applications," *Med. Phys.*, vol. 41, no. 6, pp. 064301, 2014.

[6] C. Geuzaine, and J. F. Remacle, "Gmsh: A 3- D finite element mesh generator with built- in pre- and post- processing facilities," *Numerical Meth Engineering*, vol. 79, no. 11, pp. 1309-1331, 2009.

[7] E. M. Childress, and C. Kleinstreuer, "Impact of fluid-structure interaction on direct tumor-targeting in a representative hepatic artery system," *Ann Biomed Eng*, vol. 42, no. 3, pp. 461-74, 2014.

[8] H. Ryde, P. Thieberger, and T. Alväger, "Two-Photon De-Excitation of the 0+ Level in Zr^90," *Physical Review Letters*, vol. 6, no. 9, pp. 475-478, 1961.

[9] G. C. A. Costa *et al.*, "Radioembolization Dosimetry with Total-Body Y-90 PET," *J. Nucl. Med.*, pp. jnumed.121.263145, 2021.

[10] B. A. Spencer, N. Omidvari, G. Costa, M. Rusnak, D. T. Caudle, C. Foster, C. T. Vu, and E. Roncali, "Comparison Between EXPLORER Total-body PET with Conventional PET Imaging of Y-90 Radioembolization: A Phantom and Clinical Study," in IEEE Nuclear Science Symposium and Medical Imaging Conference 2020.

[11] C. Ruvalcaba, A. Rajamani, and E. Roncali, "Characteristics of catheter injection for predictive particle transport modeling in Y-90 transarterial radioembolization procedures," *Bulletin of the American Physical Society*, vol. 11/21/2023, 2023.

Partial-ring PET Image Correction using Implicit Neural Representation Learning

Shubhangi Makkar[1,2], Siqi Ye[3], Marina Béguin[2], Günther Dissertori[2], Jan Hrbacek[1], Antony Lomax[1,2], Keegan McNamara[1,2], Christian Ritzer[2], Damien C. Weber[1,4,5], Lei Xing[3], and Carla Winterhalter[1]

[1]Center for Proton Therapy, PSI, Switzerland
[2]Department of Physics, ETH Zürich, Switzerland
[3]Department of Radiation Oncology, Stanford University, Palo Alto, California, USA
[4]Radiation Oncology Department of the University Hospital of Zurich, Switzerland
[5]Department of Radiation Oncology, Inselspital, Bern University Hospital, University of Bern, Switzerland

Abstract This work introduces a deep learning-based approach to enhance image quality in partial-ring PET scanners, which often suffer from significant artifacts due to incomplete angular field of view. Our method utilizes implicit neural representation (INR) learning with coordinate-based multilayer perceptrons (MLPs) featuring sinusoidal activations. These MLPs parameterize partial-ring PET images, training the patient-specific network to predict fully sampled images from partially sampled ones. This approach was validated using twenty digital brain phantoms, Monte Carlo simulated Derenzo phantom, and experimentally acquired hot rod phantom data from the partial-ring PETITION PET scanner. The results show that the INR learning method significantly improves image quality, achieving a PSNR above 30dB and SSIM over 0.95 for all cases. This method presents a promising solution for enhancing PET images from partial-ring scanners.

1 Introduction

Positron Emission Tomography (PET) plays an important role in functional imaging. Traditional PET systems rely on detectors arranged in a ring configuration [1, 2] to capture comprehensive imaging data across all angular views. However, the emergence of partial-ring PET systems has introduced both economical and practical advantages, particularly in specialized applications such as locoregional imaging [3] or online heavy-ion therapy verification [4]. Despite these benefits, partial-ring configurations inherently suffer from incomplete data acquisition, leading to significant image artifacts and reduced image quality.

Previously, model-based techniques have been employed to correct limited-angle PET images. For instance, sinogram completion using interpolation prior to reconstruction is an approach [5]. Another approach is the use of compressed sensing [6], which leverages the sparsity of data in a transform domain. Applying regularization constraints such as total variation [7] further helps in enforcing smoothness, thereby helping in artefact reduction. However, these model-based approaches have limitations, particularly when it comes to parameter optimization for regularization and computation times. To address these limitations, recent advances in deep learning (DL) and medical imaging have opened new pathways. The concept of implicit neural representation (INR) [8] presents a promising approach to obtain high-fidelity images from incomplete data sets [9]. The potential of deep learning, particularly in the context of INR learning, offers an innovative approach to correct and enhance PET images acquired from partial-ring systems. Unlike conventional DL approaches that depend on large labeled datasets, INR's ability to train unsupervised with sparse data makes it ideal for enhancing image quality in partial-ring PET systems. Here, we propose a framework that uses INR learning to effectively correct partial-ring PET images, particularly focusing on the developed partial-ring PETITION (PET for InTensive care units and Innovative protON therapy) scanner's design for proton therapy [10].

2 Materials and Methods

2.1 INR-based PET Image Correction

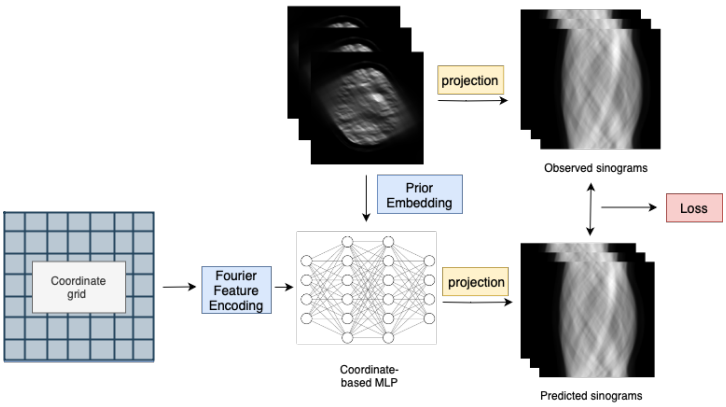


Figure 1: Network architecture for partial-ring PET image correction.

INR learning involves representing images using a neural network. We used a coordinate-based multi-layer perceptron (MLP) [8], $M(\theta)$ that transforms pixel coordinates $\{\mathbf{c}_i\}$ into their corresponding intensities.Our model consists of an MLP with 8 fully connected lay-

ers, 256 neurons each, employing sinusoidal activation functions. For efficient high-frequency signal representation, we used Fourier feature encoding, given by $F(\mathbf{c}_i) = [\cos(2\pi\mathbf{B}\mathbf{c}_i), \sin(2\pi\mathbf{B}\mathbf{c}_i)]^T$. The matrix $\mathbf{B}$, drawn from a Gaussian distribution with $\sigma = 4$ controls the frequency range. Inspired by a previous work that initialises the INR network parameters by pre-initialising the network with a prior image [9], we pre-trained the network with the uncorrected PET image from the limited-angle scan, which is the only accessible PET image of the scanned object in the real clinical scenario.

The PET reconstruction problem is often described as

$$\hat{\mathbf{x}} = \underset{\mathbf{x}}{\operatorname{argmin}} \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2, \tag{1}$$

where $\mathbf{x}$ is the reconstructed image, $\mathbf{y}$ is the measured data, and $\mathbf{A}$ is the forward model encapsulating the imaging physics.

With the INR learning, we parameterised the image $\mathbf{x}$ by the INR network, i.e., $\mathbf{x} = \mathcal{M}(\theta)$. The corresponding training loss function is therefore designed as

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \|\mathbf{A}\mathcal{M}(\hat{\theta}) - \mathbf{y}\|_2^2. \tag{2}$$

Once trained, the recovered image $\hat{\mathbf{x}}$ is inferred by $\hat{\mathbf{x}} = \mathcal{M}(\hat{\theta})$. The network architecture is depicted in Figure 1.

2.2 Datasets

Three distinct datasets employed for evaluation of our network are as follows:

- **MRI-based Brain Phantoms:** Twenty brain phantoms were sourced from BrainWeb [11, 12], and were adjusted to the PETITION scanner's resolution. Grey to white matter ratio was set at $4:1$, mimicking ^{18}F[FDG] PET and a hyperintense lesion was added onto each phantom. The geometry and the angular field of view of the PETITION PET scanner [10] was replicated.

- **Monte Carlo Simulation:** A Derenzo phantom with spheres of diameter 2.0, 2.2, 2.6, 3.2, 4.0 and 5.0 mm was simulated in GATE(v9.0) [13], where each sphere was uniformly filled with ^{18}F. The total simulated activity was 19 MBq with the acquisition time being 1800 seconds.

- **Real-Data Acquisition:** Utilizing a hot-rod phantom with uniform ^{18}F distribution, data was experimentally acquired with the PETITION PET scanner to reflect a realistic scenario of the physical effects.

2.3 Model Training

Model training was performed on a NVIDIA GV100GL GPU using PyTorch with the Adam optimiser. The model's performance was assessed using mean squared error (MSE), peak signal-to-noise ratio (PSNR), structural similarity index measure (SSIM) for different datasets in comparison to the ground-truth image that would have been obtained with a full-ring PET scanner. For the digital brain phantom dataset, we also compared the INR model with the widely used 2D U-Net [14]. Unlike the INR, the U-Net, being a data-dependent model, required division of the data into training, testing, and validation sets. Of the twenty brain phantoms, fifteen were allocated for training, four for testing, and one for validation. For the Derenzo phantom, peak-to-valley (PTV) ratios were computed for line profiles generated through different sphere diameters to assess the quantitative performance of the model. PTV ratio was calculated as the average intensity ratio of the uptake and no uptake region for each sphere.

3 Results

3.1 Digital Brain Phantoms

Figure 2 presents a comparison of various images for a slice of one of the brain phantoms. Quantitative comparison shows that the input partial-ring image lacks clarity (low PSNR and SSIM) and cannot reconstruct the lesion effectively. The U-Net and INR yielded similar SSIM and PSNR values, but INR-based correction outperformed U-Net in lesion reconstruction visually (figure 2) along with the required training times (15 minutes vs. 2 hours for U-Net).

3.2 MC Simulated Derenzo Phantom

Figure 3 displays the INR-corrected Derenzo phantom image, showing visually clear distinction between uptake and no-uptake regions. With prior initialization, INR achieves a PSNR of 31.73dB and an MSE error of 4×10^{-4}. Without prior initialization, it achieves a PSNR of 27.95dB and an MSE error of 1×10^{-3}. INR also exhibits higher SSIM values (0.95 with prior, 0.82 without) compared to the ground truth full-ring image. Quantitative analysis using PTV ratios (Figure 4) further validates the superior performance of INR, particularly when initialized with a prior. PTV ratios for the 2mm spheres are not indicated as this is beyond the resolution capabilities of the PETITION scanner.

3.3 Measured Hot-Rod Phantom

Visual comparison in Figure 5 demonstrates the improved image quality of the INR-reconstructed image

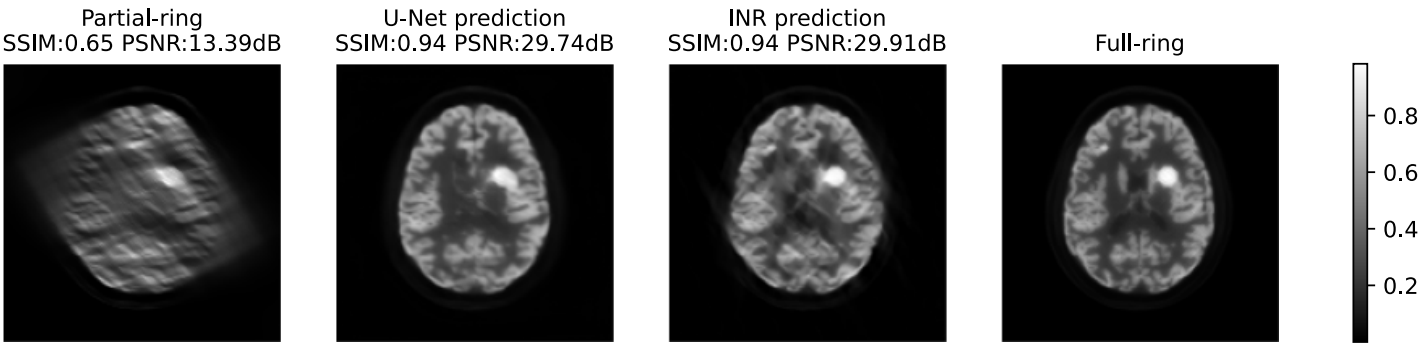


Figure 2: The central slice of a digital brain phantom. From left to right: original partial-ring MLEM reconstruction, corrected image using U-Net, corrected image using the proposed INR, and reference ground-truth image obtained with the full-ring scan.

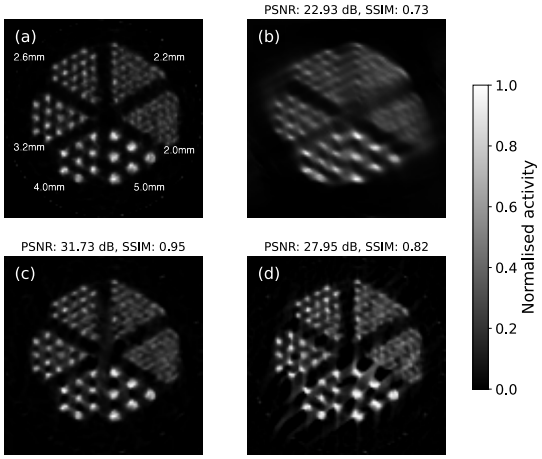


Figure 3: Results of the INR learning in representing the MC simulated Derenzo phantom.(a) Ground truth full-ring image, annotated with sphere diameters.(b) Partial-ring image.(c) Prior initialized INR corrected image and (d) INR corrected image without a prior.

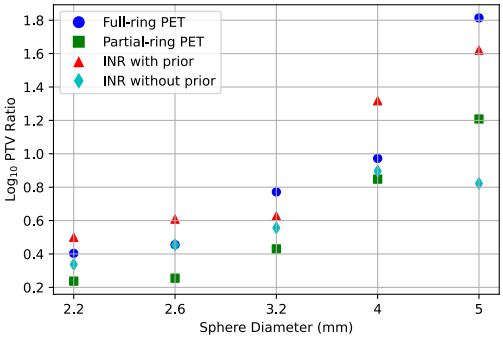


Figure 4: Peak-to-valley(PTV) ratio comparison amongst the four reconstructed Derenzo phantom images shown in Figure 3 for different sphere diameters.

compared to the initial partial-ring reconstruction. INR achieved a PSNR of 35.49 dB and an SSIM of 0.98 compared to the ground-truth full-ring PET image. Line profiles also illustrate INR's ability to recover intensities comparable to the ground truth full-ring image.

4 Discussion

This study presents an INR-based approach for partial-ring PET image correction. Unlike traditional data-driven deep learning methods that often suffer from overfitting issues and require extensive datasets, the proposed INR model does not require any external dataset but only uses the acquired partial-ring image for model training. This patient-specific nature of INR enables optimized imaging for each individual scan.

The efficiency of INR learning is seen by its faster training times compared to the 2D U-Net model. Additionally, the size of the model is dependent on the complexity of the signal rather than the discrete voxel grid size.

Our evaluation across three distinct datasets (analytically simulated brain phantom data, Monte Carlo simulated data, and phantom measurement data) demonstrates its robustness. For the digital brain phantom data, the INR model consistently achieved high PSNR and SSIM values, comparable or superior to U-Net, but with the added advantage of reduced training time. The model's ability to recover signal in artificially-inserted lesions in the brain phantom dataset, as seen in figure 2, illustrates its effectiveness in capturing structural changes and recovering missing signals. Inclusion of scanner effects in the dataset further extends the application of the method to realistic imaging scenarios. The performance improves with prior initialization of the INR model, especially in intensity prediction and separation of individual spheres in the Derenzo phantom. Similar performance was observed in measured data, reinforcing the application to different scenarios.

Furthermore, the proposed network architecture can significantly benefit longitudinal PET studies. For in-

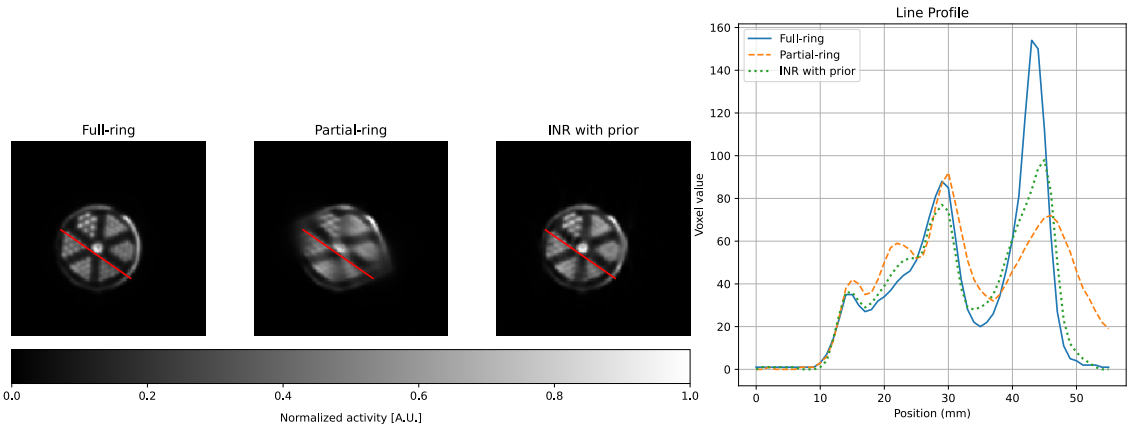


Figure 5: Depiction of full-ring, partial-ring and INR reconstructed images for the experimental hot-rod phantom. Line profile through the red line indicated in the three images is shown in the fourth column.

stance, in fractionated treatments where a new PET scan is required for verification or adaptation, the model trained during the first fraction can be reused for subsequent fractions. This not only saves inference time but also utilizes prior information to fine-tune important features.

Despite its advantages, the model has only been tested for phantom studies, due to the constraints of the PETITION PET scanner. This highlights the need for further validation with measured patient data. Future work will explore the model's performance in patient imaging and its extendibility to higher dimensionalities, such as 3D PET imaging.

5 Conclusion

This study demonstrates an INR-based approach for PET image correction, offering significant advantages in terms of efficiency and image quality. The model has been evaluated for partial-ring PET scanners without requiring a large training dataset. This approach helps in the development of further dedicated limited angle PET scanners or tomographs for varying applications.

Acknowledgements

This project was supported by the Swiss National Science Foundation, Grant No. CRSII5 189969.

References

[1] R. D. Badawi, H. Shi, P. Hu, et al. "First human imaging studies with the EXPLORER total-body PET scanner". *Journal of Nuclear Medicine* 60.3 (2019), pp. 299–303.

[2] M. Watanabe, K. Shimizu, T. Omura, et al. "A new high-resolution PET scanner dedicated to brain research". *IEEE Transactions on Nuclear Science* 49.3 (2002), pp. 634–639.

[3] S. Surti and J. S. Karp. "Design considerations for a limited angle, dedicated breast, TOF PET scanner". *Physics in Medicine & Biology* 53.11 (2008), p. 2911.

[4] M. G. Bisogni, A. Attili, G. Battistoni, et al. "INSIDE in-beam positron emission tomography system for particle range monitoring in hadrontherapy". *Journal of medical imaging* 4.1 (2017), pp. 011005–011005.

[5] H. W. De Jong, R. Boellaard, C. Knoess, et al. "Correction methods for missing data in sinograms of the HRRT PET scanner". *IEEE Transactions on Nuclear Science* 50.5 (2003), pp. 1452–1456.

[6] S. M. Valiollahzadeh, T. Chang, J. W. Clark, et al. "Image recovery in pet scanners with partial detector rings using compressive sensing". *2012 IEEE Nuclear Science Symposium and Medical Imaging Conference Record (NSS/MIC)*. IEEE. 2012, pp. 3036–3039.

[7] M. Burger, J. Müller, E. Papoutsellis, et al. "Total variation regularization in measurement and image space for PET reconstruction". *Inverse Problems* 30.10 (2014), p. 105003.

[8] V. Sitzmann, J. Martel, A. Bergman, et al. "Implicit neural representations with periodic activation functions". *Advances in neural information processing systems* 33 (2020), pp. 7462–7473.

[9] L. Shen, J. Pauly, and L. Xing. "NeRP: implicit neural representation learning with prior embedding for sparsely sampled image reconstruction". *IEEE Transactions on Neural Networks and Learning Systems* (2022).

[10] C. E. Ritzer, V. Commichau, J. Flock, et al. "The PETITION Project-Initial Monte-Carlo Simulations" (2021).

[11] B. Aubert-Broche, M. Griffin, G. B. Pike, et al. "Twenty new digital brain phantoms for creation of validation image data bases". *IEEE transactions on medical imaging* 25.11 (2006), pp. 1410–1416.

[12] C. Da Costa-Luis. "BrainWeb-based multimodal models of 20 normal brains". *Jul-2019 [Online]. Available: https://doi.org/10.5281/zenodo* 3269888 (2019).

[13] S. Jan, G. Santin, D. Strul, et al. "GATE: a simulation toolkit for PET and SPECT". *Physics in Medicine & Biology* 49.19 (2004), p. 4543.

[14] C.-C. Liu and H.-M. Huang. "Partial-ring PET image restoration using a deep learning based method". *Physics in Medicine & Biology* 64.22 (2019), p. 225014.

Artificial-intelligence enhanced image reconstruction for proton range verification

Lena M. Setterdahl[1], William Lionheart[2], Sean Holman[2], and Ilker Meric[1]

[1]Western Norway University of Applied Sciences, Department of Computer science, Electrical engineering and Mathematical sciences, Bergen, Norway
[2]Department of Mathematics, University of Manchester, Manchester, United Kingdom

Abstract Proton therapy is a cancer treatment method that utilizes proton beams to irradiate cancerous tissue, with a key objective of minimizing the dose to surrounding healthy tissue. Its full potential is limited by range uncertainties stemming from causes such as variations in patient anatomy and CT Hounsfield unit conversion uncertainties. In 2023, the NOVO collaboration introduced the NOVCoDA system, utilizing fast neutron and prompt-gamma rays emitted along the beam's path to detect beam range shifts. However, limitations in NOVCoDA, such as measurement uncertainties, pose challenges to achieving desired millimeter range shift detection precision. In general, conventional methods like MLEM and MAP result in noisy or less informative reconstructed images for proton range verification. In this work, we propose a hybrid approach, using images predicted by an encoder-decoder network (EDN) trained on Monte-Carlo data to guide MAP reconstruction. Tested on a single case, the MAP-EDN algorithm yields an image with nearly 4 times smaller mean squared error than the implemented list-mode MAP. Further exploration is warranted to assess the utility of the MAP-EDN algorithm for range shift identification.

1 Introduction

Proton therapy is a highly attractive radiotherapy treatment for cancer due to sharp dose gradients and greater healthy tissue sparing potential that it offers. However, its full potential is limited by range uncertainties caused by anatomical changes, organ motion, and day-to-day variations; tissue and tumor changes in response to treatment, patient setup or positioning errors; and inevitable uncertainties in the conversion of CT Hounsfield units to relative proton stopping powers. Numerous non-invasive range verification systems have been proposed, relying on imaging the emission probability distribution of secondary particles from proton interactions with tissue. Challenges include achieving millimeter range detection precision amidst limited counts and high statistical noise. Early 2023, the NOVO (NeutrOn and gamma-ray imaging for real-time range VerificatiOn and image guidance in particle therapy) collaboration demonstrated a hybrid multi-particle approach to range assessment-based treatment verification in particle therapy where a 1 mm range shift precision was achieved at relatively low proton intensities (10^7 protons per spot) [1], using list-mode MLEM, but noteably without the consideration of detector resolution effects. A core component of NOVO is the compact detector array known as NOVCoDA, composed of optically segmented, densely stacked organic scintillator bars. Imaging using the NOVCoDA relies on the assumption that the origin of detected particles, fast neutrons (FNs) and prompt-gamma rays (PGs), lie on the surface of a so-called event cone. Reconstructed event cones serve as input to the tomographic reconstruction of the particle emission origins used for range verification. Tomographic reconstruction for systems using event cones, such as Compton Cameras, neutron scattering cameras, and, in our case, the NOVCoDA, are usually based on algorithms like Maximum Likelihood Expectation Maximization (MLEM) and Origin Ensemble (OE), and variations thereof [1–3]. In the case of few detected events, these conventional methods produce extremely noisy images due to high statistical variations [4] and inevitable detector resolution effects like uncertainties in energy, time, and position measurements. Noise effects can be handled by introducing a regularizing function to the reconstruction, based on, for instance, prior information about the unknown emission distribution one seeks to image. In the context of proton beam range verification, the formulation of a suitable regularization function is still in its infancy, and to the best of our knowledge, the available literature on the subject is scarce. In this work we present a novel approach to image reconstruction for proton range verification using an encoder-decoder network (EDN) trained on Monte-Carlo simulated data in combination with a Maximum A-Posteriori (MAP) algorithm, referred to as the EDN-MAP algorithm. Traditional imaging relies on models of the physics and statistical nature of the measured signal and noise, e.g. assuming a Poisson model for the noise. On the other hand, a direct EDN-driven reconstruction solely relies on an EDN model whose weights are learned based on high-quality examples. Such a model operates as a "black box" and has low interpretability of the imaging process. Our proposed hybrid approach combines the traditional MAP and EDN reconstruction (heavily inspired by the use of AI in PET reconstruction [5]), leveraging the EDN-predicted image to guide the iterative MAP update. Specifically focusing on 2D imaging and with a list-mode approach, we compare the effects of EDN-MAP on the calculated range landmark (RL), defined in the next section, to the conventional reconstruction algorithms Simple Backprojection (SBP), MLEM, and MAP. We do not evaluate OE algorithms in this work and consider only imaging of FNs, leaving PGs, and combined FN and PG imaging for future studies.

2 Methods

As a benchmark for the proposed EDN-MAP algorithm, we used the list-mode approach to SBP, MLEM, and MAP. A direct reconstruction algorithm, SBP simply projects event cones onto a discretized image plane positioned on and whose surface normal is perpendicular to the proton beam axis. The intersection of multiple cone surfaces with the plane provides an estimate of the particle emission distribution. The iterative methods MLEM and MAP are based on a Bayesian approach, where the probability distribution of the sought emission distribution, represented by the image vector $\boldsymbol{\lambda}$, given the data vector $\boldsymbol{y}$ is expressed using Bayes' rule

$$p(\boldsymbol{\lambda}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{\lambda})p(\boldsymbol{\lambda})}{p(\boldsymbol{y})}. \quad (1)$$

MAP aims to find the solution that maximizes the a-posteriori probability distribution $p(\boldsymbol{\lambda}|\boldsymbol{y})$, while MLEM searches for the solution that maximizes the likelihood $p(\boldsymbol{y}|\boldsymbol{\lambda})$. The objective function in MAP is

$$\max_{\boldsymbol{\lambda}} \log p(\boldsymbol{\lambda}|\boldsymbol{y}) = \max_{\boldsymbol{\lambda}} \log p(\boldsymbol{y}|\boldsymbol{\lambda}) + \log p(\boldsymbol{\lambda}). \quad (2)$$

Given that the list-mode data is modeled as independent Poisson random variables, the point-wise update equation (derived for binned data in [6]) for the MAP estimate is

$$\lambda_j^{new} = \frac{\lambda_j^{old}}{s_j + \beta \frac{\partial V(\boldsymbol{\lambda})}{\partial \lambda_j}\Big|_{\boldsymbol{\lambda}^{old}}} \sum_i \frac{a_{ij}}{\sum_b a_{ib}\lambda_b^{old}}. \quad (3)$$

where λ_j^{old} is the old value of pixel j, s_j is the pixel sensitivity (set to $s_j = 1, \forall j$ for simplicity), a_{ij} is the system matrix element indicating the probability that the ith event of the list was emitted from pixel j. System matrix elements were computed using an approach similar to the marching algorithm presented in [7]. A common means of describing $p(\boldsymbol{\lambda})$ is with a Gibbs prior

$$p(\boldsymbol{\lambda}) \propto e^{-\beta V(\boldsymbol{\lambda})} = e^{-\beta \sum_j \sum_k (w_{jk}\phi(\lambda_j - \lambda_k))} \quad (4)$$

where the potential function $V(\boldsymbol{\lambda})$ describes desired properties of the unknown image, $\phi(u)$ is a non-negative function and symmetric about 0, and the weight w_{jk} is a measure of the strength of the relationship between pixels j and k: $w_{jk} = 1$ for orthogonal nearest neighbors, $1/\sqrt{2}$ for diagonal neighbors and 0 otherwise. As the parameter $\beta \to 0$, (3) tends to the MLEM update function. In this work we have chosen the potential function examined in [8]

$$V(\boldsymbol{\lambda}) = \sum_j \sum_k w_{jk} \log\left(1 + \left(\frac{\lambda_j - \lambda_k}{\mu}\right)^2\right) \quad (5)$$

whose partial derivative with respect to λ_j, needed in (3), is

$$\frac{\partial V(\boldsymbol{\lambda})}{\partial \lambda_j} = \sum_k w_{jk} \frac{2(\lambda_j - \lambda_k)}{\mu^2 + (\lambda_j - \lambda_k)^2}. \quad (6)$$

This potential function discourages the separation of uniform regions, penalizing high-frequency noise, while avoiding excessive penalties for larger separations that are likely to occur at the boundary between two distinct regions of the image, such as the edge of the FN emission distribution. The parameter μ adjusts how much the prior is influenced by the difference between nearby pixel values. In the EDN-MAP algorithm, we use an EDN-predicted image $\boldsymbol{\lambda}^*$ to guide MAP reconstruction by replacing λ_k in (6) with λ_k^*.

Encoder-Decoder Network

The EDN was created using the machine learning framework PyTorch (v. 2.0.1) [9]. The architecture of the EDN considered in this work comprises 3 convolutional layers in the encoder, a linear (fully connected) layer before and after the latent space that is of dimension 20, and 3 transpose convolutional layers in the decoder. The convolutional layers utilize 4x4 kernels with (2x2) pixel strides. Training data was generated using GATE (v.9.0) simulations [10] of a Gaussian profiled ($\sigma_y = \sigma_z = 2$ mm) proton beam fired at a $40 \times 40 \times 40$ cm^3 water phantom. The proton beam energy was set to 85 MeV and contained 10^9 protons (corresponding to a high-intensity treatment beam spot). Produced within the phantom by primary proton interactions, FNs were detected by a $30 \times 20 \times 20$ cm^3 detector volume with a material composition similar to that of plastic scintillators (with carbon-to-hydrogen atom ratio 10:11 and density 1.099 g/cm^3). A total of 11 range shifts were simulated by adding and removing material in steps of 1 mm from the phantom of the dimension parallel to the beam. Following simulations, FN hits in the detector volume were processed in ROOT v.6.28/06 [11], involving the introduction of detector resolution to the hit data (using resolutions of an organic glass scintillator (OGS) - a candidate scintillator for the NOVCoDA: 10% energy resolution, 10 mm depth-of-interaction resolution, and 500 ps time resolution [12]), the pairing of two consecutive detector hits caused by the same neutron - i.e., double coincident events (DCE) - and construction of FN event cones. For each processed range shift simulation, bootstrapping was used to randomly sample (with replacement) n event cones to represent lower treatment beam intensities ($N = 10^6, 10^7, 10^8$). The number of bootstrap samples was determined using the observed DCE efficiency $\eta = 1.7 \cdot 10^{-4}$ per primary proton and desired beam intensity N

$$n = \eta N. \quad (7)$$

Using different random seeds, 100 bootstraps were performed for each range shift and the three proton intensity. The SBP and ground truth (GT) image of the FN emission distribution was generated for all bootstraps, yielding $11 \times 3 \times 100$ = 3300 sets of SBP-GT image pairs. The images measured (400×400) pixels, with pixel width and height equaling 1 mm, and were normalized such that pixel intensities were in the range of [0, 1]. A single SBP-GT image pair was

isolated for each range shift and proton intensity, saved for later used in evaluation of the EDN-MAP algorithm. The remaining dataset was split into 70:30 training and validation subsets. In the training step, the SBP image serves as the input and the GT as the target image. The optimization of model weights involved minimizing the mean squared error (MSE) loss between the model's output and the associated GT. This optimization process utilized the 'Adam' optimizer [13] with an arbitrarily selected learning rate of 0.001. A batch size of 10 was selected without specific criteria, and the network underwent training for 10 epochs. Training concluded when the validation loss showed signs of stagnation. Training and validation of the model took roughly 1 hour to complete and was conducted on a MacBook Pro (2017) running macOS Monterey (version 12.5.1) with a 3.1 GHz Intel Core i5 processor, 8 GB of memory, and 256 GB of storage. The machine was equipped with an Intel Iris Plus Graphics 650 graphics card.

Evaluation

To compare the performance of the EDN-MAP and conventional image reconstruction algorithms, we use the MSE as a measure of agreement to GT image and the standard deviation (SD) to quantify noise levels of the image in question.

$$MSE = \frac{1}{J}\sum_{j=0}^{J-1}(\lambda_j^{recon} - \lambda_j^{GT})^2 \tag{8}$$

where J represents the number of pixels, λ_j^{recon} is the reconstructed image, and λ_j^{GT} is the GT image. The SD of the image, whose average pixel value is $\bar{\lambda}$, was calculated by

$$SD_{\boldsymbol{\lambda}} = \sqrt{\frac{1}{J}\sum_{j=0}^{J-1}(\lambda_j - \bar{\lambda})^2} \tag{9}$$

We introduce a range landmark (RL) parameter to assess algorithms performance on range prediction performance, defining RL as the weighted average of the normalized 1D emission distributions, obtained by collapsing the image onto the axis parallel to the beam axis,

$$RL = \frac{\sum_m (h_m bin_m)}{\sum_m h_m} \tag{10}$$

where h_m is the weight and bin_m is the position (in mm) of the mth bin. The standard deviation of the estimated RL

$$SD_{RL} = \sqrt{\frac{\sum_m h_m (bin_m - RL)^2}{\sum_m h_m}} \tag{11}$$

The relative RL error, using $\Delta RL = |RL - RL^{GT}|$ was computed as

$$\delta = \frac{\Delta RL}{RL^{GT}} \times 100\% \tag{12}$$

Evaluation metrics were computed for images cropped at 100 mm depth into the phantom.

3 Results

The EDN-MAP algorithm's performance was evaluated against SBP, MLEM, and MAP reconstructions for a single range shift case: a 0 mm range-shifted proton beam of 10^8 primaries (corresponding to $1.7 \cdot 10^4$ imageable FN event cones). Hyperparameters were determined through a series of trials and subsequent adjustments based on performance outcomes. In (3), β was set to 0, 0.2, and 0.1 for MLEM, MAP, and EDN-MAP reconstruction, respectively, and μ in (6) was set to 1. The iterative methods were stopped after an arbitrary number of 20 iterations. For the GT and the reconstructed images in Fig. 1, the corresponding evaluation metrics are reported in Tab. 1 and 1D emission distribution are shown in Fig. 2. Among the methods evaluated, EDN-MAP emerged as the top performer, generating an image that closely resembled the GT with the lowest MSE, nearly 4 times lower than the MAP image, and relative RL error. In contrast, conventional methods exhibited anticipated interpretability challenges, such as severe blurriness (SBP), noise amplification (MLEM), or a combination of both issues (MAP). Note that the reconstructed 1D distributions' tail make their respective RL value larger than the GT's.

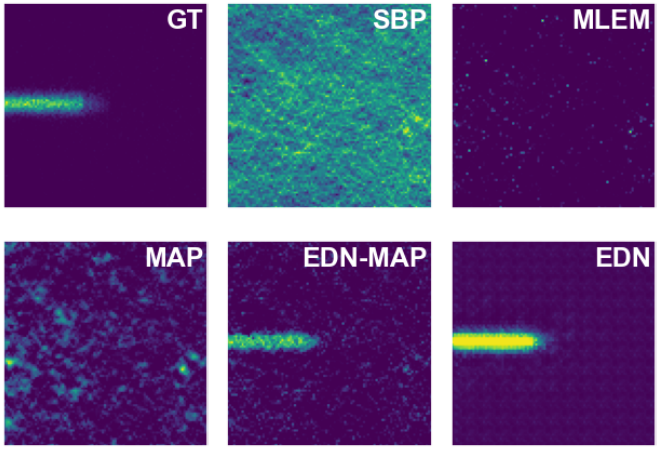


Figure 1: Ground truth (GT) and reconstructed images for 10^8 primary protons with 0 mm range shift. Images cropped are at 100 mm into the phantom.

4 Discussion

Despite promising results, acknowledging challenges and limitations is crucial. The MAP update function in (3) is not guaranteed to converge in general and depends on the choice of potential function $V(\boldsymbol{\lambda})$ and the value of β, previously pointed out in [6]. Additionally, limited training data and EDN model complexity and hyperparameters (e.g., depth, loss function, optimizer, learning rate) may impact performance. Future studies will explore different network archi-

Table 1: Evaluation metrics of various algorithms compared to ground truth (GT).

Alg.	MSE	SD_{λ}	RL [mm]	SD_{RL} [mm]	ΔRL [mm]	δ [%]
GT	0.0	0.99	23.26	2.52	0.0	0.0
SBP	2611.16	1.02	50.81	2.92	27.54	118.41
MLEM	28.23	5.20	42.89	5.64	19.63	84.37
MAP	1.37	0.69	42.43	4.85	19.17	82.39
EDN-MAP	0.35	1.02	36.75	3.92	13.49	57.99
EDN	0.68	1.58	34.27	3.47	11.01	47.34

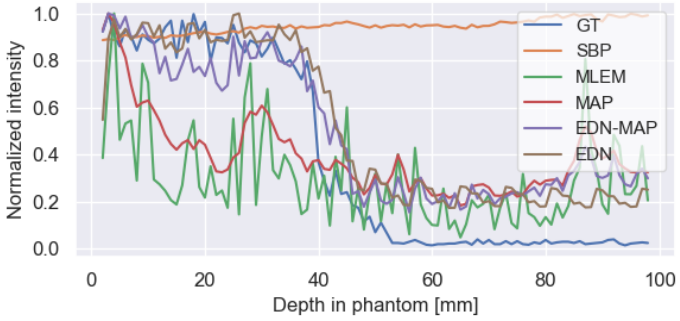


Figure 2: 1D projection of the reconstructed and ground truth (GT) FN production distribution, for 10^8 primary protons with 0 mm range shift.

tectures, hyperparameter optimization, and the generation of realistic training data, considering phantom heterogeneities and various clinical proton beam energies. Moreover, there will be a focus on investigating whether range shifts can be successfully identified and at what precision levels this can be achieved.

5 Conclusion

The EDN-MAP produces an improved image as indicated by the MSE metric calculated for this single case. However, more efforts are needed to produce more conclusive evidence for the usefulness of the proposed method. The current work is the preliminary stage of an extensive project, and additional efforts are required to comprehensively evaluate the efficacy of the method for range shift identification in both homogeneous and even more importantly, heterogeneous phantoms.

6 Acknowledgments

This work is supported by The Research Council of Norway (Grant no. 301459).

References

[1] I. Meric, T. Kögler, H. E. S. Pettersen, et al. “A hybrid multi-particle approach to range assessment-based treatment verification in particle therapy”. *Sci Rep* 13 (2023), p. 6709. DOI: 10.1038/s41598-023-33777-w.

[2] I. Valencia Lozano, G. Dedes, S. Peterson, et al. “Comparison of reconstructed prompt gamma emissions using maximum likelihood estimation and origin ensemble algorithms for a Compton camera system tailored to proton range monitoring”. *Z Med Phys* 33.2 (2023), pp. 124–134. DOI: https://doi.org/10.1016/j.zemedi.2022.04.005.

[3] J. Braverman, J. Brennan, E. Brubaker, et al. “Single-Volume Neutron Scatter Camera for High-Efficiency Neutron Imaging and Spectroscopy” (2018). ArXiv Instrum. and Detect. arXiv:1802.05261.

[4] N. Kohlhase, T. Wegener, M. Schaar, et al. “Capability of MLEM and OE to Detect Range Shifts With a Compton Camera in Particle Therapy”. *IEEE Trans Radiat Plasma Med Sci* 4.2 (2020), pp. 233–242. DOI: 10.1109/TRPMS.2019.2937675.

[5] A. J. Reader and B. Pan. “AI for PET image reconstruction”. *Brit J Radiol* 96.1150 (Oct. 2023), p. 20230292. DOI: 10.1259/bjr.20230292.

[6] P. Green. “Bayesian reconstructions from emission tomography data using a modified EM algorithm”. *IEEE T Med Imaging* 9.1 (1990), pp. 84–93. DOI: 10.1109/42.52985.

[7] S. Wilderman, W. Rogers, G. Knoll, et al. “Fast algorithm for list mode back-projection of Compton scatter camera data”. *IEEE T Nucl Sci* 45.3 (1998), pp. 957–962. DOI: 10.1109/23.682685.

[8] T. Hebert and R. Leahy. “A generalized EM algorithm for 3-D Bayesian reconstruction from Poisson data using Gibbs priors”. *IEEE T Med Imaging* 8.2 (1989), pp. 194–202. DOI: 10.1109/42.24868.

[9] A. Paszke, S. Gross, F. Massa, et al. *PyTorch: An Imperative Style, High-Performance Deep Learning Library*. 2019.

[10] D. Sarrut, M. Bardiès, N. Boussion, et al. “A review of the use and potential of the GATE Monte Carlo simulation code for radiation therapy and dosimetry applications”. *Med Phys* 41 (June 2014), p. 064301. DOI: 10.1118/1.4871617.

[11] R. Brun and F. Rademakers. “ROOT - An Object Oriented Data Analysis Framework”. *Nucl Instr Meth A* 389 (Sept. 1997). Proceedings AIHENP’96 Workshop, Lausanne, pp. 81–86. DOI: 10.1016/S0168-9002(97)00048-X.

[12] J. Turko, R. Beyer, A. Junghans, et al. “Characterization of organic glass scintillator bars and their potential for a hybrid neutron/gamma ray imaging system for proton radiotherapy range verification”. *J Instrum* 19.01 (Jan. 2024), P01008. DOI: 10.1088/1748-0221/19/01/P01008.

[13] D. Kingma and J. Ba. “Adam: A Method for Stochastic Optimization”. *ICLR* (Dec. 2014).

Head and Neck Deformable Registration Using Self-Attention with Positional Encoding

Elizabeth McKenzie[1], Taman Upadhaya[1], Hengjie Liu[2], and Ke Sheng[3]

[1]Department of Radiation Oncology, Cedars-Sinai Medical Center, Los Angeles CA, USA
[2]Department of Radiation Oncology, University of California Los Angeles, Los Angeles CA, USA
[3]Department of Radiation Oncology, University of California San Francisco, San Francisco CA, USA

Abstract This study proposes a completely attention-based deformable image registration (DIR) model which we call AmorWarp. An attention-based model allows one to incorporate long-range dependencies in registration techniques. Large motion is not well handled in DIR, and we endeavor to harness attention's unique abilities to improve DIR in the setting of large head-and-neck motion. Our resulting model demonstrates improved robustness for both inter- and intra-patient DIR relative to an established CNN model and a B-spline technique. In the setting of large motion, our method reduced the average 95% Hausdorff distance across 17 contours from 17.12mm (pre-registration) to 6.88mm for intra-patient registrations, and from 15.52mm to 7.92mm for inter-patient registrations. Our method reduced the average target registration error across 4 landmarks from 9.80mm to 3.37mm. Our attention-based model performed similarly to more established methods, and we hope that this proof of concept initiates further DIR applications of this powerful technique.

1 Introduction

Deformable image registration (DIR) remains a challenging problem in medical imaging. Despite its importance in tracking disease progression, combining imaging modalities, and adapting radiation therapy to changes in patient anatomy, there are still many opportunities for improvement, especially in the presence of large motion[1]. Network-based DIR, primarily using convolutional neural networks (CNNs)[2-10], has shown several advantages in registering images with small deformation but is limited by a local approach to image processing and struggles with large deformations.

To address these limitations, we propose the use of self-attention networks, a technique initially developed for natural language processing but recently shows promise in image processing[11]. Self-attention networks, unlike CNNs, consider the entire image to determine responses at specific points, facilitating long-range data connections[12]. We leverage this strength to build a network to take cues from the entire 3D volume, thus potentially providing more accurate registration in scenarios with significant anatomical movement, such as in head and neck radiation therapy.

However, self-attention networks are computationally expensive, especially for large 3D images. To overcome this, we use computationally efficient axial attention, which breaks down images into 1D components and computes attention sequentially, reducing computational complexity[12].

We extend this idea to build a fully self-attentional DIR model for 3D images, predicting deformation vector fields (DVF) for a pair of images. This model was tested on head and neck CT scans, and its performance was compared with a traditional CNN-based DIR model[3,] and a B-spline DIR[13-14] approach. The results demonstrate the potential of self-attention networks to accurately handle large motion and complex deformations in medical image registration.

2 Materials and Methods

We collected head and neck CT images from The Cancer Imaging Archive[15] to create a training dataset of 409 volumes, with testing and validation datasets of 44 and 53 volumes, respectively. Additionally, we collected paired CT's from 14 patients for intra-patient registration testing. All volumes were rigidly registered to a template, downsized to 128x128x128 (4mm isotropic voxels), clipped to an intensity range of [-1024, 3000], and then normalized to [-1,+1].

Our network, which we call *AmorWarp*, takes a pair of 3D images (source and target) as input and outputs a predicted DVF, which is then used to deform the source image. To combine the two input volumes, we first pass each image through two convolution layers, concatenate, and perform max pooling. This results in a 64x32x32x32 tensor. After the images are combined, they enter the U-shaped portion of the network. The number of features going down the encoder branch was 256, 512, 1024, and 2048 for each level, respectively. The decoder branch reversed these feature numbers. Each level of the U-net was constructed using a self-attention block with positional encoding and a residual identity path, followed by a non-linearity. Each level of the U-net was connected via skip connections with concatenation and convolutions of kernel size 1. The attention block was both preceded and followed by 1x1x1 kernel 3D convolution. Each axial attention component was constructed by reshaping the input into a 1D tensor along the selected dimension. For example, the axial depth tensor had the shape [*batch size* x *height* x *width*, *features*, *depth*]. By factoring a 3D tensor into sequential 1D operations along the height, width, and depth, we efficiently trained a network with a global receptive field for attention. The axial attention formulism with positional encoding is given by:

$$y_o = \sum_{p \in \mathcal{N}_{1xm}(o)} softmax_p\left(q_o^T k_p + q_o^T r_{p-o}^q + k_p^T r_{p-o}^k\right)\left(v_p + r_{p-o}^v\right)$$

where m is the span of the entire input at that layer, and r_{p-o}^q, r_{p-o}^k, and r_{p-o}^v are the relative learned positional

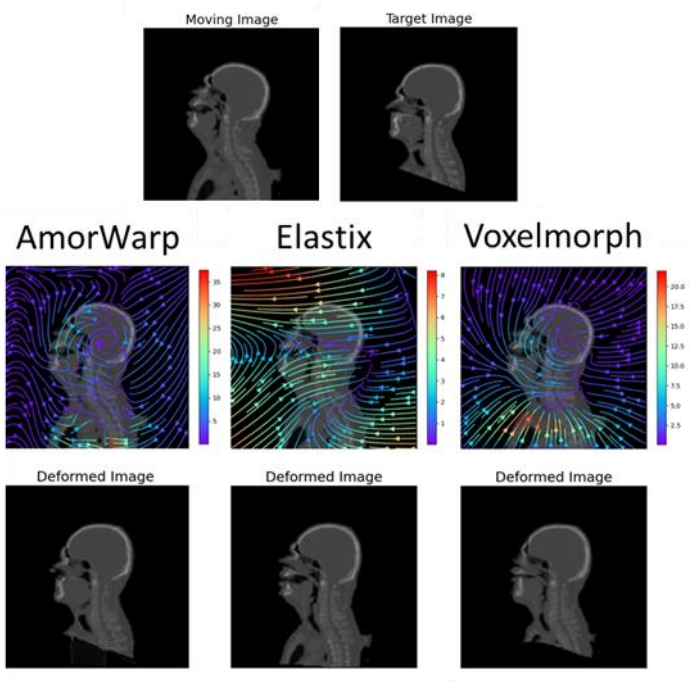


Figure 1 Comparison among registration methods: Amorwarp (axial attention), Elastix (B-spline), and Voxelmorph (CNN). The deformation vector fields are overlaid on the moving image [middle row]. The resulting deformed moving images [bottom row] show our methods ability to match the target image even in the presence of large neck flexion.

encodings for the queries (q), keys (k), and values (v), respectively. Note that the subscript is $p - o$ for the positional encodings, indicating that these positions are relative relationships between the input and output locations. The result is a map of attention that looks to the entire input for guidance, incorporating information about all the input values and their positions. The above formulation was first proposed by Wang et al[12], and we have adapted it for 3D inputs and outputs in DIR.

Our loss function contained 3 terms: a mean square error fidelity term between the deformed source image and target image, a bending energy term to promote a smoothly varying DVF, and a deep feature VGG loss term[16-17] to improve detail matching between the deformed and target images. Additionally, we trained the network with and without an average dice term derived from 17 previously auto-segmented contours[18] (note that the contours were not provided as input).

The network was trained using stochastic gradient descent, 10 warm-up epochs, and an initial learning rate of 0.0257 with a cosine learning rate schedule. The model was trained for 280 epochs, with random pairs of source and target images sampled from the 409 samples for a total of 167,281 potential pairs.

We analyzed the network's performance for both inter and intra-patient deformable image registration. We compared the accuracies of the registrations using 95% Hausdorff distance of 17 automatically created head and neck contours, as well as the target registration error (TRE) of 4 manually placed landmarks (nose, front teeth, chin, and



Figure 2 Saliency map showing where the network is attending. Without input contours, the network has learned to pay attention to anatomically relevant portions of the image when predicting the deformation vector field.

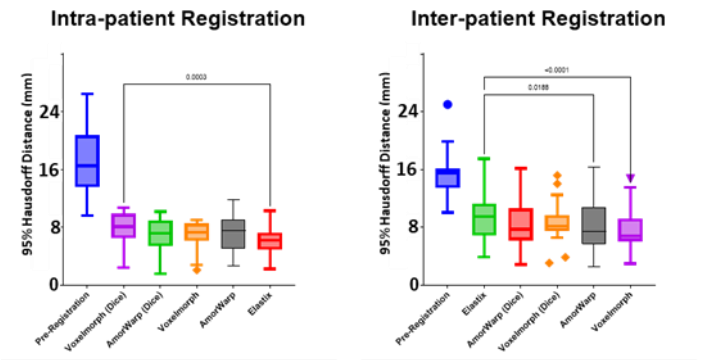
epiglottis). These results were compared among our axial attention-based network, a CNN-based network (Voxelmorph[3]), and a B-spline non-network registration application (Elastix[13-14]).

3 Results

Qualitatively, our network matched the head and neck CT volumes well even in the presence of gross anatomical distortion (e.g., opening and closing the mouth). Inspecting the DVF's showed that it learned discontinuities around the motion of connected components, demonstrating the model's ability to associate disparate parts of the image. The B-spline and CNN methods showed relatively unanatomical smooth DVF's, and a corresponding lower aptitude in regions of large deformation, such as going from a chin-up position to a chin down [Fig. 1].

In attention-based networks, the saliency maps are more interpretable since they reveal what the network has learned to "pay attention to."[11] When examining the saliency maps, we discovered that our network learned to pay attention to relevant portions of the anatomy without any contouring input. For example, it learned to attend to the mandible, shoulders, and skull [Fig. 2]. Of note, the shoulders and skull were not contours included in the average dice loss term, and thus emerged strictly from the training.

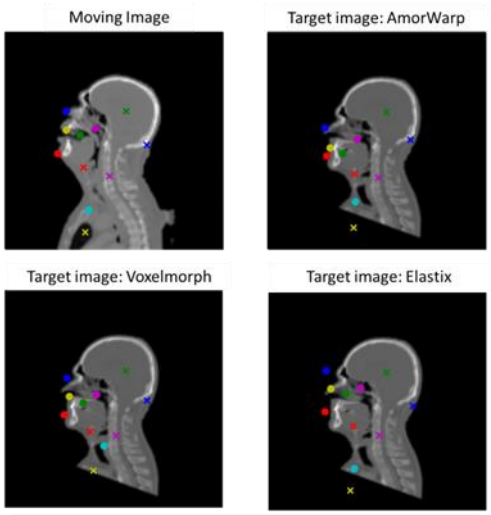
Quantitatively, we compared contour matching using 95% Hausdorff distance averaged across 17 head and neck contours (brainstem, cord, chiasm, right and left orbits, right and left submandibular glands, right and left optic nerves, right and left parotids, right and left cochleae, mandible, larynx, pharynx, and esophagus) [Fig. 3]. For the inter-patient comparisons, the best performers were the CNN and our attention-based techniques, which were both significantly better than the B-spline technique. For the intra-patient comparisons, the best was the B-spline technique, followed by our attention-based technique. Interestingly, we found that including an average dice loss



Type	Mean (mm)	Stdev (mm)
Pre-Registration	17.12	4.96
Voxelmorph (Dice)	7.76	2.48
AmorWarp (Dice)	6.92	2.4
Voxelmorph	6.92	2
AmorWarp	6.88	2.48
Elastix	6.12	2.04

Type	Mean (mm)	Stdev (mm)
Pre-Registration	15.52	3.36
Elastix	9.36	3.68
AmorWarp (Dice)	8.64	3.72
Voxelmorph (Dice)	8.64	3.12
AmorWarp	8.36	3.92
Voxelmorph	7.92	3.2

Figure 3 Evaluation of registration accuracy using 95% Hausdorff distance across 17 automatically generated head and neck contours. The left column gives the accuracy when registering two images of the same patient (intra-patient registration). The right column is for registrations between images from two separate patients (inter-patient). Pre-Registration indicates the amount of contour matching prior to deformable registration. Voxelmorph is the CNN-based deformable registration, Elastix is the B-spline-based iterative deformable registration, and AmorWarp is our proposed axial attention approach. Networks trained with a contour dice loss are indicated with a (Dice) suffix. Each boxplot is summarized in a table below the figure. Our proposed method is second-best for both intra- and inter- patient registration, demonstrating robustness in both large and small deformations.



Type	Mean (mm)	Stdev (mm)
Elastix	9.80	2.80
Pre-Registration	7.66	1.63
Voxelmorph (Dice)	5.44	2.04
Voxelmorph	4.03	1.29
AmorWarp	**3.56**	**1.67**
AmorWarp (Dice)	**3.37**	**1.62**

Figure 4 Evaluation of intra-patient registration accuracy using target registration error (TRE) for a selection of manually created landmarks. The top of the figure shows an example comparing the deformable image registrations (DIR) for the proposed (AmorWarp), CNN-based (Voxelmorph), and B-spline-based (Elastix) methods. The registration has been applied to the moving image's landmarks, and each method's deformed points are superimposed on the target image. The average and standard deviations of the TRE across the 14 test patients is shown in the bottom table. Our proposed registration method was superior to both the CNN and B-spline based techniques.

term led to poorer performance for both the CNN and attention-based networks.

In addition to contour matching, we also investigated intra-patient landmark matching which gives a finer analysis of how well the individual voxels are deformed. Our analysis revealed that the axial attention-based network outperformed both the CNN and B-spline method. Figure 4 visualizes this performance, demonstrating how our method more accurately places anatomical landmarks such as the nose, teeth, and epiglottis.

4 Discussion

Our novel 3D DIR model relies on a self-attention backbone to compute the DVF as a representation of the relationship between two inputs. The major advantage of self-attention is its ability to model long range dependencies in the input. Since our input is the combination of moving and target images, understanding long range relationships within this amalgamation is crucial for appropriate deformation prediction. Additionally, attention provides a unique way to aggregate information from the input. Instead of learning a series of static filters, attention dynamically computes filters based on what it sees in the input[12]. This is due to its elegant system of keys and queries. This system is extended to an additional spatial advantage. Whereas convolution kernels spatially represent relationships through relative value placement in their kernels and layer down-sampling, attention with positional encoding learns dynamic representations of relative positional encoding across the entire input and output[12]. This allows for greater flexibility for the network to learn long-range spatial relationships prominent in the head and neck anatomy.

We harnessed these advantages to create the first fully attention-based DIR network. We showed that it performs comparatively or better than a selection of current DIR solutions. Notably, it performed well in *both* an inter- and intra-patient registration setting despite their somewhat conflicting goals for a single algorithm to manage. These results suggest an increased robustness in our network to both large and small distortions. We contend that this is because our network better learns matching points at large distances in its input, which is a unique feature of attention-based networks.

While our network exhibited strong results in inter- and intra- patient DIR, we recognize that this is an early implementation of a promising technique. Many design choices of this network could be further optimized for better results. Our proposed method was compared to more established techniques which have been extensively optimized. With further research into attention-based DIR, the exciting advantages of this technique can lead to a new paradigm in medical 3D deformable image registration.

5 Conclusion

A novel self-attention based deformable image registration model was developed for 3D medical imaging. We leveraged axial factorization to overcome the computational complexity of attention networks and to allow us to apply them to fully 3D input. This work opens the door for further applications of attention techniques in medical DIR.

References

[1] Shi L, Chen Q, Barley S, et al. Benchmarking of Deformable Image Registration for Multiple Anatomic Sites Using Digital Data Sets With Ground-Truth Deformation Vector Fields. *Pract Radiat Oncol*. 2021;11(5):404-414. doi:10.1016/j.prro.2021.02.012

[2] de Vos BD, Berendsen FF, Viergever MA, Staring M, Išgum I. End-to-End Unsupervised Deformable Image Registration with a Convolutional Neural Network. In: Vol 1.; 2017:204-212. doi:10.1007/978-3-319-67558-9_24

[3] Balakrishnan G, Zhao A, Sabuncu MR, Dalca A V., Guttag J. An Unsupervised Learning Model for Deformable Medical Image Registration. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE; 2018:9252-9260. doi:10.1109/CVPR.2018.00964

[4] Dalca A V., Balakrishnan G, Guttag J, Sabuncu MR. Unsupervised learning of probabilistic diffeomorphic registration for images and surfaces. Med Image Anal. 2019; 57:226-236. doi:10.1016/j.media.2019.07.006

[5] Balakrishnan G, Zhao A, Sabuncu MR, Guttag J, Dalca A V. VoxelMorph: A Learning Framework for Deformable Medical Image Registration. IEEE Trans Med Imaging. 2019;38(8):1788-1800. doi:10.1109/TMI.2019.2897538

[6] Neylon J, Min Y, Low DA, Santhanam A. A neural network approach for fast, automated quantification of DIR performance. Med Phys. 2017;44(8):4126-4138. doi:10.1002/mp.12321

[7] Yang X, Kwitt R, Niethammer M. Fast predictive image registration. Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics). 2016;10008 LNCS:48-57. doi:10.1007/978-3-319-46976-8_6

[8] Kuppala K, Banda S, Barige TR. An overview of deep learning methods for image registration with focus on feature-based approaches. Int J Image Data Fusion. 2020;00(00):1-23. doi:10.1080/19479832.2019.1707720

[9] Zhu X, Ding M, Huang T, Jin X, Zhang X. PCANet-based structural representation for nonrigid multimodal medical image registration. Sensors (Switzerland). 2018;18(5). doi:10.3390/s18051477

[10] Haskins G, Kruger U, Yan P. Deep learning in medical image registration: a survey. Mach Vis Appl. 2020;31(1):1-18. doi:10.1007/s00138-020-01060-x

[11] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. Adv Neural Inf Process Syst. 2017;2017-Decem(Nips):5999-6009. doi: 10.48550/arXiv.1706.03762

[12] Wang H, Zhu Y, Green B, Adam H, Yuille A, Chen L-C. Axial-DeepLab: Stand-Alone Axial-Attention for Panoptic Segmentation. In: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Vol 12349 LNCS; 2020:108-126. doi:10.1007/978-3-030-58548-8_7

[13] Shamonin D, Bron E, Lelieveldt B, Smits M, Klein S, Staring M. Fast parallel image registration on CPU and GPU for diagnostic classification of Alzheimer's disease. Front Neuroinform. 2014;7(January):1-15. doi:10.3389/fninf.2013.00050

[14] Klein S, Staring M, Murphy K, Viergever MA, Pluim JPW. elastix : A Toolbox for Intensity-Based Medical Image Registration. 2010;29(1):196-205. doi: 10.1109/TMI.2009.2035616

[15] Clark K, Vendt B, Smith K, et al. The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository. J Digit Imaging. 2013;26(6):1045-1057. doi:10.1007/s10278-013-9622-7

[16] Hui Z, Li J, Wang X, Gao X. Image fine-grained inpainting. arXiv. 2020;(2):1-11. doi: 10.48550/arXiv.2002.02609

[17] McKenzie EM, Tong N, Ruan D, Cao M, Chin RK, Sheng K. Using neural networks to extend cropped medical images for deformable registration among images with differing scan extents. Med Phys. 2021;48(8):4459-4471. doi:10.1002/mp.15039

[18] Tong N, Gou S, Yang S, Ruan D, Sheng K. Fully automatic multi-organ segmentation for head and neck cancer radiotherapy using shape representation model constrained fully convolutional neural networks. *Med Phys*. 2018;45(10):4558-4567. doi:10.1002/mp.13147

The Prognostic Value of Foundational Radiomics and Clinical Features in Locally Advanced Non-Small Cell Lung Cancer in NRG Oncology Trial RTOG 0617

Taman Upadhaya, Elizabeth M. McKenzie, Indrin J. Chetty and Katelyn M. Atkins

Department of Radiation Oncology, Cedars-Sinai Medical Center, Los Angeles, USA

Abstract

This study investigated the feasibility of the foundational radiomics technique on the downstream task of predicting overall survival outcomes in patients with locally advanced non-small cell lung cancer (NSCLC). In an analysis of 449 patients from NRG Oncology/Radiation Therapy Oncology Group (ROTG) 0617, univariate cox regression identified 241 foundational radiomic and 19 clinical features as significant (p-value≤0.05) independent predictors of survival. Subsequently, the performance of popular predictive models, such as Least Absolute Shrinkage and Selection Operator (LASSO) in conjunction with a generalized linear model (GLM), Random Forest (RF), and Support Vector Machine (SVM) were evaluated using 5-repeat 10-fold cross-validation (n=278) and in an unseen independent test set (n=140), yielding an AUC of 0.67 (95% confidence interval [CI], 0.64-0.70) and 0.64 (95% CI, 0.54-0.73) based on foundational radiomics model solely employing SVM and 0.66 (95% CI, 0.73-0.68) and 0.67 (95% CI, 0.58-0.76) based on the clinical features model solely employing LASSO/GLM, respectively. When the foundational radiomics features were combined the clinical features, the AUC using a RF model increased to 0.74 (95% CI, 0.72-0.77) for the validation datasets and 0.67 (95% CI, 0.58-0.76) for the unseen independent dataset. This suggests that radiomics signatures provide complimentary information to clinical features for prediction of survival outcomes in patients with locally advanced NSCLC.

1. Introduction

A foundational model is an artificial intelligence (AI) model that uses a self-supervised learning technique to build the deep learning model on a vast amount of unlabeled data that can be applied to a wide range of downstream use cases. Recently, foundational models have demonstrated significantly enhanced performance in language and vision, driving successes in applications of ChatGPT, BERT, and CLIP in natural language processing (NLP), and SimCLR, YOLO, DINO, SAM and DALLE in computer vision [1, 2]. Furthermore, leveraging a foundational model for downstream tasks helps address the challenge of both limited and labeled data for modeling. This is particularly true in the domain of medical image analysis as colleting a substantial amount of data for modeling can be challenging. Thus, a foundational model capable of learning from diverse datasets through self-supervised learning, independent of specific outcomes, and capable of predicting various downstream tasks is highly desirable. To date, the only known foundational model study for 3D medical imaging was published by Pai *et el.,* [3] which was trained using a comprehensive dataset of 11,467 radiographic cancers on computed tomography (CT) imaging for various clinical relevance tasks. While there are many studies related to hand crafted radiomics features (intensity, texture, morphological) and deep radiomics (based on deep learning) the research for foundation radiomics (based on a foundational model) for prediction of oncologic outcomes is largely unexplored [4, 5].

In this work we exploit a pretrained foundational model from Pai *et el.,* for the downstream task of predicting 2-year overall survival in patients with locally advanced NSCLC. Our approach utilizes the foundational model as a feature extractor with various minor and major modification. In addition, we evaluated the complementary value of foundational radiomics features with clinical features using popular machine learning (ML) classifier namely generalized linear model with LASSO penalty, Random Forest (RF) and Support Vector Machine (SVM) for their performance.

2. Materials and Methods

2.1. Clinical cohort data

The NRG Oncology/Radiation Therapy Oncology Group (ROTG) 0617 data set from the National Clinical Trials Network (NCTN) was utilized. [6]. All relevant CT images, RT plans and clinical features such as Zubrod performance status, tumor histology, age, race, gender, smoking history, pneumonitis, esophagitis, chemotherapy, cetuximab, and RT dose groups (60 vs. 74 Gy) were utilized as per the NCTN dataset [7]. Additionally, dose statistics from heart, left anterior descending coronary artery [8], planning tumor volume (PTV), clinical tumor volume CTV), and ipsilateral lung structures were included in the analysis. Survival time and vital status were used to estimate 2-year overall survival (OS).

2.2. Foundational radiomics extraction

We employed a pretrained foundational model developed by Pai *et el.,* which was pre-trained on 11,467 diverse annotated cancers identified on computed tomography (CT) imaging. This foundational model is the modified version of the original SimCLR framework to effectively represent the nature of medical images [9]. It adopts ResNet50 architecture as encoder with the contrastive loss function in a self-supervised manner to learn to identify positive pairs of 3D patches surrounding the lesion and negative pairs, randomly sampled 3D patches from the rest of the scan [3, 10]. Once the model is trained, it can be used on various

downstream tasks, such as demonstrated in the original study. Furthermore, for our specific task, even though we used the same pretrained model, we modified the image transformation step by removing the seed-based crop to lesion-based crop to accommodate cancers of various sizes and multiple lesions within the region of interest. We then used these regions of interest from the CT images to extract 4098 features for our downstream task of predicting 2-year OS, which was termed features foundational radiomics.

2.3 Feature selection and machine learning

Univariate cox regression and correlation were applied to identify significant and independent predictors of 2-year OS. Subsequently, popular predictive models, such as LASSO, RF, and SVM were trained. Pre-processing for removing zero-variance features and centering (mean of zero and standard deviation of one) was applied before feeding features to the algorithms. Class imbalance was handled using the Synthetic Minority Over-sampling Technique (SMOTE) and tuning of the parameters for SVM, RF and GLM/LASSO were done using 5-repeat 10-fold cross-validation. All independent predictors were rank based on the importance score from LASSO, RF and SVM and then stepwise forward feature selection based on the previous ranking is used to select the subset of features minimizing the validation error. The final model for testing was chosen based on the highest area under the curve (AUC). For technical details, we direct readers to our prior publications [11, 12]. Three distinct experimental setups were carried out: (i) in the first approach, all RTOG 0617 patients (n=449) were included (for both high and low dose treatment arms), and after preprocessing split into training-testing cohorts (n=278 training, n=140 testing); (ii) the second approach was limited to patients from high dose arm (n=184 training, n=46 testing); and (iii) the third approach involved was limited to patients from low dose arm (n=150 patients training, n=38 testing).

3. Results

The univariate Cox analysis identified 822 foundational radiomics features (hazard ratios ranging from 0.14 to 3.35, and corresponding p-values ≤0.05) and 26 clinical features (hazard ratios ranging from 0.56 to 6.26 and p-values ≤0.05), as significant. Among them, 241 foundational radiomics features, along with 19 clinical features, were identified as independent predictors of survival.

Figure 1 illustrates the comprehensive performance results derived from our experiments. In the initial set of experiments, which incorporated all patients regardless of radiation dose, the model leveraging solely 12 foundational radiomics features achieved an AUC of 0.67 (95% confidence interval [CI], 0.64-0.70) in cross-validation and 0.64 (95% CI, 0.54-0.73) in unseen independent testing

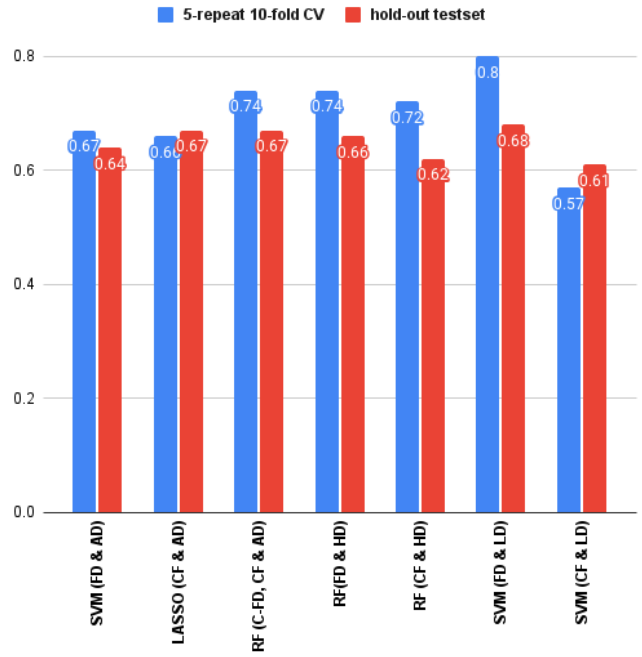


Figure 1. Performance of foundational radiomics feature (FD) based model, clinical feature (CF) based model and their combination (C-FD and CF) using all patients irrespective of dose (AD), as well as separately within high dose (HD) arm and low dose (LD) arm as assessed by 5-repeat 10-fold cross validation and in unseen independent test set.

using the SVM algorithm. In contrast, the model relying solely on 6 clinical features achieved an AUC of 0.66 (95% CI, 0.73-0.68) in cross-validation and 0.67 (95% CI, 0.58-0.76) in unseen independent testing using the LASSO algorithm. The combination of 7 foundational radiomics features and 4 clinical features (grade 3 esophagitis, PTV volume, ipsilateral lung mean dose, v5 heart and v5 lung percentage) produced the highest AUC of 0.74 (12 features) (95% CI, 0.72-0.77) in cross-validation and 0.67 (95% CI, 0.58-0.76) in unseen independent testing RF algorithm.

Similarly, in the subsequent set of experiments involving patients from the high dose arm only, the model based solely on 5 foundational radiomics features achieved the highest AUC of 0.74 (95% CI, 0.70-0.767) in cross-validation and 0.66 (95% CI, 0.50-0.83) in unseen independent testing using the RF algorithm. Conversely, the model relying solely on 6 clinical features (age, PTV log volume, chemotherapy, Left Anterior Descending Coronary Artery (LAD) mean dose, v5 lung and V95 PTV percentage) achieved an AUC of 0.72 (95% CI, 0.69-0.75) in cross-validation and 0.62 (95% CI, 0.45-0.78) in unseen independent testing, once again using the RF algorithm. However, the integration of foundational radiomics and clinical features did not enhance the AUC performance.

In the final experiment which incorporated patients from the low dose arm only, the model relying solely on 12 foundational radiomics achieved the highest AUC of 0.80 (95% CI, 0.77-0.83) in cross-validation and 0.68 (95% CI, 0.50-0.85) in unseen independent testing using the SVM algorithm. Meanwhile, the model relying solely on 9 clinical features (PTV log volume, ipsilateral lung mean dose, grade5 toxicity, LAD mean dose, heart mean dose, heart median dose, V5 and V20 lung percentage, V60 esophagus) achieved the lowest performance, with an AUC of 0.57 (95% CI, 0.53-0.61) in cross-validation and 0.61 (95% CI, 0.42-0.79) in unseen independent testing, once again using the SVM algorithm. Similar to above, the combination of foundational radiomics and clinical features did not enhance the AUC performance.

4. Discussion

In this study, we demonstrated that prediction models for OS can be built using foundational radiomics with ML techniques. In our first experiments involving all patients irrespective of dose received, the model incorporating both foundational radiomics and clinical features outperformed the model replying solely on foundational radiomics features or clinical features. Overall, all methods led to models with modest AUC (around or above 0.66). As we shifted our approach from dose-agnostic to those receiving high vs low dose treatments separately, we observed noticeable variations in model performance, worth of further study. For the experiment involving only patients from the high-dose arm, the foundational radiomics-based model solely achieved a performance of 0.74 in cross-validation and 0.66 in the unseen test set, outperforming the clinical feature-based model. Similarly in the experiment involving only patients from the low dose arm, the foundational radiomics-based model reached a performance of 0.80 in cross-validation and 0.68 in unseen independent test set, again outperforming the clinical features-based model. In both analyses by dose arm, the combination of foundational radiomics and clinical features did not improve the model performance.

To our best knowledge, this study is the first to investigate the prognostic and complementary value of foundational radiomics with clinical features using various popular machine learning algorithms. To date, only one study has been reported using foundational radiomics for prognostication by Pai et al., however they only included one linear classifier. Our study achieved similar performance to the previous study using foundational radiomics, though our approach uniquely explored the additional benefits of combining these foundational radiomics features with clinical features.

Our study has notable limitations. In the model development, we did not perform fine-tuning on our data set that could potentially improve the performance. However, it is important to note as these foundational models get bigger- hosting, building, and training a foundational is an incredibly expensive endeavor—often costing millions of dollars. Therefore, investigating downstream tasks of models such as these would be crucial. Our analysis also did not include handcrafted radiomics features. We are actively working on expanding our work to fine tune the foundational model, include handcrafted radiomics features and additional anatomical structures such as the heart and lung, which could provide further insights into tissue characteristics, as our ML framework is easily amenable to incorporation of additional data [13].

5. Conclusion

In conclusion, a self-supervised foundational model based on radiomics signatures showed promising prediction of 2-year OS for patients with locally advanced NSCLC from RTOG 0617. The combination of foundational radiomics with conventional clinical features provided even better prediction, suggesting the complimentary value of foundational radiomics in outcomes modeling.

References

[1] Wang, H., Li, J., Wu, H., Hovy, E. and Sun, Y., 2022. Pre-trained language models and their applications. Engineering. DOI: https://doi.org/10.1016/j.eng.2022.04.024

[2] Zhang, J., Huang, J., Jin, S. and Lu, S., 2023. Vision-language models for vision tasks: A survey. arXiv preprint arXiv:2304.00685. DOI: https://doi.org/10.48550/arXiv.2304.00685

[3] Pai, S., Bontempi, D., Prudente, V., Hadzic, I., Sokač, M., Chaunzwa, T.L., Bernatz, S., Hosny, A., Mak, R.H., Birkbak, N.J. and Aerts, H.J., 2023. Foundation models for quantitative biomarker discovery in cancer imaging. medRxiv.

[4] Aerts, H.J., Velazquez, E.R., Leijenaar, R.T., Parmar, C., Grossmann, P., Carvalho, S., Bussink, J., Monshouwer, R., Haibe-Kains, B., Rietveld, D. and Hoebers, F., 2014. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nature communications, 5(1), p.4006.

[5] Afshar, P., Mohammadi, A., Plataniotis, K.N., Oikonomou, A. and Benali, H., 2019. From handcrafted to deep-learning-based cancer radiomics: challenges and opportunities. IEEE Signal Processing Magazine, 36(4), pp.132-160.

[6] Bradley, J.D., Hu, C., Komaki, R.R., Masters, G.A., Blumenschein, G.R., Schild, S.E., Bogart, J.A., Forster, K.M., Magliocco, A.M., Kavadi, V.S. and Narayan, S., 2020. Long-term results of NRG oncology RTOG 0617: standard-versus high-dose chemoradiotherapy with or without cetuximab for unresectable stage III non–small-cell lung cancer. Journal of Clinical Oncology, 38(7), p.706.

[7] Clark, K., Vendt, B., Smith, K., Freymann, J., Kirby, J., Koppel, P., Moore, S., Phillips, S., Maffitt, D., Pringle, M. and Tarbox, L., 2013. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. Journal of digital imaging, 26, pp.1045-1057.

[8] McKenzie, E., Zakariaee, R., Zhang, S., Guthier, C.V., Hakimian, B., Mirhadi, A.J., Kamrava, M., Padda, S., Lewis, J.H., Nikolova, A. and Mak, R.H., 2022. Left Anterior Descending Coronary Artery Radiation Dose Association With All-Cause Mortality in NRG Oncology Trial RTOG 0617. International Journal of Radiation Oncology, Biology, Physics, 114(3), p.S58.

[9] Chen, T., Kornblith, S., Norouzi, M. and Hinton, G., 2020, November. A simple framework for contrastive learning of visual representations. In International conference on machine learning (pp. 1597-1607). PMLR.

[10] Uemura, T., Näppi, J.J., Hironaka, T., Kim, H. and Yoshida, H., 2020, March. Comparative performance of 3D-DenseNet, 3D-ResNet, and 3D-VGG models in polyp detection for CT colonography. In Medical Imaging 2020: Computer-Aided Diagnosis (Vol. 11314, pp. 736-741). SPIE.

[11] Morin, O., Vallières, M., Braunstein, S., Ginart, J.B., Upadhaya, T., Woodruff, H.C., Zwanenburg, A., Chatterjee, A., Villanueva-Meyer, J.E., Valdes, G. and Chen, W., 2021. An artificial intelligence framework integrating longitudinal electronic health records with real-world data enables continuous pan-cancer prognostication. Nature Cancer, 2(7), pp.709-722.

[12] Upadhaya, T., Vallières, M., Chatterjee, A., Lucia, F., Bonaffini, P.A., Masson, I., Mervoyer, A., Reinhold, C., Schick, U., Seuntjens, J. and Le Rest, C.C., 2018. Comparison of radiomics models built through machine learning in a multicentric context with independent testing: identical data, similar algorithms, different methodologies. IEEE Transactions on Radiation and Plasma Medical Sciences, 3(2), pp.192-200.

[13] Uemura, T., Näppi, J.J., Hironaka, T., Kim, H. and Yoshida, H., 2020, McKenzie, E., Zhang, S., Zakariaee, R., Guthier, C.V., Hakimian, B., Mirhadi, A., Kamrava, M., Padda, S.K., Lewis, J.H., Nikolova, A. and Mak, R.H., 2023. Left anterior descending coronary artery radiation dose association with all-cause mortality in NRG oncology trial RTOG 0617. International Journal of Radiation Oncology* Biology* Physics, 115(5), pp.1138-1143.

Exploring How Prostate Cancer Treatment Choices Affect Radiomic Features: A Look at Healthy Tissue vs. Dominant Lesion in MR-Guided Radiotherapy

Aaron G. Rankin[1], Martin Swinton[1,2], Priya Maitre[2,3], Jim Zhong[4,5], Angela Davey[1], Marianne Aznar[1], Ananya Choudhury[1,2], and Alan McWilliam[1]

[1]Advanved Radiotherapy Group, Division of Cancer Sciences, School of Medical Sciences, University of Manchester, Manchester, United Kingdom
[2]Department of Clinical Oncology, The Christie NHSFT, Manchester, United Kingdom
[3]Tata Memorial Centre, Homi Bhabha National University, Mumbai
[4]Department of Diagnostic and Interventional Radiology, Leeds Teaching Hospitals NHS Trust, UK
[5]Leeds Institute of Medical Research, University of Leeds, UK

Abstract In this work we contrast feature values acquired from different treatment groups for patients receiving MR-guided radiotherapy as part of treatment. A total of 69 patients with localised prostate cancer were included in this study. We explored the impact of different fractionation schemes and androgen deprivation therapy on radiomic feature values acquired from on-treatment T2W-images. Using conventional radiomic approaches, by assessing robustness against contour variation and correlation to volume, we identified 41 features out of 95 to be redundant across treatment groups. An in-house longitudinal methodology for determining representative feature trajectories was used to identify 8 features that were then used for further analysis in assessing relations between feature values within the dominant lesion and entire prostate.

1 Introduction

With the rise in MR guided radiotherapy (MRgRT) there is an increasing amount of rich longitudinal MR imaging data from patient on treatment. There is now an unmet need for models that can fully utilise this rich longitudinal MR data to provide useful clinical insight in to a patient's response to radiotherapy. Radiomic features provide quantitative metrics to derive information relating to the texture, shape and spatial heterogeneity of different tissues which could be used in the development of potential novel image-based biomarkers.

For patients with prostate cancer, radiotherapy is moving to stereotactic ablative radiotherapy (SABR) as standard of care. This is enabled by advances in image-guided radiotherapy (including MRgRT) enabling precision treatments and thus enhanced treatment individualisation For MRgRT, tracking changes in the dominate intraprostatic legion (DIPL) in the prostate becomes possible. Radiomics may allow changes to the DIPL to become discernible, informing treatment escalation if needed. These changes may inform if longer 20 fraction treatments are beneficial in some men over the shorter SABR regimes.

In this work, we investigated the influence of different treatment pathways on radiomic feature values for patients receiving MRgRT for prostate cancer. Our primary aims were to:

- discern if there were any significant differences in the distribution of feature values between each treatment group.
- investigate the trajectory of radiomic features for each ROI during the course of treatment.

We achieved this by utilising pre-existing feature reduction techniques and using our in-house methodology for handling longitudinal changes in radiomic features.

2 Materials and Methods

2.1 Patient Data

In total, 69 patients (median age of 72.5 years and range of 57-77 years) were included in this study. All patients had a diagnosis of localised prostate cancer (Cambridge Prognostic Group 2-4) and received radical MRgRT delivered on a 1.5T MR Linac. Treatment group 1 received androgen deprivation therapy (ADT) and 60 Gy in 20 fractions over 4 weeks of prostate radiotherapy, group 2 received prostate SABR (36.25 Gy in 5 fractions on alternative days over 2 weeks) and ADT and group 3 received SABR alone without ADT.

A 3D T2W-image was collected at each fraction for SABR patients and at fractions 1, 5, 10, 15, 20 for patients receiving 20 fractions. Each image was acquired with an echo time of 278 ms, repetition time 1535 ms, voxel spacing 0.83 x 0.83 mm and slice thickness 2 mm.

The entire prostate and the DIPL was contoured on the first fraction image of each patient by at least one of three clinicians; two clinical oncologists specialising in urology (Observer A and B) and a radiologist (Observer C). A full breakdown of the number of patients in each treatment group and the number of contours provided by each observer is shown in Table 1. The feature values from Observer A were used to conduct all analysis - feature values from Observer B and C were used for the sole purpose of determining robustness against contour variation, as discussed in Section 2.3.2.

Treatment Group	Obs A	Obs B	Obs C	Total
20 Frac + ADT	25	10	4	25
SABR + ADT	23	10	5	24
SABR + no ADT	20	13	5	20

Table 1: Number of patients included in each treatment group with both the whole and DIPL contoured. Patients with SABR received 36.25 Gy in 5 fractions, whilst patients receiving 20 fractions were treated with 60 Gy. Patients with ADT at least 3 months prior to the first fraction of radiotherapy. Obs refers to the clinician that provided the contour of the DIPL mask.

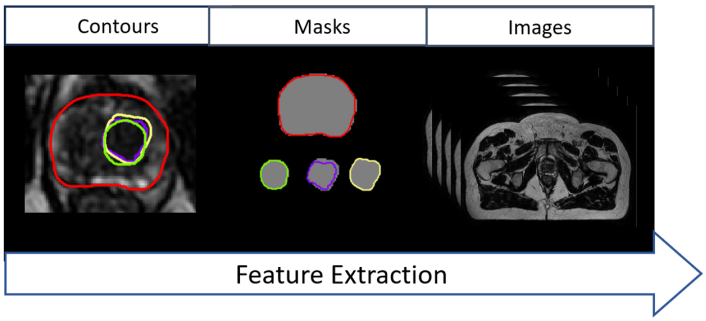


Figure 1: General radiomics feature extraction pipeline. An observer delineates the DIPL contours are generated. The masks from those contours are then used to extract radiomic features from each of the 5 T2W-images per patient acquired during treatment.

2.2 Feature Extraction

Prior to feature extraction, images were rigidly registered and to the first fraction image. Histogram-matching was also per-formed to standardise each image intensity distribution to the first fraction image for each patient. PyRadiomics (v3.1.0) was then used to extract 95 features from each available image with a fixed bin size of 25 and no voxel resampling. Following extraction, feature values were rescaled to be within [0,1] in order to improve comparability.

2.3 Feature Robustness

Multiple studies have shown the importance of feature reduction techniques in radiomics analysis. Identifying potentially useful clinical features involves removing any redundant features. Due to the high dimensionality of radiomics features, a number of feature reduction techniques are needed to determine the most robust and representative features.

2.3.1 Volume Correlation

Traveso et al. have shown that certain radiomic feature values have a strong dependence on the volume of the mask used to extract features from [1]. Feature values at Fraction 1 were correlated to volume values using Spearman rank correlation, ρ.

If $|\rho| \geq 0.6$, features were deemed to be strongly correlated to volume. Features that were identified as strongly correlated to volume in 2 out of 3 treatment groups were removed.

2.3.2 Inter-Observer Variability

The intra-class coefficient, ICC, provides a measure of agreement between different observers. Many studies have shown the importance of test-retest analysis for radiomic feature values since we aim to ensure robustness against variations introduced by different observers [2].

In this work, the ICC was used to measure the consistency of feature values calculated from a subset of patients contoured by all 3 observers. Features with an ICC < 0.75 were removed from further analysis.

2.3.3 Representative Feature Trajectories

Following the feature reduction steps above, we aimed to identify features that were the most representative when accounting for the full trajectory over treatment. Rather than look at single timepoint values or the difference between two timepoints (i.e. delta-radiomics values from fraction 1 to 5), we compared how different radiomic feature values evolved over the course of treatment.

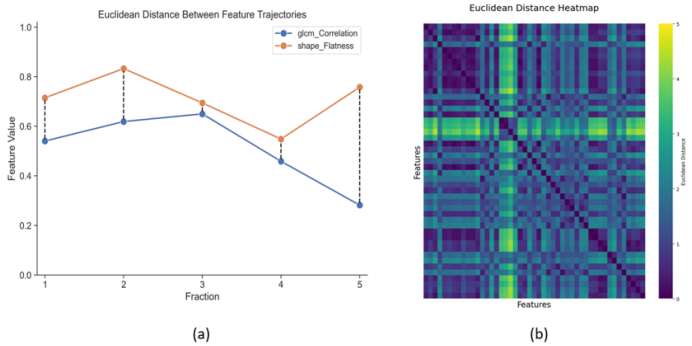


Figure 2: How feature trajectories are used to group together similar features. In (a) the Euclidean distance is used to measure the spatial separation between feature values at each timepoint. These values can then be visualised, like in (b), to see trends across feature sets. These values are then used in our hierarchical clustering algorithm.

For each patient, the Euclidean distance between feature-pair trajectories was computed. These distance values were then used to group features in to similar trajectories based on hierarchical clustering. Clusters of more than 8 features were futher broken down in to sub-clusters and clusters with less than 3 features were discarded as they deemed unstable.

Within these clusters, we performed a Pearson correlation and identified the trajectory with the highest mean correlation to all others. This feature was then passed through to the ranking stage.

The most representative features were then identified from the top 10% of features most commonly ranked across the

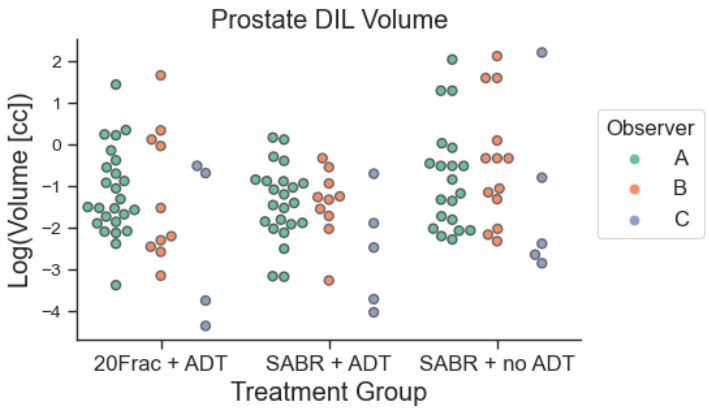


Figure 3: Swarm plot of the log(volume) of each DIPL mask for treatment groups. Patients without ADT appear to have larger volumes. Also shows the degree of consistency between different observers in relation to the contoured volumes.

patient group. Features found to be robust in 2 out of 3 treatment groups were used for further anaylsis.

2.4 Prostate vs DIPL

After identifying the most representative feature trajectories, we used the chosen features to investigate the relationship between feature values calculated from the DIPL and entire prostate mask. At the first and last fraction of treatment, we used the Pearson correlation coefficient to quantify the linear relationship between features calculated on the DIPL and the prostate. We hypothesise that due to the near-uniform treatment dose to the prostate during RT, that features within the prostate and DIPL would become increasingly correlated.

3 Results

3.1 Volume Analysis

The volume of the DIPL contours was compared across treatment groups and observers. Fig 3 shows how the distribution of the DIPL volume masks differ. For the SABR + no ADT group, the mean DIPL volume was 1.42 cm^3 with a standard deviation of 2.41 cm^3. When comparing to patients who did receive SABR + ADT, the mean volume is 0.31 cm^3 with a standard deviation of 0.28 cm^3. For the 20 fractions group that was found to be 0.63 cm^3 and 1.05 cm^3

3.2 Feature Robustness

A total of 13 patients were used to determine ICC values for each feature. Of the 93 features analysed, 30 features were found to be not robust against IOV, leaving 63 features. Upon calculating the correlation to volume, 20 features were deemed strongly correlated. Of those 20, 9 were also sensitive to IOV - totalling 41 redundant features.

The remaining 52 features were then used to determine the most representative feature trajectories over

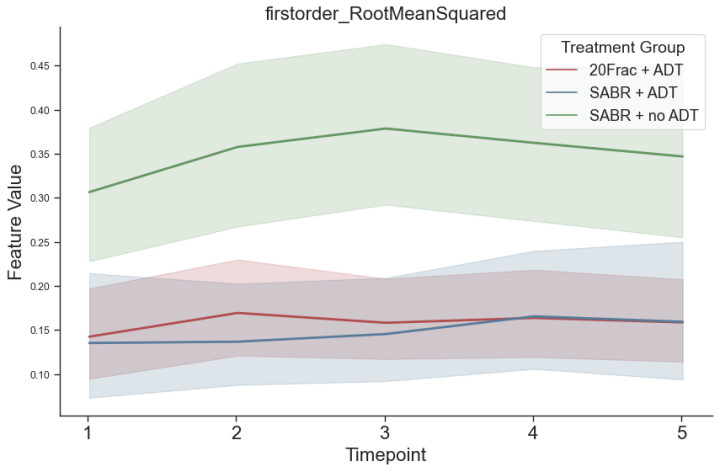


Figure 4: Signal plot of $firstorder-RootMeanSquared$ for each treatment group. Identified as robust in each treatment group. Solid represents the mean and shaded areas denote 95% CI. Clearly shown is the separation between the SABR + no ADT group.

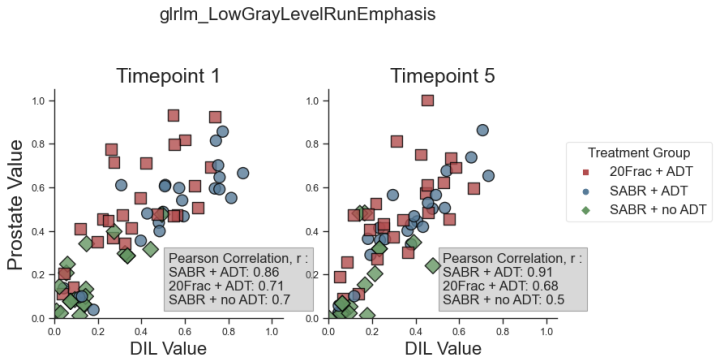


Figure 5: Distribution of feature values calculated at each timepoint. Pearson correlation was used to quantify the linear relationship between the values from the prostate and DIPL.

the 5 timepoints of treatment. Following the clustering process, we identified 4 features selected by each of the treatment groups: $firstorder-RootMeanSquared$, $glcm-JointEntropy$, $glrlm-LowGrayLevelEmphasis$ and $gldm-GrayLevelVariance$. A further 4 features were common in 2 out of 3 feature groups and were carried on to the next stage of the analysis.

3.3 Prostate vs DIPL

In the 20 Fraction + ADT only 2 out of the 8 features showed an increase in the Pearson correlation coefficient. Conversely 6 of 8 features increased in correlation in the SABR + ADT group and 4 features in the SABR + no ADT group showed an increase in Pearson correlation coefficient. Figure 5 shows an example of the feature distributions at the first and last fractions of treatment.

4 Discussion

We were able to identify features robust against observer variations and volume correlation for feature values calcu-

lated between different radiotherapy treatment schemes for prostate cancer. Despite this though, there are stark differences in the distribution of features as shown by Figs 5 and 4. There is a clear distinction between feature values from the SABR + no ADT group and the other two treatment groups. ADT is known to shrink and homogenise signal within the prostate and so this may be the driving force between the relative differences [3].

Another area where the impact of ADT is visible is the volume of the DIPL contours. The mean DIPL contour was over 1 cm^3 larger than contours from the other 2 treatment groups too. Further analysis is required to discern if the pronounced differences is influenced by volume. Another factor when considering the volume of the masks, is the difficulty in generating contours of the dominant lesion, given their small size. This may also be a factor in making them more susceptible to noise.

When comparing feature distributions between the two groups that received ADT, as in Fig 4, we found that there was no clear distinction between patients receiving 5 and 20 fractions of radiotherapy, despite the treatment duration lasting 20 days longer for 20 fraction patients.

Here we have provided a framework for longitudinal analysis of radiomic features acquired from on-treatment images. Future work could involve correlating clinical endpoints and to robust features in order to help further treatment personalisation for patients.

5 Conclusion

To conclude, we showed the strong impact that ADT has on feature values and contrasted the differences in distributions across patients receiving different courses of treatment for prostate cancer. We have also demonstrated a method to identify robust radiomic features across treatment groups and find representative feature trajectories from routine images acquired as part of MRgRT.

References

[1] A. Traverso, M. Kazmierski, I. Zhovannik, et al. “Machine learning helps identifying volume-confounding effects in radiomics”. *Physica Medica* 71 (Mar. 2020), pp. 24–30. DOI: 10.1016/j.ejmp.2020.02.010.

[2] C. Xue, J. Yuan, D. M. Poon, et al. “Reliability of MRI radiomics features in mr-guided radiotherapy for prostate cancer: Repeatability, reproducibility, and within-subject agreement”. *Medical Physics* 48.11 (Sept. 2021), pp. 6976–6986. DOI: 10.1002/mp.15232.

[3] C. Gunner, A. Gulamhusein, and D. J. Rosario. “The modern role of androgen deprivation therapy in the management of localised and locally advanced prostate cancer”. *Journal of Clinical Urology* 9.2$_{s}$*uppl* (Dec. 2016), pp. 24–29. DOI: 10.1177/2051415816654048.

On Computational Challenges in Proton Therapy Treatment Planning

Mohammadhossein Mohammadisiahroudi[1], Wei Zou[2], Lei Dong[2], Yuriy Zinchenko[3], and Tamás Terlaky[2]

[1]Department of Industrial and System Engineering, Lehigh University, Bethlehem, PA, USA
[2]Department of Radiation Oncology, University of Pennsylvania, Philadelphia, PA, USA
[3]Department of Mathematics and Statistics, University of Calgary, Calgary, Canada

Abstract In radiation therapy treatment planning, finding optimal machine configurations, to get maximal coverage of the tumor and minimize damage to critical and healthy organs, is an important optimization problem. Proton Therapy is a new technology enabling high-precision treatment of the tumor close to critical organs. However, treatment planning for proton therapy comes with computational challenges arising from the large number of decision variables and the sensitivity of the treatment to the movements of the patient. In this research, we experiment with different optimization models for proton therapy and evaluate their performance. To tackle large-scale problems, we developed sparsifying and size reduction techniques, including cropping CTs, and prolonging voxels in the interior of organs. In addition, we develop an efficient successive relaxation method for the mixed integer formulation that provides an optimal plan, achieving prescribed DVH constraints. Numerical results show a high level of comformality to DVH goals.

1 Introduction

Creating and delivering an optimal treatment plan, where the target volume receives the prescribed dosage while minimizing exposure to other tissues represents the ultimate goal in radiation therapy. Achieving perfect equilibrium in this regard, however, proves to be an unattainable goal. Instead, the process necessitates making various trade-offs and developing optimization models that deliver clinically sound treatment plans. Numerous factors come into play that influence the degree to which a patient's treatment can be optimized [1]. This research delves into the complexities of optimizing a proton therapy (PT) plan, with a specific focus on challenges related to building and solving large-scale optimization models arising in PT treatment planning.
The characteristics of PT enhance the potential for achieving a treatment with minimal negative impact, primarily due to the precise concentration of dose in the characteristic Bragg peak. This is particularly relevant to the capability of administering elevated radiation doses in close proximity to normal tissue structures, which would typically be dose-limiting in conventional X-ray treatments. Additionally, the finite range of protons contributes to a much reduced integral dose to the surrounding healthy tissues [2].
A primary hurdle in the planning and optimization of radiation treatment lies in integrating prescribed dose-volume goals to ensure adequate coverage of the target and proper sparing of healthy structures. Conventional optimization models incorporating dose-volume histogram (DVH) constraints are computationally challenging. Recently, the planning process in clinical settings has largely become computerized, but the efficiency in meeting specific dose-volume requirements still heavily relies on the planner's experience. Consequently, a typical treatment planning process is resource intensive, as it necessitates numerous iterative and manual interventions from the planner [3].
For PT, efficiently building and solving conventional treatment planning optimization models seemed to be impossible in practice, because of the high number of variables to be optimized and the higher sensitivity of PT to uncertainties and motions. To be more precise, in PT optimization models, numerous beamlets are defined for each spot with specific energy levels for each field. On the other hand, using a fine grid for fully discretized problems, calculating the dose for all voxels, building the optimization model, and solving it efficiently is referred to as one of the major challenges for PT [4]. To mitigate this issue, some authors try to reduce the size of the problem by increasing grid spacing [5] or iteratively resampling from pencil beams [6]. However, such approaches impact the quality of the plan.
In this research, we employ three techniques to reduce both the number of voxels and increase efficiency of optimization:

- Cropping CTs to remove irrelevant voxels.
- Sparsifying dose-influence matrix, which is calculated by MatRad.
- Enlarging voxel size in the interior of organs at risk.

Accordingly, we use three optimization approaches to evaluate their performance for PT. In addition, a successive relaxation algorithm is developed to find an optimal plan, satisfying prescribed DVH constraints.

2 Materials and Methods

Our approach has three major parts: 1) preparing and processing data; 2) building optimization models to find optimal treatment plans; 3) methods to solve the optimization problems. Our methodology is explained in the following sections.

2.1 Data Preparation

Let V denote the set of voxels, and O the set of Organ-at-risk (OARs). The set of voxels in the target volume is denoted by

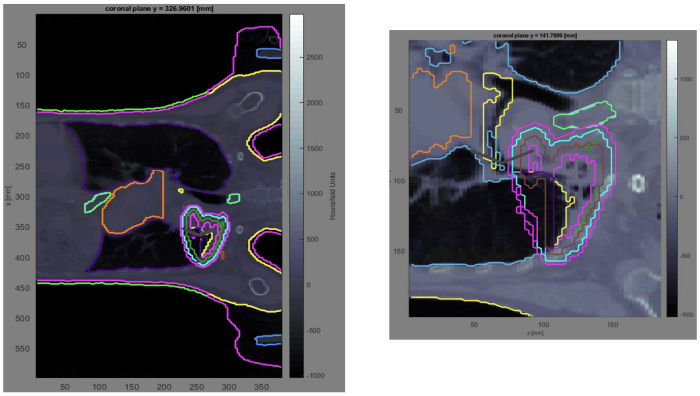


Figure 1: A sample of cropping CT for a Lung cancer case.

V_T, and the set of voxels in OAR o by V_o. Let B denote the set of beamlets. To build the optimization model, one needs to calculate the dose d_{ij} received at voxel i from the unit intensity of beam j. We use MatRad [7] for this calculation. We store the calculated dose d_{ij} in matrix $D \in \mathbb{R}^{|V|\times|B|}$, which is known as the dose-influence matrix. In treatment planning, we are interested in finding the optimal intensity of beamlets, x_j for all $j \in B$. For each beamlet, we can calculate the accumulated dose by $\mathscr{D}_i = \sum_{j\in B} d_{ij}x_j$.
MatRad and other APIs use dose calculation algorithms to calculate the dose-influence matrix in a reasonable time. Although this matrix is usually sparse, the large size of this matrix and the large number of nonzero elements make the solution of the optimization problems impossible. We use three techniques to reduce the size of this matrix.

2.1.1 CT Scans Truncation

Some parts of CT scans are not useful for treatment planning, as that part of the body is not affected by any beams. Thus, we crop parts of CT scans that a clinical expert marked as unnecessary parts. For example, Figure 1 shows how many voxels are removed by cropping a CT scan of a Lung patient.

2.1.2 Sparsification

The dose-influence matrix usually has many extremely small numbers, that create severe numerical problems. Thus, those elements in the dose-influence matrix that are less than a constant times the maximum of the corresponding column are made zero. This way the matrix gets sparser and better conditioned for plan optimization.

2.1.3 Inner Voxel Prolongation

To reduce the number of voxels, one can increase the size of voxels. However, this will affect the quality of the plan. To have a high-quality plan, we use a fine grid with appropriate voxel size. For voxels in the interior of the organs, we double the voxel size by combining adjacent voxels. In this way, the voxels on or near the boundary of organs are small and enable precise calculation for that area. On the other hand, the total number of voxels will significantly decrease as the number of interior voxels by up to 87.5%. We avoid prolongation for voxels in the target volume to have a high-quality plan.

2.2 Optimization Approaches

The general idea of treatment optimization is minimizing the deviation of dose on target from the prescribed bound, while the dose received in OARs is bounded by prescribed DVH constraints. We present three optimization approaches.

2.2.1 Linear Optimization

In the linear optimization (LO) approach, we have prescribed upper bound $T_{\max}$ and lower bound $T_{\min}$ for the dose received by the target volume. Each OAR has an upper bound MD_o.

$$\begin{aligned}
\min\ & \sum_{i\in V_T} s_i + \sum_{i\in V_T} s'_i \\
\text{s.t.}\ & \sum_{j\in B} d_{ij}x_j - s_i \le T_{\max} \text{ for all } i \in V_T \\
& \sum_{j\in B} d_{ij}x_j + s'_i \ge T_{\min} \text{ for all } i \in V_T \\
& \sum_{j\in B} d_{ij}x_j \le MD_o \text{ for all } i \in V_o \text{ and } o \in O \\
& x, s, s' \ge 0.
\end{aligned}$$

As the objective function and the constraints are linear, we have a linear optimization (LO) model. It can be solved efficiently by commercial optimization solvers, such as Gurobi.

2.2.2 Mixed-integer Optimization

The major disadvantage of the LO approach is that the quality of the plan is not good, since we use a bound on all voxels in the OAR. In clinical practice, physicians usually prescribe a dose limit to a volumetric fraction of the OARs, i.e., limit a point on the DVH curve. More precisely, for OAR o, we have (UD_o, UV_o), which means that UV_0 fraction of voxels in OAR o can NOT receive a dose higher than UD_o. To incorporate such DVH constraints, we need a mixed-integer optimization (MIO) model. Let M be the highest dose a voxel can get, and n_o is the total number of voxels in OAR o.

$$\begin{aligned}
\min\ & \sum_{i\in V_T} s_i + \sum_{i\in V_T} s'_i \\
\text{s.t.}\ & \sum_{j\in B} d_{ij}x_j - s_i \le T_{\max} \text{ for all } i \in V_T \\
& \sum_{j\in B} d_{ij}x_j + s'_i \ge T_{\min} \text{ for all } i \in V_T \\
& \sum_{j\in B} d_{ij}x_j \le UD_o + (M - UD_o)y_i \text{ for all } i \in V_o,\ o \in O \\
& \sum_{i\in V_o} y_i \le UV_o n_o \text{ for all } o \in O \\
& x, s, s' \ge 0.
\end{aligned}$$

Although the MIO formulation provides an ideal plan, there is no efficient way to solve it, as it is a large-scale mixed-integer optimization problem.

2.2.3 Second-order Cone Optimization

To bridge the LO and MIO approaches, a moment-based approach was proposed in [8]. Consider the DVH curve as a function. Let $\mathscr{D}$ be the resulting DVH curve and $\bar{\mathscr{D}}$ be the prescribed one. Let $\mu_k(f)$ be the kth moment of function f. In the moment-based approaches, the moments of the resulting DVH are bounded by the moments of the prescribed DVH.

$$\begin{aligned}
\min\ & \sum_{i\in V_T} s_i + \sum_{i\in V_T} s'_i \\
\text{s.t.}\ & \sum_{j\in B} d_{ij}x_j - s_i \le T_{\max} \text{ for all } i \in V_T \\
& \sum_{j\in B} d_{ij}x_j + s'_i \ge T_{\min} \text{ for all } i \in V_T \\
& \mu_1(d_j) \le \mu_1(\bar{d}_j) \text{ for all } i \in V_o \text{ and } o \in O \\
& \mu_2(d_j) \le \mu_2(\bar{d}_j) \text{ for all } i \in V_o \text{ and } o \in O \\
& x, s, s' \ge 0.
\end{aligned}$$

It is shown in [8] that the above model is a second-order cone optimization (SOCO) model and commercial solvers, e.g., Gurobi, can solve it efficiently.

2.3 Successive Relaxation Method

As mentioned, there are efficient solvers for LO and SOCO, but still solving MIO model is challenging. Here we develop the Successive Relaxation Method (SRM) to solve MIO model as follows. Let $m_o = [UV_o n_o]$ for all $o \in O$.

Step 1. Solve the initial linear problem as

$$\begin{aligned}
\min\ & \sum_{i\in V_T} s_i + \sum_{i\in V_T} s'_i && (1)\\
\text{s.t.}\ & \sum_{j\in B} d_{ij}x_j - s_i \le T_{\max} \text{ for all } i \in V_T && (2)\\
& \sum_{j\in B} d_{ij}x_j + s'_i \ge T_{\min} \text{ for all } i \in V_T && (3)\\
& \sum_{j\in B} d_{ij}x_j \le UD_o \text{ for all } i \in V_o \text{ and } o \in O && (4)\\
& x, s, s' \ge 0. && (5)
\end{aligned}$$

Calculate the dual variables, also called shadow prices, of constraints (4) as sp_{io} for all $i \in V_o$ and $o \in O$.

Step 2. OV_o : the list of m_o voxels with the largest shadow prices for $o \in O$, and $LV_o = V_o - OV_o$.

	Before Sparsifying	After Sparsifying
# of Nonzeros	13,475,201	6,824,686
Min Nonzeros	1.6428e-20	4.01105e-05
Max Nonzeros	4.0110e-03	4.0110e-03

Table 1: Effect of sparsification on the dimension of the data from a Lung patient.

Step 3. Solve the following linear optimization problem as

$$\begin{aligned}
\min\ & \sum_{i\in V_T} s_i + \sum_{i\in V_T} s'_i \\
\text{s.t.}\ & \sum_{j\in B} d_{ij}x_j - s_i \le T_{\max} \text{ for all } i \in V_T \\
& \sum_{j\in B} d_{ij}x_j + s'_i \ge T_{\min} \text{ for all } i \in V_T \\
& \sum_{j\in B} d_{ij}x_j \le UD_o \text{ for all } i \in LV_o \text{ and } o \in O \\
& \sum_{j\in B} d_{ij}x_j \le M \text{ for all } i \in OV_o \text{ and } o \in O \\
& x, s, s' \ge 0.
\end{aligned} \qquad (6)$$

Update sp_{io} for all $i \in V_o$, OV_o, and LV_o for $o \in O$.

Step 4. If OV_o is changed, go to **Step 3.**
Otherwise, terminate the algorithm.

We proved the convergence of this algorithm for the MIO approach of proton treatment planning. Numerical results demonstrate its high practical proficiency.

3 Results

We implement our approach in MATLAB. We use data from 7 patients. Five of them are Lung cancer patients and the other two are Brain cancer patients. Gurobi's barrier methods are used to solve the LO and SOCO models. For solving MIO models, we use the proposed SRM. Both the MIO and SOCO models are solved in minutes. Figure 2 illustrates that they provide high-quality plans.

4 Discussion

Table 1 shows that the dose-influence matrix calculated by MatRad contains many tiny elements, as small as 10^{-20}. These tiny elements significantly increase solution time and cause numerical challenges in the optimization process. Our sparsification technique resolve these issue, and Table 1 shows the effect of sparsifying dose-influence matrix on the size of the problem. Computational results show that smart sparsification also allows high-quality plans.

Table 2 shows that by cropping unnecessary parts of CTs, we can reduce the number of rows and columns by half which leads to smaller data and faster optimization. We tried to build LO and SOCO models without sparsification and CT truncation, but the state-of-the-art solvers reached out-of-memory errors even on a computer with 132 GB of memory.

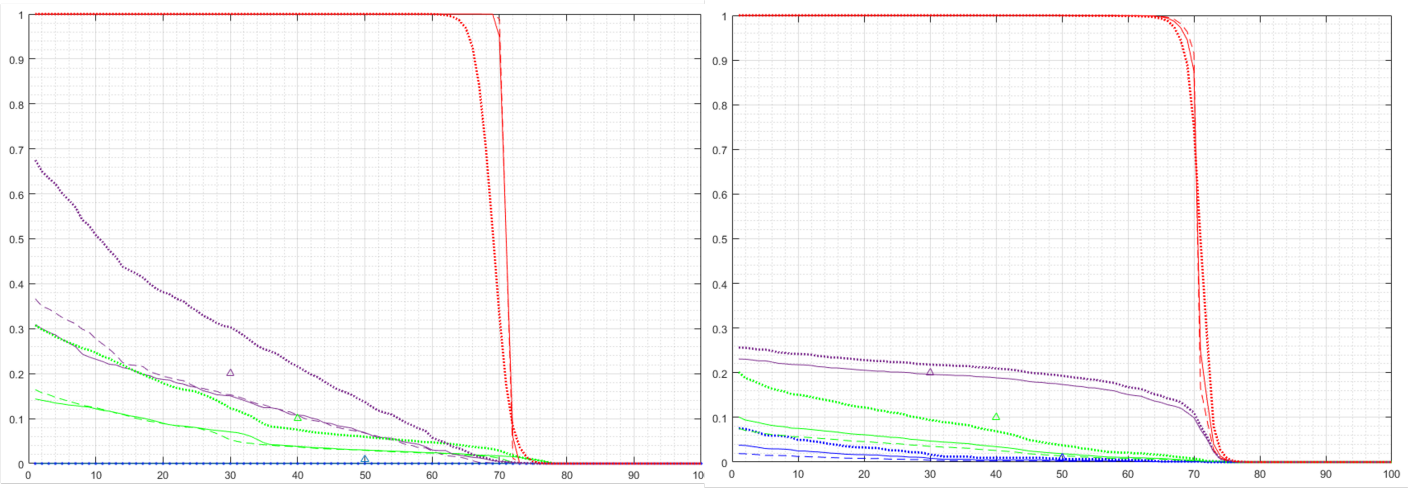


Figure 2: DVH curves of plans for two lung patients. Dotted lines: LO, Solid lines: SOCO, Dashed lines: MIO. Red: Target, Purple: Esophagus, Green: Lungs, Blue: Heart. Triangles indicate prescribed DVH constraints.

	Rows	Columns	Nonzeros
Before truncation	382,247	478,548	89,213,704
After truncation	166,835	222,403	16,717,861
Reduction	56.3541%	53.5254%	81.2608%

Table 2: Effect of cropping CTs on the dimension of the SOCO model for a Lung patient.

	Variables	Constraints	Solution time (s)
Uniform voxel size	414,088	389,395	28,872.22
Different voxel sizes	90,416	65,723	246.36
Reduction	78.16%	83.12%	99.14%

Table 3: Effect of prolonging the size of interior voxels on the dimension of the SOCO model for a Lung patient. For both cases, matrices are sparsified and CTs are cropped.

By sparsification and CT truncation, we solve the LO and SOCO models in a few hours.
To further reduce optimization model size, we increase the size of interior voxels. As Table 3 demonstrates, this significantly reduces problem size, and SOCO solution time reduces to minutes, however MIO models are still solved in hours. To solve the MIO models efficiently, we develop a successive relation technique that solves MIO models in less than one hour. Figure 2 shows that both the MIO and SOCO approaches provide high-quality plans, although plans by the LO approach are not satisfying.

5 Conclusion

In this paper, we proposed proton therapy treatment planning approaches that provide optimal plans satisfying the DVH constraints with a few minutes of optimization time. This process is completely automated and no need for extra intervention from clinical staff. To speed up the process of optimizing a treatment plan, we reduced problem size by cropping CT scans, sparsifying the dose-influence matrix, and enlarging the size of interior voxels. All these preprocessing operations are done in a smart way to ensure a high-quality plan. We present three optimization approaches. Based on the results, the second-order cone and the mixed-integer models provide similarly conformal plans. To solve the mixed-integer model, we develop a successive relaxation method that provides a high-quality plan with respect to DVH goals within minutes.

References

[1] S. McGowan, N. Burnet, and A. Lomax. “Treatment planning optimisation in proton therapy”. *The British Journal of Radiology* 86.1021 (2013), pp. 20120288–20120288.

[2] T. E. Merchant, C.-h. Hua, H. Shukla, et al. “Proton versus photon radiotherapy for common pediatric brain tumors: comparison of models of dose characteristics and their relationship to cognitive function”. *Pediatric Blood & Cancer* 51.1 (2008), pp. 110–117.

[3] D. M. Shepard, M. C. Ferris, G. H. Olivera, et al. “Optimizing the delivery of radiation therapy to cancer patients”. *SIAM Review* 41.4 (1999), pp. 721–744.

[4] H. Nystrom, M. F. Jensen, and P. W. Nystrom. “Treatment planning for proton therapy: what is needed in the next 10 years?” *The British Journal of Radiology* 93.1107 (2020), p. 20190304.

[5] M. Hillbrand and D. Georg. “Assessing a set of optimal user interface parameters for intensity-modulated proton therapy planning”. *J. of Applied Clinical Medical Physics* 11.4 (2010), pp. 93–104.

[6] S. van de Water, A. Kraan, S. Breedveld, et al. “Improved efficiency of multi-criteria IMPT treatment planning using iterative resampling of randomly placed pencil beams”. *Physics in Medicine & Biology* 58.19 (2013), p. 6969.

[7] H.-P. Wieser, E. Cisternas, N. Wahl, et al. “Development of the open-source dose calculation and optimization toolkit matRad”. *Medical Physics* 44.6 (2017), pp. 2556–2568.

[8] Y. Zinchenko, T. Craig, H. Keller, et al. “Controlling the dose distribution with gEUD-type constraints within the convex radiotherapy optimization framework”. *Physics in Medicine & Biology* 53.12 (2008), p. 3231.

A neural network-based dose calculation engine for proton spots dose: a clinical implement study

Xuying Shang[2, 1, 3, #], Yaoying Liu[2, 1, 3, #], Zishen Wang[3], Nan Li[3], Chunfeng Fang[3], Yue Zou[3], Wei Zhao[2, 4], Xiaoyun Le[2], Gaolong Zhang[2], Shouping Xu[1, 2, *]

[1] National Cancer Center/ National Clinical Research Center for Cancer/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, 100021, People's Republic of China

[2] School of Physics, Beihang University, Beijing 102206, People's Republic of China

[3] Hebei Yizhou Tumor Hospital, Hebei zhuozhou 072750, People's Republic of China

[4] Beihang Hangzhou Innovation Institute, Yuhang Xixi Octagon City, Hangzhou, 310030, People's Republic of China

[#] The authors contributed equally to this work

[*] Authors to whom correspondence should be addressed

Abstract A fast-accurate dose calculation algorithm can benefit planning and QA in proton therapy with pencil beam scanning. Recently, a study proposed a neural network (NN)-based dose calculation method for proton spots to accelerate calculation speed. However, there is a lack of a method to apply the NN method in calculating spot dose with complex configurations, such as specific energy, range shifter, isocenter, and scanning location. The study presented a detailed method to calculate the spot dose of clinical plans. In the phase of NN training, a total of 119244 pairs of MC spots dose and CT array were generated. The results showed significant accuracy of the NN. Testing spots' gamma passing rates (GPRs, 3mm/2%) were 98.76 %. The global GPR of 2 untrained patient doses was 97.76 %, and the mean time cost was 0.5s. We successfully established a frame for using NN to calculate complex plan doses.

1 Introduction

Both accuracy and speed are significant in dose calculations for planning and QA in proton therapy with pencil beam scanning (PBS) [1]. However, in clinics, accurate dose algorithms such as Monte Carlo (MC) are always time-consuming [2], which is not suggested in planning. In contrast, analytical algorithms such as the pencil beam algorithm (PBA) [3] have an outstanding speed performance that is widely applied in the calculations of dose influence matrix (d_{ij}) [4] in planning. However, PBA can quickly increase dose errors in heterogeneous tissues, leading to dose variations between optimization dose and final dose QA, and it will decrease the quality of plans.

Recent studies have proposed a neural network (NN) — based dose calculation method for proton spots to accelerate calculation speed [5,6], and the accuracy is higher than PBA. However, the studies were only focused on calculating doses with specific energy but without a method of computing doses in clinical scenes, such as with range shifters or specific gantry angles and emission angles.

Therefore, the study proposed a nova model named spots dose neural network (SDNN); the model input data involves energy and CT images and involves the name of range shifters. The SDNN dose calculation model is towards a clinically implemented proton dose calculation with PBS.

2 Materials and Methods

In establishing the dose calculation model, we generated 119244 data pairs for training and evaluating the model by random MC dose calculations.

2.1 Clinical Data

A clinical dataset was collected that contains 40 clinically implemented cases with PBS, which has CT images (slice size of 512 × 512, with pixel spacing of 1mm × 1mm, and slice thickness ranging from 2mm to 3mm), dose, plan, and contour files.

The 38 cases for generating random MC dose to train the model and the left 2 cases (a case with range shifter, another without) were used to validate the model's accuracy in calculating the global dose based on RTPLAN.

2.2 Random Monte Carlo Dose Generations

To implement NN in the calculated spot dose, we sampled full CT images into 16 × 256 × 256 size CT bars (pixel spacing of 3mm × 3mm × 3mm), with a random gantry angle and spot position, but we only reserved the sampled CT bars that passed the Body area of the patient.

A well-configured TOPAS-based MC frame [7] for PBS was used to calculate the MC dose on CT bars. The MC frame was based on the source plane model; thus, we located the source plane at the first slice of the CT bar with random energies. Moreover, we also computed the dose with a range shifter, which was close to the source plane in the MC dose computation.

In the phase of MC dose generation, a total of 119244 pairs of dose and CT bars were generated, with specific energies and range shifters—111296 pairs for training, 3974 pairs for validating and 3974 pairs for testing.

2.3 The Neural Network Structure

The backbone of the SDNN is an ′encoder-transformer-decoder' structure; the encoder converts 2D CT slice into 1D features; in contrast, the decoder refines the transformer

output 1D array into a 2D-slice array, slice by slice, then forms the 3D dose. The transformer structure takes CT features, energy, and range shifter switch (if the MC dose is calculated with the range shifter, the switch equals 1; it equals 0) and outputs 1D dose features. The overview of the SDNN is in Figure 1. The MSE loss function and Adam optimizer were used to optimize the weights of SDNN.

2.4 The Method of Evaluation

The gamma passing rates (GPRs, 3mm/2%) were used to evaluate the NN's accuracy in the untrained testing dataset and two patient plans. In the phase of computing plan dose by NN, we sampled CT bars according to the plan Dicom files, such as isocenter, gantry angle, spot position, and range shifter. When all the spot doses were calculated on CT bars by NN, a re-sample operation was made to align all spot doses to the dose coordination, resulting in the planned dose.

3 Results

The mean GPRs between the NN calculated, and MC doses in 3974 testing data were 98.76±1.54 %; The GPR for the plan dose without range shifter was 98.85%, for the plan dose with range shifter (water equivalent thickness equal 40 mm) was 96.67%, which shown a significant accuracy in a PBS plan dose calculation, however, the NN achieved an outstanding speed, that outcome a plan dose within 0.5s on an RTX 3090 GPU, in contrast, the TOPAS MC will cost beyond 5 hours to compute a plan dose with the comparable accuracy. The NN method shows a decisive advantage in accurate dose calculation.

4 Discussion

The study proposed a method with NN to compute spot dose in complex clinical scenes, such as with range shifters in specific gantry angles and spot positions. The feasibility study showed the application of the NN method in clinical proton dose calculation. In future studies, we should collect more clinical plan doses to evaluate the accuracy of our NN dose calculation method. Moreover, we should consider more clinical details, such as the couch angle, which was not considered in our study. Another possible attempt to boost the accuracy of NN is using clinical spot dose as training data; in contrast, the spot dose in our study is all generated by random MC dose calculated on random CT bars with random energies, whether clinical spot dose can benefit NN's accuracy should be considered in future. In the last, although we have simplified NN's structure, such as only using a single convolution/deconvolution in encoder/decoder to accelerate the calculation speed, there is still the possibility of simplifying the computation flow of the NN, such as token-selection tech [8] which only reserve the token that effective to the NN's output, to reduce the cost of computation.

5 Conclusion

The study proposed a nova neural network structure that can compute proton spot dose with/without range shifters. Our NN method of spot dose calculation is accurate in both random spot and clinical plan doses. The feasibility study shows that our NN method of proton dose calculation has the potential for clinical implementation.

Funding

This study was funded by the National Natural Science Foundation of China (No. 12375359).

References

[1] Fippel M, Soukup M. A Monte Carlo dose calculation algorithm for proton therapy[J]. Medical physics, 2004, 31(8): 2263-2273.

[2] Wu C, Nguyen D, Xing Y, et al. Improving proton dose calculation accuracy by using deep learning[J]. Machine learning: science and technology, 2021, 2(1): 015017.

[3] Da Silva J, Ansorge R, Jena R. Sub-second pencil beam dose calculation on GPU for adaptive proton therapy[J]. Physics in Medicine & Biology, 2015, 60(12): 4777.

[4] Liu W, Zhang X, Li Y, et al. Robust optimization of intensity modulated proton therapy[J]. Medical physics, 2012, 39(2): 1079-1091.

[5] Neishabouri A, Wahl N, Mairani A, et al. Long short-term memory networks for proton dose calculation in highly heterogeneous tissues[J]. Medical physics, 2021, 48(4): 1893-1908.

[6] Pastor-Serrano O, Perkó Z. Millisecond speed deep learning based proton dose calculation with Monte Carlo accuracy[J]. Physics in Medicine & Biology, 2022, 67(10): 105006.

[7] Liu Y, Shang X, Zhao W, et al. Commissioning dose computation model for proton source in pencil beam scanning therapy by convolution neural networks[J]. Physics in Medicine & Biology, 2023, 68(15): 155010.

[8] Xu Y, Zhang Z, Zhang M, et al. Evo-vit: Slow-fast token evolution for dynamic vision transformer[C]. Proceedings of the AAAI Conference on Artificial Intelligence. 2022, 36(3): 2964-2972.

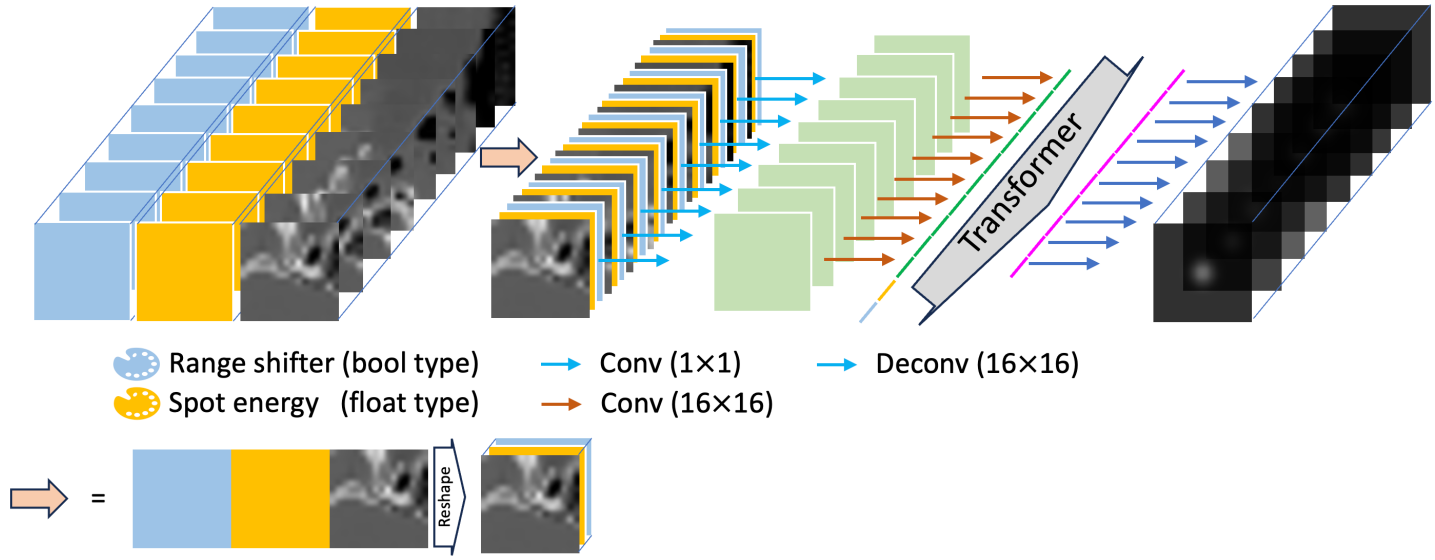


Figure 1. The structure of SDNN.

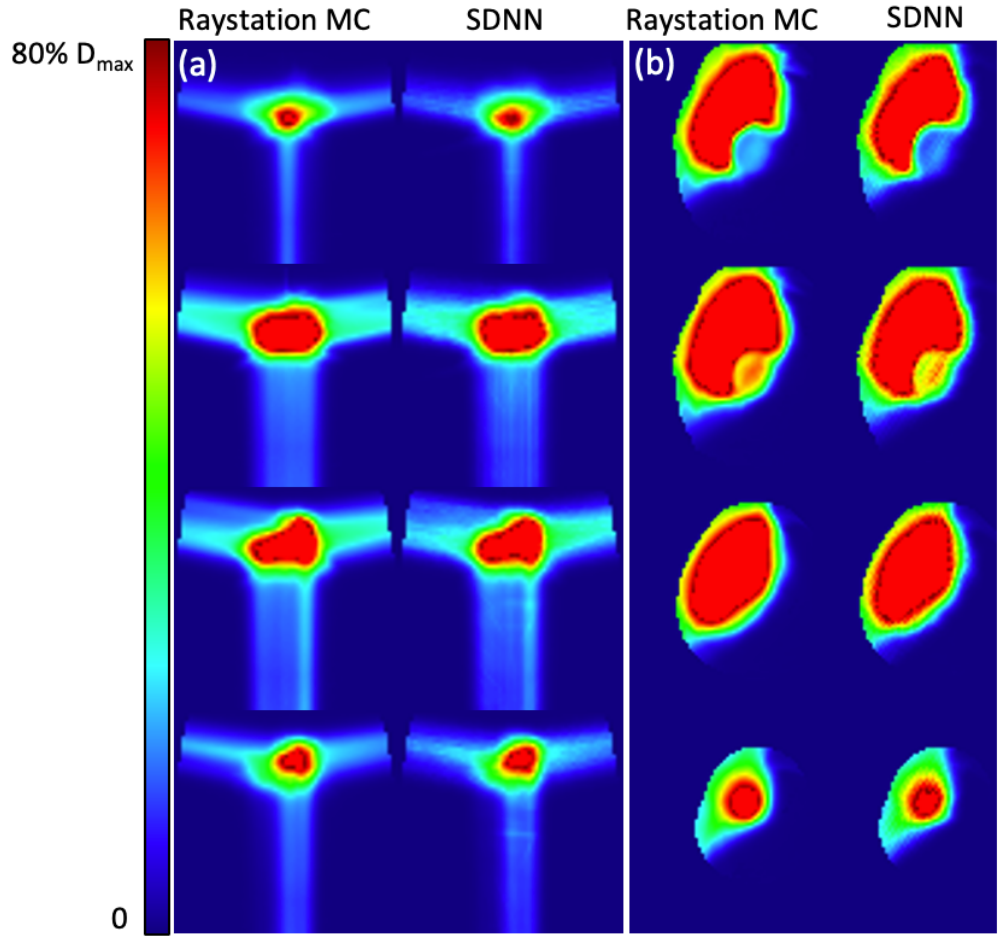


Figure 2. A comparison between Raystation MC and SDNN calculation. (a) a plan dose without range shifter, (b) a plan dose with range shifter.

JulianA.jl - A flexible open source Julia package for pencil beam scanning proton therapy

Renato Bellotti[1,2], Tony J. Lomax[1,2], Andreas Adelmann[2,3], Damien C. Weber[1, 4, 5], and Jan Hrbacek[2,*]

[1]Center for Proton Therapy, Paul Scherrer Institut
[2]Department of Physics, ETH Zürich
[3]Laboratory for Simulation and Modelling, Paul Scherrer Institut
[4]Department of Radiation Oncology, University Hospital of Zurich, USZ, Zurich
[5]Department of Radiation Oncology, University Hospital of Bern, Inselspital, Bern
[*]Corresponding author: jan.hrbacek@psi.ch, Forschungsstrasse 111, 5232 Villigen PSI, Switzerland

Abstract We present JulianA.jl, an open source Julia package tailored to the needs of proton radiotherapy. JulianA accelerates proton radiotherapy research in multiple ways. First, it enriches the proton radiotherapy community with the full power of the Julia ecosystem for scientific computing. This allows researchers to use efficient implementations of the latest algorithms to improve treatment success and efficiency. Second, JulianA provides a flexible and modular toolkit of common functionality that relieves researchers from reinventing the wheel. Third, it includes a powerful fully automatic spot weight optimisation algorithm for proton radiotherapy. Fourth, its code is easily readable, usable and extensible even by domain experts with limited programming skills and easy to use even for students with a limited time budget. We demonstrate how a treatment plan of clinical relevance can be generated in only 19 min with just 38 lines of code, fully automatically and without providing objective weights or training data. Finally, we demonstrate JulianA's general utility for proton radiotherapy research with an application in deep learning.

1 Introduction

Software plays an increasingly important role in radiotherapy. From selecting the most beneficial treatment modality [1] to autosegmentation [2, 3] and autoplanning [4] - software already is shaping and advancing the state of the art in radiotherapy and the workflows in clinics. The surge of machine learning and artificial intelligence in radiotherapy [5] will further emphasise the importance of easily available, efficient and well-tested algorithms.

The complexity of radiotherapy workflows requires the combination of multiple specialised systems, sometimes even a mix of in-house and commercial software. Data transfer between these systems is often performed manually in a cumbersome and error-prone process, slowing down research in algorithms and imposing problems with regard to reproducibility. All major vendors offer a scripting API to their treatment planning systems (TPS), which removes the need for some manual steps for operations within the same planning system. However, data transfer to other systems still remains a manual process. The scripts are typically tailored to a specific TPS, sometimes even a specific version. This strong dependence forces clinicians and researchers to reimplement their scripts from scratch after changing a vendor, or even worse, prohibits changing the vendor altogether. This imposes a barrier to code reuse and wastes resources that could be spent on patient care or further clinical research. Finally, the scripting APIs only allow a superficial access to the planning systems, but not to the underlying algorithms. Researcher are limited to using the entire proprietary pipelines, but cannot examine or improve them one piece at a time. This hinders the proliferation of novel algorithms and the reproducibility of results across manufacturers.

Open source TPS efficiently address the problems of limited code reuse and modularity in two ways. First, the unrestricted access to source code fosters code reuse and supports researchers when developing novel algorithms. Second, open source TPS allow to combine functionality in ways that were not foreseen by the original authors. The two most popular open source TPS are matRad [6] and CERR [7]. Both have enabled a plethora of papers, are written in MATLAB and aim to provide TPS functionality to researchers. However, MATLAB, once an obvious choice for scientists without computer science background, is not the pinnacle of scientific computing for domain experts anymore. Specifically, the language is comparatively slow unless the problem can be formulated as matrix operations, which is not always the case in proton radiotherapy (e. g. calculating radiological path lengths, depth-dose curve lookups).

We present JulianA.jl, a novel open source package written in the Julia programming language [8]. It is tailored for applications in proton radiotherapy clinics and research environments and offers highest flexibility. The name JulianA is a portmanteau of the Julia programming language and FIonA, the name of our in-house TPS. JulianA's architecture and design prioritise ease of use, flexibility, extensibility and modularity. JulianA employs a script-first approach: It aims to provides a robust and modern software stack that enables researchers and clinicians to build their own tools. However, it does not contain a graphical user interface (GUI). Rather, it provides clean interfaces to TPS, which allows building TPS-agnostic workflows and implementing algorithms without vendor lock in. On top, it contains an automatic treatment planning algorithm for proton radiotherapy that does not rely on patient-specific objective weights or training data, enabling researchers to create high-quality

treatment plans within minutes [9]. The algorithm was only tested for intensity-modulated proton therapy, but in principle it could be applied to any other radiotherapy modality that relies on an influence matrix for optimisation, such as intensity-modulated radiotherapy. While the autoplanning algorithm has already been published, this work presents the underlying framework as a software package. The following sections explain our reasoning for the choice of the Julia language, outline an overview of the functionality and provide an example code including results to demonstrate the utility of JulianA in a research setting. Finally, we conclude with a presentation of first applications of the package in our research institute.

2 Materials and Methods

2.1 Design and Goals

The target audience of JulianA.jl are medical physicists, dosimetrists as well as bachelor, master and PhD students. This audience often has limited software development skills and limited time for tasks that involve substantial coding. Therefore, any software targeted towards them should be as simple to use as possible. The users should be empowered to quickly try out new ideas, exchange every building block of a workflow and automate any task they might need as a part of their research or clinical workflow. Therefore, JulianA was developed with three main characteristics in mind. First, it is written in a clear and readable way, following principles of modern software design such as SOLID [10], DRY (don't repeat yourself) [11] and KISS (keep it short and simple), to remain future-proof and easy to learn. This also lowers maintenance work and allows quicker addition of features. Second, JulianA is written in a modular way, which fosters code reuse and therefore promises synergies within and hopefully across research facilities. Further, modularity allows to build complex workflows without having to rewrite any code. Third, JulianA is vendor agnostic, yet allows interfacing with any treatment planning system. This property is especially important because it lays the foundation of code reuse across facilities, accelerating the adoption of new algorithms and kindling healthy competition among commercial vendors.

JulianA was implemented in the Julia programming language because it is easy to use and leads to fast code. These properties recently greatly increased Julia's popularity for computational sciences. The increased popularity in turn gave rise to a vast ecosystem of packages in the language. More and more developers of algorithms decide to use Julia for their proof-of-principle codes, which puts Julia at the forefront of computational advances. Additionally, Julia can be interfaced to Python and C code (two-way interfaces), which broadens the wealth of available packages even more. Julia comes with an advanced package management system that not only improves reproducibility, but also facilitates installation and deployment. Finally, Julia makes it easy for domain experts such as medical physicists to implement GPU code without diving into the depths of low-level programming languages. Further advantages of Julia can be found in literature [8, 12].

JulianA.jl is not another open source TPS. This niche is already occupied by excellent codes such as matRad and CERR. Experiences with our in-house TPS FIonA have shown however, that a substantial amount of work is needed to implement, maintain and extend graphical user interfaces (GUIs). In order to streamline quick and flexible development of novel algorithms and powerful workflows, we have employed a a scripting-first philosophy and deliberately abstained from including a GUI. Instead, JulianA.jl provides a stable and flexible foundation on top of which other tools (potentially with a GUI) can be built.

2.2 Features

JulianA caters to both the needs of users who need to automate parts of their workflows and users who need to write their own algorithms. For the former, it provides easy-to use high-level functions. For the latter, it provides a plethora of low-level operations such as calculating binary masks from contours, calculating in-slice distances from a structure, calculating water-equivalent depth and more. This architecture provides a quick entry path into the framework, but still allows users to build their own algorithms based on the low-level building blocks. At the same time, the high-level API serves as an example how to use the underlying low-level building blocks.

JulianA.jl supports many basic building blocks in proton radiotherapy. It can read and write DICOM CTs, structure sets and RT dose distributions. By default, all the metadata is removed to facilitate research, but meta-data can easily be retained or added if required.

JulianA.jl facilitates working with structures by calculating in-slice and full 3D distances to contours. The availability of the distance to the structure allows computing binary masks, which simplify any operations like inside-outside tests, calculating dose-volume histograms, mean and maximum doses to structures, structure expansion and shrinkage and boolean operations such as union and difference. Both the distances and binary masks are part of the structure's representation and can be easily accessed as attributes, which greatly facilitates the implementation of all algorithms that rely on structures.

JulianA.jl contains code to calculate the radiological path for photons and protons and dose distributions and influence matrices for proton radiotherapy. Further, it is possible to interface JulianA.jl with any TPS that allows scripting. The current version supports our in-house TPS FIonA, but we are currently investigating an integration to Eclipse and RayStation.

JulianA.jl also provides reporting and plotting functions. The former enables researchers to compare target coverage and compliance with organ-at-risk (OAR) constraints as tables with just two function calls and export as full LaTeXreports.

The latter creates publication-grade dose distribution and dose-volume histogram plots.
All this functionality has allowed us to quickly develop a novel autoplanning algorithm for proton spot weight optimisation, which is also included in JulianA.jl [9].

2.3 TPS Interface

JulianA can be interfaced with any TPS by specialising the following three functions

```
place_spots(
    tps,
    ct,
    field_definitions,
    # Structures that defines the iso-centers.
    field_structures,
)
calculate_Dij_matrix(
    tps,
    ct,
    plan,
    optimisation_points,
)
calculate_dose(
    tps,
    resolution,
    ct,
    plan,
),
```

where `tps` is a subtype of `Juliana.AbstractTps`. We have implemented them for our in-house TPS FIonA. Running the same JulianA.jl code in another institute with a different TPS can be realised by changing the type of the `tps` struct. No other changes are needed.

3 Results

3.1 Example code

The following code demonstrates how little code is needed to run a fully automatic spot weight optimisation using JulianA.jl:

```
# Load data.
ct, structures =
    Juliana.load_dicom_directory(data_dir)
constraints = Juliana.parse_oar_constraints_file(
    "$(data_dir)/constraints.csv",
    target_dose,
    structures,
)
prescriptions = Juliana.Prescriptions(
    [
        ("PTV_05_5412", 54.12f0),
        ("PTV_05_5940", 59.40f0),
        ("PTV_05_6996", 69.96f0),
    ],
    constraints,
)
tps = Juliana.FionaStandalone.FionaStandaloneTps(
    fiona_standalone_bin_path, fiona_jar_path,
    output_dir,
)
fields, center_structures =
    load_fields("$(data_dir)/beam_arrangement.csv",
    structures)

# Place spots.
plan = Juliana.FionaStandalone.place_spots(
    tps,
    ct,
    fields,
    center_structures,
)
 # Optimise.
optimal_weights, _, _ =
    Juliana.optimise_head_and_neck(ct, structures,
    prescriptions, plan, tps, tps.work_dir)
plan_optimised =
    Juliana.FionaStandalone.update_spot_weights(plan,
    optimal_weights)
# Calculate a dose distribution.
dose = Juliana.FionaStandalone.calculate_dose(
    tps,
    0.35f0, # dose grid resolution
    ct,
    plan_optimised,
)
```

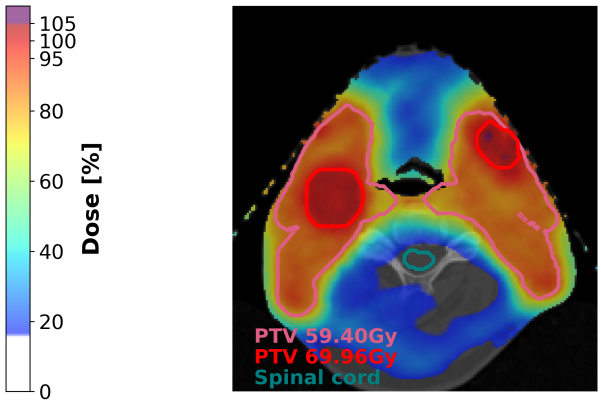


Figure 1: Pencil beam scanning proton dose distribution of a nasal cavity tumor generated and plotted using JulianA.jl. Color wash isodose levels of the prescribed dose are displayed on the left-hand side of the figure.

This code runs in approx. 19 min on an NVIDIA DGX A100 GPU, optimising 6 fields that contain a total of 152′989 spots. The resulting dose distribution is depicted in Fig. 1.

3.2 Applications

The validation study of the JulianA autoplanning algorithm [9] is currently under peer-review. Another study that

examines its capabilities for head-and-neck planning is in the writing stage. Based on the preliminary results, we are exploring the utility of JulianA.jl as a quality assurance tool and for patient selection at our institute. Finally, the JulianA.jl package has accelerated the validation of a deep learning model [13] (currently undergoing peer review) considerably. JulianA relieved the deep learning expert from learning how to use a TPS and manual treatment planning for each patient. In fact, JulianA.jl enabled us to perform clinical quality treatment planning for 12 patients within a single day and is able to scale to even larger datasets.

4 Discussion and Conclusion

We have presented the JulianA.jl package for proton radiotherapy. The package is easy to use and extend and allows researchers without special training to generate clinics-grade treatment plans. We have demonstrated that the package is useful to accelerate research on artifical intelligence in proton radiotherapy and intend to use it as a base package for more research endeavours at our institute. There are plans to extend our TPS interface to automatically upload JulianA.jl treatment plans back to our production TPS and add support for additional commercial TPS. Finally, we hope that our contribution encourages other researchers to write their code in a TPS-agnostic fashion and share it with the community.

References

[1] J. A. Langendijk, P. Lambin, D. De Ruysscher, et al. “Selection of patients for radiotherapy with protons aiming at reduction of side effects: the model-based approach”. en. *Radiother. Oncol.* 107.3 (June 2013), pp. 267–273.

[2] C. E. Cardenas, J. Yang, B. M. Anderson, et al. “Advances in auto-segmentation”. en. *Semin. Radiat. Oncol.* 29.3 (July 2019), pp. 185–197.

[3] K. Harrison, H. Pullen, C. Welsh, et al. “Machine learning for auto-segmentation in radiotherapy planning”. en. *Clin. Oncol. (R Coll. Radiol.)* 34.2 (Feb. 2022), pp. 74–88.

[4] S. Momin, Y. Fu, Y. Lei, et al. “Knowledge-based radiation treatment planning: A data-driven method survey”. en. *J. Appl. Clin. Med. Phys.* 22.8 (Aug. 2021), pp. 16–44.

[5] G. Francolini, I. Desideri, G. Stocchi, et al. “Artificial Intelligence in radiotherapy: state of the art and future directions”. en. *Med. Oncol.* 37.6 (Apr. 2020), p. 50.

[6] H.-P. Wieser, E. Cisternas, N. Wahl, et al. “Development of the open-source dose calculation and optimization toolkit matRad”. *Medical Physics* 44.6 (2017), pp. 2556–2568. DOI: 10.1002/mp.12251.

[7] J. O. Deasy, A. I. Blanco, and V. H. Clark. “CERR: A computational environment for radiotherapy research”. *Medical Physics* 30.5 (2003), pp. 979–985. DOI: https://doi.org/10.1118/1.1568978.

[8] J. Bezanson, A. Edelman, S. Karpinski, et al. “Julia: A Fresh Approach to Numerical Computing”. *SIAM Review* 59.1 (2017), pp. 65–98. DOI: 10.1137/141000671.

[9] R. Bellotti, J. Willmann, A. J. Lomax, et al. *JulianA: An automatic treatment planning platform for intensity-modulated proton therapy and its application to intra- and extracerebral neoplasms*. 2023.

[10] R. C. Martin. *Agile software development, principles, patterns, and practices*. en. Alan Apt series. Upper Saddle River, NJ: Pearson, Oct. 2002.

[11] A. Hunt and D. Thomas. *The pragmatic programmer*. en. Boston, MA: Addison Wesley, Oct. 1999.

[12] E. Roesch, J. G. Greener, A. L. MacLean, et al. “Julia for biologists”. en. *Nat. Methods* 20.5 (May 2023), pp. 655–664.

[13] X. Li, R. Bellotti, B. Bachtiary, et al. *A Unified Generation-Registration Framework for Improved MR-based CT Synthesis in Proton Therapy*. 2024. DOI: 10.48550/ARXIV.2401.12878.

[14] R. Bellotti. *JulianA.jl*. 2023.

A Code availability

The latest version of the Juliana.jl code can be found on Zenodo [14].

Monte Carlo verification of proton treatment planning

K. Badraoui Čuprová[1], O. Hrstka[2] and V. Vondráček[1]

[1]Department of Clinical Physics, Proton Therapy Center, Prague, Czech Republic
[2]Department of IBA, Proton Therapy Center, Prague, Czech Republic

Abstract
In Proton Therapy Center in Prague we treat patients since 12.12.2012 with pencil beam scanning technique. We treated more than 10 000 patients so far. For quality control of treatment planning we have developed an independent dose calculation check. Our automated workflow is based on Monte Carlo simulations and enables verification of dose computational algorithms of local treatment planning system RayStation 12A (RaySearch, Sweden).

1 Introduction

Explicit modelling of particle transport makes the Monte Carlo method the method with the highest calculation accuracy. Dose distribution is calculated in real inhomogeneous patient environment, therefore MC provides complementary information to QA measurement in water. Linear energy transfer distribution in patient can be simulated alongside.
Our project is built on open-source software GATE - a Monte Carlo simulation toolkit for medical physics applications (http://www.opengatecollaboration.org). High precision and accuracy of this powerful method is paid by large computation costs. We reduce the computation time through parallel computing with HTCondor (https://htcondor.org) – an open-source high-throughput computing software. Currently we compute on 324 CPU cores.
The workflow is fully automated. The user sends DICOM files from PACS to dedicated server where the pre-processing, the simulation and the post-processing are done automatically. The output files are sent back to PACS and are ready to be imported to TPS or OmniPro I'mRT software (www.iba-dosimetry.com) for further analysis.
The independent dose verification system was validated against the TPS and shows very good agreement. Currently we are in the phase of integrating the system into clinical workflow.

2 Materials and Methods

For a realistic radiotherapy simulation in patient, information stored in DICOM files (RT Plan, RT Structure Set, CT Scans, RT Dose) must be read and processed to specific input files for GATE simulation. DICOM files exported from TPS are not directly readable by GATE, so we have developed various scripts in Python (version 2.7 and 3.8) to process input and output data.
The automated workflow can be divided into three stages: pre-processing, simulation and post-processing.

Pre-processing
1) The patient´s dataset (RTP, RTSS, RTD and CT) is exported from TPS to PACS
2) The entire dataset is sent from PACS to dedicated server.
3) The server automatically checks for the presence of the required DICOM files every 10 minutes and starts to run a sequence of python scripts. The patient directory has a defined structure, the DICOM files and the patient non-specific files are copied to appropriate locations.
4) The first script parses CT directory and RTStructureSet to create voxelized 3D patient phantom. The image is cropped with respect to body contour and PTV and it is resized with respect to slice resolution. Mhd and raw files are generated.
5) The second script parses RTPlan and CT image, it loads main.mac template and it creates main.mac file for each treatment field individually. (Main.mac file contains the main instructions and parameters for running a simulation.)
6) The third script parses RTPlan and creates beam.mac file for each treatment field. Each field is described by gantry angle and collection of spots. Each single spot is defined by its position, weight and energy.
7) The fourth script parses RTPlan, it loads actors.mac template and it creates actors.mac file for each treatment field. Dose actor, LET actor, simulation statistic actor and calculation grid are defined.
8) All the files are prepared for MC simulation. The job is split for parallel processing with HTCondor. The computational task is submitted to 324 cores.

Simulation
Monte Carlo simulations are run in open source software Gate (version 9.2). Gate is a GEANT4 based simulation platform for nuclear medicine. The treatment nozzle is not simulated. For the pencil beam scanning technique only few elements are present in the nozzle and they do not actively participate in the shaping of the beam. Optical properties of each pencil beam in relationship with energy are defined in the source description file. The source

description file is specific for each treatment room and is based on measurements of spots and PDDs for the beam model. The file is a collection of polynomial equations for optical properties (spot size, divergence, emittance), energy properties (energy and energy spread) and MU calibration.
To reduce the computation time parallel computing is utilized. For this purpose HT Condor software is used and the simulation is performed by 324 computational cores. For high simulation efficiency, each core is assigned a number of histories according to its performance. The simulation on each machine finishes at a similar time.
Calculation grid is set to 2mm and physic parameters are set to achieve both clinical level of dose accuracy (simulation statistical uncertainty below 2%) and simulation efficiency.

Post-processing
The post-processing is automatically initiated once all the simulations are completed.
The dose scored by the Monte Carlo method is converted to absolute. From RTP, the total number of Monitor Units (MU) per beam and the number of MU per energy layer can be obtained. The relationship between the number of protons per MU for a given energy is known, therefore the number of protons per given energy layer and per beam can be calculated. From the simulation statistics a total number of simulated protons is obtained and used to determine conversion factor. The absolute dose is calculated.
Dose averaged linear energy transfer is calculated alongside. Dose and LET distributions are converted to DICOM dose planes and they are sent automatically to PACS. They can be imported to TPS or to OmniPro I‘mRT software for comparison with real plan and for further analysis.

3 Results

The automated system was validated on series of patients of various locations (head and neck, prostate, mamma). The evaluation was performed in TPS and in OmniPro I’mRT software. For gamma analysis criteria of 3% and 3mm were used with a threshold of 95% pixels passing. All plans met this criterion. For TPS, dose differences, profiles, dose volume histograms and values for clinical goals were evaluated.

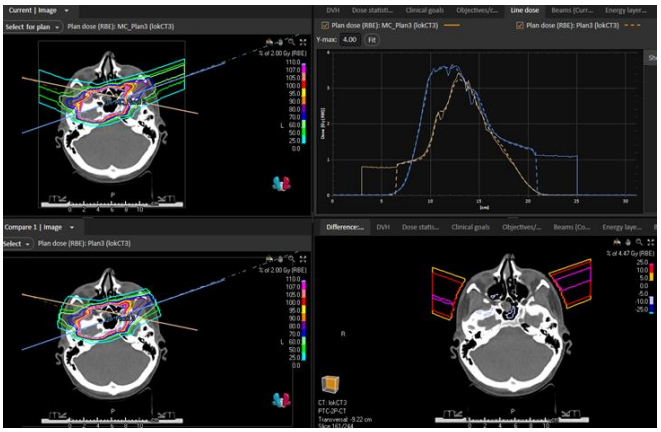


Figure 1: Dose distribution (MC x TPS), profiles, dose difference

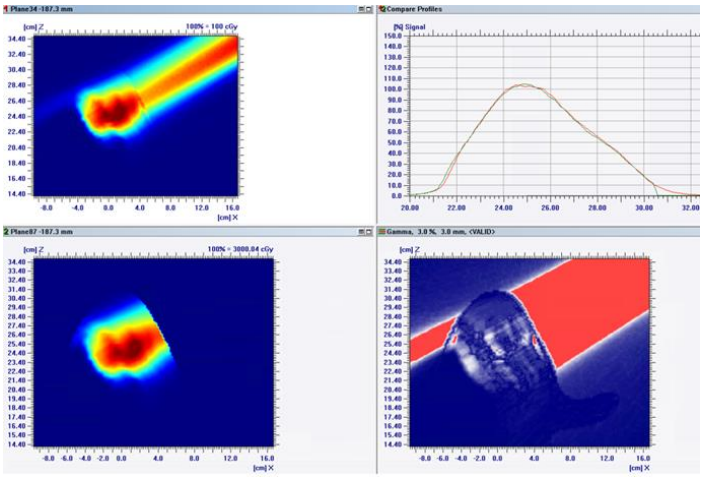


Figure 2: Dose distributions (MC-red x TPS-green), profiles and gamma analysis (3%,3mm)

4 Discussion

An in-house MC method for independent calculation and verification of proton treatment plans was presented. Explicit modelling of particle transport in MC allows more accurate dose calculations in regions with heterogeneities and can be used to benchmark TPS. Our approach calculates dose distributions and dose weighted LET in a patient based on information stored in DICOM files.
The automated system requires minimum interaction, the user just exports DICOM datasets to a specific server. The pre-processing, simulation and post-processing are performed automatically. The resulting Monte Carlo distributions appear in PACS and can be imported for further analysis into TPS or specific software for gamma analysis.

5 Conclusion

The developed MC system was validated and can be used for independent verification of proton treatment planning.
Currently we are in the phase of integrating the system into clinical workflow and prepare protocols.
Our approach can be adapted to other treatment planning systems, in fact this workflow can be implemented at any proton therapy center that uses spot scanning.

References

[1] Grevillot, L., Bertrand, D., Dessy, F., Freud, N., & Sarrut, D. (2011). A Monte Carlo pencil beam scanning model for proton treatment plan simulation using GATE/GEANT4. Physics in Medicine and Biology, 56(16), 5203–5219. https://doi.org/10.1088/0031-9155/56/16/008

RadiSeq: A whole-genome DNA sequencing simulation pipeline to investigate radiation-induced carcinogenesis

Felix Mathew[1,2] and John Kildea[1,2]

[1]Medical Physics Unit, McGill University, Montreal, Canada
[2]Research Institute, McGill University Health Centre, Montreal, Canada

Abstract DNA sequencing holds significant promise for probing the biophysical mechanisms underlying radiation carcinogenesis. However, experimental implementation is hampered by cost and time constraints. To address this, we present a computational pipeline for simulating radiation exposure and subsequent whole-genome DNA sequencing of human cells. Utilizing the TOPAS-nBIO Monte Carlo framework, we constructed a single-cell geometric nuclear DNA model for radiation-exposure simulations. Subsequently, we developed "RadiSeq", an application in C++, to enable bulk-cell and single-cell whole-genome DNA sequencing simulations of radiation-exposed cell models. RadiSeq exhibits acceptable performance, with early validation results showing promise. Upon validation completion, we believe that RadiSeq will be a useful tool for further investigations into the mechanisms of radiation carcinogenesis.

1 Introduction

Iatrogenic cancer risk is often a concern, especially during high-energy radiotherapy [1,2]. Radiation-induced carcinogenesis has long been an area of interest within the scientific community [3,4]. Nevertheless, our understanding of the intricate biophysical mechanisms underpinning radiation-induced carcinogenesis remains incomplete. One promising avenue for advancing our understanding of radiation's genomic impact is through the utilization of DNA sequencing technologies. With the decreasing cost of DNA sequencing, there has been substantial progress in employing this tool to investigate further the radiation-induced mutational processes. Notably, the milestone discovery of a radiation signature in tumors associated with radiation exposure by Behjati et al. in 2016 [5] underscored the effectiveness of bulk-cell whole genome sequencing in identifying radiation-induced mutational profiles.

In our own research program, we have assessed the efficacy of the single-cell whole genome sequencing (ScWGS) technique [6] in discerning radiation-induced genomic alterations within human cells subjected to varying in-vitro radiation doses. Preliminary results from our investigation reveal the potential of ScWGS as a tool for directly assessing the genomic landscape of radiation in exposed cells, offering insights into the early stages preceding mutational cell transformation. Despite its promising applications, it is noteworthy that ScWGS still remains a resource-intensive undertaking, both in terms of expense and time, thereby limiting its widespread adoption for mutational investigations. Addressing these challenges will be imperative for advancing the utility of single-cell sequencing in radiation-induced mutational studies.

Therefore, we developed a comprehensive computational simulation pipeline designed to model ScWGS of radiation-exposed cells, providing a cost-effective alternative to experimental sequencing procedures. Our approach involved the development of a dedicated stand-alone tool named "RadiSeq," designed to simulate both single-cell and bulk-cell whole genome DNA sequencing processes from first principles. Through this tool, we have established a functional pipeline to replicate key experimental procedures: (i) simulating the radiation exposure of cells, (ii) emulating the sequencing library preparation of irradiated cells, and (iii) replicating the whole-genome sequencing of these libraries. This simulation framework closely mirrors the experimental protocol, demonstrating reasonable agreement between simulated and actual procedures.

2 Materials and Methods

Our simulation pipeline for whole-genome DNA sequencing of radiation-exposed cells is executed in two steps : (i) Monte Carlo simulation of cell irradiation and (ii) library preparation and sequencing simulations of the irradiated cells. In this section, each of these aspects is described.

2.1 CELL IRRADIATION SIMULATION

We developed an in-house single-cell geometric nuclear DNA model [7] using the TOPAS-nBio toolkit [8] for Monte Carlo simulations, as shown in Figure 1. This open-source model [9] allows for the stochastic exposure of varying radiation doses, with flexibility in the choice of radiation quality. Within the model, nucleotides, the fundamental structural units of the genome, are constructed using simple spherical volumes. The nucleus, for simulation purposes, is simplified into a cubic grid. Importantly, the number of nucleotide base pairs and the DNA density within the model are maintained to be consistent with that of a human nucleus.

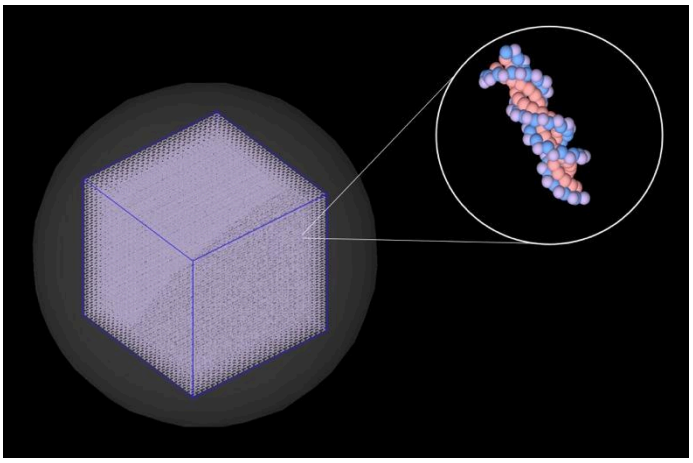


Figure 1: Our geometric nuclear DNA model developed in TOPAS-nBio[7].

In the track structure simulation of secondary particle interactions with our developed model, the underlying physics was managed by the G4EmDNAPhysics_hybrid2and4 physics constructor. Also, the chemistry governing the indirect action to track secondary chemical species was handled by the TsEmDNAChemistry chemistry constructor. Our cell model facilitates the scoring of dose depositions within different geometric volumes throughout the radiation exposure simulation. Furthermore, the model is capable of generating a genomic map detailing all incurred damages, presented as output in the standard DNA damage (SDD) data file format [10]. This approach ensures a detailed and structured representation of the simulated radiation-induced effects on the genome.

2.2 SEQUENCING IRRADIATED CELLS

In the realm of simulating genomic sequencing data from various sequencers, a plethora of computational tools exist [11]. However, when confronted with the task of simulating ScWGS of radiation-exposed cells using our model, none of these tools emerged as a suitable candidate. These artificial read-generating tools are mostly built for application with small, single-stranded reference genomes and are ill-suited for use with our double-stranded, radiation-damaged whole-genome human cell model. Moreover, several of these tools lack adequate testing and validation, necessitating the development of a dedicated tool capable of simulating both sequencing library preparation and the subsequent sequencing process.

To address this gap, we developed RadiSeq in C++, drawing inspiration from the open-source ART simulating toolkit [12]. RadiSeq allows users to simulate bulk-cell or single-cell DNA sequencing of any cell or group of cells, accommodating cell models capable of generating SDD data file outputs. Users have the flexibility to specify parameters such as the Illumina sequencer model (Illumina Inc., San Diego, CA, USA), sequencing coverage, number of cells to be sequenced, and sequencing protocol, among other fine-tuning parameters through a single input file. The software is optimized for handling extensive data files using memory mapping techniques, and its multi-threading capability enables parallelized computation. Incorporating a user-friendly interface, RadiSeq efficiently simulates library preparation and sequencing of irradiated cells from first principles.

Post-sequencing simulation, RadiSeq produces two outputs: (1) FASTQ files containing reads—data derived from small genomic segments analyzed during sequencing—and (2) a run summary file detailing simulation parameters and essential information for result reproduction. Currently undergoing validation with experimentally obtained data, FASTQ files produced by RadiSeq can be processed in a manner similar to those obtained through experimental means. Mutational analysis using diverse bioinformatics tools is expected to yield comparable results, validating the efficacy of RadiSeq in emulating real-world sequencing scenarios.

3 Results

Figure 2 illustrates the processing performance of RadiSeq, comparing the optimized version utilizing memory mapping and multi-threading against the earlier unoptimized iteration of the software. The plot demonstrates that with optimization, RadiSeq can efficiently generate millions of reads within a short time frame, allowing for the simulation of real experimental scenarios in a matter of hours, depending upon the number of CPU cores available for the simulation.

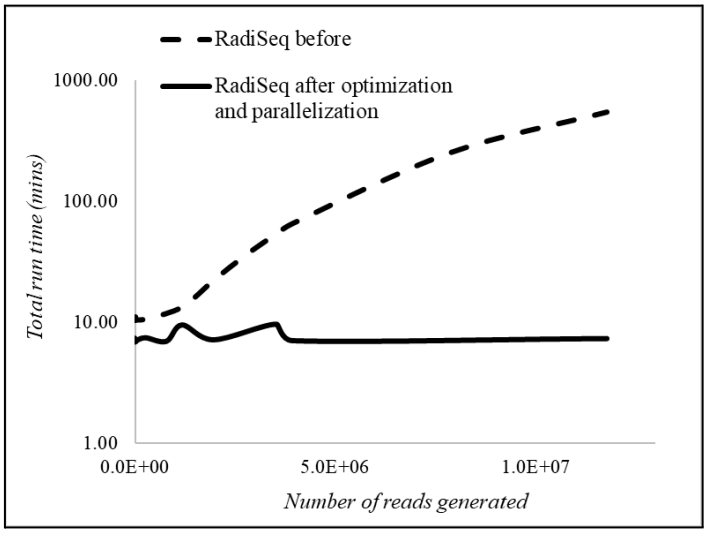


Figure 2: Comparison of RadiSeq performance before and after optimization and multi-threaded parallelization.

In Figure 3, we present preliminary test results from our validation of RadiSeq. The graph compares the distribution of reads across the genome—specifically, the percentage of the genome covered by varying numbers of reads—for both experimentally obtained data and data simulated by RadiSeq. Notably, a reasonable agreement is evident, suggesting the reliability and validity of RadiSeq in faithfully reproducing read distribution patterns observed in actual experimental scenarios.

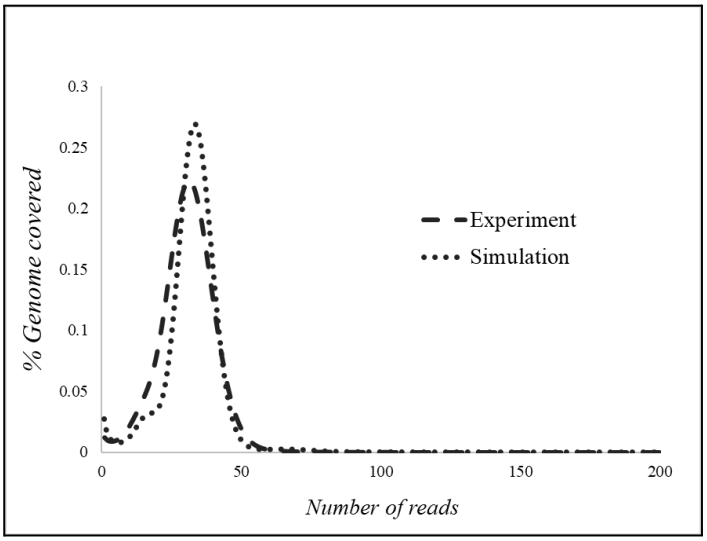


Figure 3: Read coverage obtained from a sequencing experiment and from a sequencing simulation using RadiSeq.

4 Discussion

Our geometric nuclear DNA model, constructed in TOPAS-nBio, is not an exact replication of a human cell; it relies on numerous assumptions and oversimplifications of cellular complexities. This inherent simplification introduces uncertainties and variations in the simulated sequenced data, potentially accounting for observed differences in the comparison of simulation and experimental results. However, it's crucial to note that RadiSeq demonstrates compatibility with any cell model capable of producing an SDD output file from an irradiation simulation. Consequently, researchers have the flexibility to integrate more intricate cell models [13,14] into this pipeline.

Although the preliminary data from RadiSeq's validation are promising, more investigation must be done to verify the accuracy of this tool across multiple metrics. This work is currently ongoing and once RadiSeq is fully validated, we plan to release it under an open-source license.

A notable limitation of our work lies in the omission of considerations for cell repair and cell death following radiation damage. Presently, the cell genome is sequenced with all radiation-induced genomic alterations, akin to performing sequencing immediately after radiation exposure. An enhanced pipeline would encompass the possibility of damage repair after radiation exposure, recognizing that repair processes often rectify simple damages and occasionally result in misrepair, ultimately leading to mutations. While there are tools available for modelling DNA repair subsequent to radiation damage [15,16], these mechanistic models currently suppress the spatial information of damages post-repair, posing challenges in accurately modelling the cell genome. Consequently, within the current framework, our proposed pipeline with RadiSeq is unable to accommodate cell repair.

5 Conclusion

Sequencing analysis holds promising potential in aiding the investigation of radiation carcinogenesis. However, the accessibility of this technology is hindered by its expense and its time-consuming nature. Therefore, we have built RadiSeq to simulate bulk-cell and single-cell DNA sequencing of radiation-exposed human cell models. With this software in place, we have now established a working pipeline that can be used to computationally simulate cell irradiations and irradiated cell sequencing for the purpose of investigating the underlying mechanisms of radiation carcinogenesis at the genomic level. RadiSeq's performance is very promising and the early results from the validation studies of the software suggest that the pipeline is reasonably comparable to the experimental procedure. With RadiSeq, we will be able to identify the optimum experimental irradiation and sequencing conditions prior to commencing experimental work. This will allow us to perform sequencing experiments of irradiated cells cheaper and faster and more efficiently learn about radiation interactions with the genome.

Acknowledgement

This research was supported by the Fonds de Recherche du Québec - Santé (FRQS), the Canadian Space Agency Grant #19FAMCGB25, and by the CREATE Responsible Health and Healthcare Data Science (SDRDS) grant of the Natural Sciences and Engineering Research Council of Canada.

References

[1] Allan JM, Travis LB. Mechanisms of therapy-related carcinogenesis. Nature Reviews Cancer. 2005. pp. 943–955. doi:10.1038/nrc1749

[2] Imaoka T, Nishimura M, Daino K, Takabatake M, Moriyama H, Nishimura Y, et al. Risk of second cancer after ion beam radiotherapy: insights from animal carcinogenesis studies. Int J Radiat Biol. 2019;95: 1431–1440. DOI: 10.1080/09553002.2018.1547848

[3] Yuhas, John M, et al. Biology of Radiation Carcinogenesis. Raven Press, 1976.

[4] Burns, F. J, et al. Radiation Carcinogenesis and DNA Alterations. Springer US, 1986. DOI: 10.1007/978-1-4684-5269-3.

[5] Behjati, S., Gundem, G., Wedge, D. C., Roberts, N. D., Tarpey, P. S., Cooke, S. L., ... & Campbell, P. J.. Mutational signatures of ionizing radiation in second malignancies. Nature communications, 2016; 7(1), 1-8. DOI: 10.1038/ncomms12605

[6] Mathew, F., Manalad, J., Yeo, J., Galarneau, L., Ybarra, N., Wang, Y. C., ... & Kildea, J. (2023). Single-cell DNA sequencing—a potential dosimetric tool. *Radiation Protection Dosimetry*, *199*(15-16), 2047-2052. DOI: 10.1093/rpd/ncad055

[7] Montgomery, L., Lund, C. M., Landry, A., & Kildea, J. Towards the characterization of neutron carcinogenesis through direct action simulations of clustered DNA damage. Physics in Medicine & Biology, 66(20), 205011 (2021). DOI: 10.1088/1361-6560/ac2998

[8] Schuemann, J., McNamara, A. L., Ramos-Méndez, J., Perl, J., Held, K. D., Paganetti, H., ... & Faddegon, B. TOPAS-nBio: an extension to the TOPAS simulation toolkit for cellular and sub-cellular radiobiology. Radiation research, 191(2), 125-138 (2019). DOI: 10.1667/RR15226.1

[9] Montgomery L, Lund C M, Manalad J, Landry A and Kildea J TOPAS Clustered DNA Damage GitHub repository (2021). DOI: 10.5281/zenodo.5090104

[10] Schuemann, J., McNamara, A. L., Warmenhoven, J. W., Henthorn, N. T., Kirkby, K. J., Merchant, M. J., ... & McMahon, S. J. (2019). A new standard DNA damage (SDD) data format. *Radiation research*, *191*(1), 76-92. DOI: 10.1667/RR15209.1

[11] Escalona, M., Rocha, S., & Posada, D. (2016). A comparison of tools for the simulation of genomic next-generation sequencing data. *Nature Reviews Genetics*, *17*(8), 459-469. DOI: 10.1038/nrg.2016.57

[12] Huang, W., Li, L., Myers, J. R., & Marth, G. T. (2012). ART: a next-generation sequencing read simulator. *Bioinformatics*, *28*(4), 593-594. DOI: 10.1093/bioinformatics/btr708

[13] McNamara, A. L., Ramos-Méndez, J., Perl, J., Held, K., Dominguez, N., Moreno, E., ... & Schuemann, J. (2018). Geometrical structures for radiation biology research as implemented in the TOPAS-nBio toolkit. *Physics in Medicine & Biology*, *63*(17), 175018. DOI: 10.1088/1361-6560/aad8eb

[14] Schuemann, J., McNamara, A. L., Ramos-Méndez, J., Perl, J., Held, K. D., Paganetti, H., ... & Faddegon, B. (2019). TOPAS-nBio: an extension to the TOPAS simulation toolkit for cellular and sub-cellular radiobiology. *Radiation research*, *191*(2), 125-138. DOI: 10.1667/RR15226.1

[15] Ingram, S. P., Warmenhoven, J. W., Henthorn, N. T., Smith, E. A. K., Chadwick, A. L., Burnet, N. G., ... & Merchant, M. J. (2019). Mechanistic modelling supports entwined rather than exclusively competitive DNA double-strand break repair pathway. *Scientific Reports*, *9*(1), 6359. DOI: 10.1038/s41598-019-42901-8

[16] McMahon, S. J., Schuemann, J., Paganetti, H., & Prise, K. M. (2016). Mechanistic modelling of DNA repair and cellular survival following radiation-induced DNA damage. *Scientific reports*, *6*(1), 33290. DOI: 10.1038/srep33290

Detectability of Iodine Contrast Agent with Ultra-low Concentration by Photon-counting CT: A Simulation Study

Xiaoyu Hu[1], Yuncheng Zhong[1], Kai Yang[2], and Xun Jia[1]

[1]Department of Radiation Oncology and Molecular Radiation Sciences, Johns Hopkins University, Baltimore, MD, USA
[2]Division of Diagnostic Imaging Physics, Department of Radiology, Massachusetts General Hospital, Boston, MA, USA

Abstract To investigate the contrast-to-noise (CNR) of the iodine contrast agent with low concentrations in CT images acquired with photon-counting detector (PCD), we developed a simulation pipeline to model a multi-channel PCD CT and an energy-integrating detector (EID) CT. A 2D digital phantom was generated with iodine solution inserts of 0.25 to 10 mgI/mL concentrations as the object, and 1/2/3-channel PCD and EID CTs were simulated to compute the CNR of these iodine solutions under a 120 kV x-ray tube voltage with various CT dose levels. Numerical results show that PCD CT outperforms EID CT in terms of CNR in general. Furthermore, the multi-channel PCD-CT provides extra degrees of freedom in the energy dimension for finding the optimal channel, by which an improved CNR can be obtained. When the $\text{CTDI}_{\text{free air}}$ is 3.3 mGy, the PCD CNR averaged over all iodine inserts were 1.26 to 2.13 times greater than the CNR from EID CT. When the $\text{CTDI}_{\text{free air}}$ is 98.7 mGy, the CNR from 1-channel PCD is close to that of EID, but the 2/3-channel PCD CT is able to obtain about 17% higher CNR.

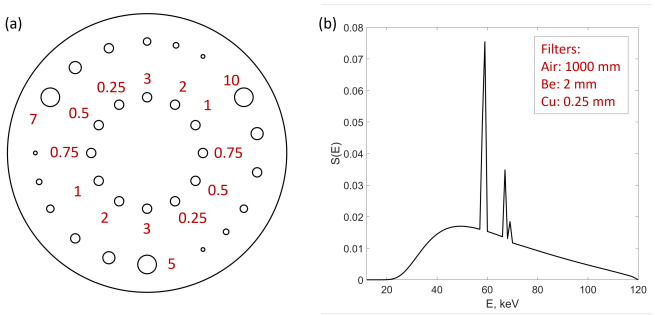


Figure 1: (a) X-ray spectrum at 120 kVp. (b) The digital phantom. The red number indicates the concentration of iodine solutions in the unit of mgI/mL.

1 Introduction

The advent of photon counting technology in computed tomography (CT) has ushered in a new era in medical imaging [1–3]. Unlike traditional energy-integrating detectors (EID), photon counting detectors (PCD) measure the number of incoming photons at discrete energy levels, enabling superior energy discrimination and improved signal-to-noise ratios. This innovative technology holds great promise for providing detailed anatomical information while minimizing patients' radiation exposure dose.

In this context, the application of contrast enhanced CT scans gains even greater significance, offering unprecedented opportunities for precise diagnostic imaging and targeted therapy [4]. Contrast agents help increase targeted abnormal regions and backgrounds, as well as provide functional information, for more effective and sensitive diagnosis. In therapy, the use of contrast agents enhances the delineation of target volumes, facilitating the development of personalized and targeted treatment plans. it also allows monitoring disease responses to treatment over time. Yet, clinical challenges associated with contrast agents, such as potential allergic reactions, nephrotoxicity, and contraindications still persist [5]. It is imperative to enhance the sensitivity of imaging technologies to allow the reduction of contrast agent concentration while maintaining its clinical effectiveness.

The motivation of this study is to investigate thedetectibility of iodine contrast agent in the CT images acquired with PCD. We performed extensive numerical simulations to find the lowest concentration of iodine that can be detected by PCD. The significance of this work is that we show that under a given condition, PCD outperforms EID in detecting the iodine solutions with ultra-low concentrations (< 3 mgI/mL). We focused on iodine contrast due to the wide clinical application, but we expect the method is generalizable to other agents.

2 Method

2.1 Simulation pipeline

We designed a two-dimensional digital phantom for this work as shown in Figure 1(a). The base of the phantom is a pure water cylinder with a 30 cm diameter. Three sets of inserts with different diameters ($0.4 \sim 2$ cm) were filled with 5, 7, 10 mgI/mL iodine solutions, and the 12 equal-sized (1 cm diameter) inserts were uniformly distributed in the inner ring, and iodine solutions with lower concentrations were filled.

The setup of the PCD-CT system was based on a Siemens Force scanner with the source-to-isocenter distance of 59.5 cm [6], and the source-to-detector distance of 108.5 cm. The CdTe PCD had 1500 pixels with a 0.4 mm pixel size and 1 mm thickness. A 120 kVp x-ray spectrum $S(E)$ with a 0.25 mm Copper and 2 mm Beryllium filtration was considered (Figure 1(b)). Additionally, a bow tie filter was added to create a more uniform x-ray beam delivered to the detector.

For a CT system, the signal acquired can be expressed as

$$M(u) = c \int W(E)S(E)\frac{1}{D^2(u)}B(E,u)R(E) \exp\left[-\int_L \mu(E,x)\mathrm{d}l\right]\mathrm{d}E. \quad (1)$$

where u is the detector pixel position, $W(E)$ is the weight of detected photons, $D(u)$ is the distance from the source to a detector pixel in meter, and the term $D^{-2}(u)$ is the inverse square law since the source fluence $S(E)$ was defined at 1 m. $R(E)$ is detector response function. L is the path connecting the source to pixel u, and $\mu(E,\mathbf{x})$ is the x-ray attenuation image of the phantom at energy E and $\mathbf{x}$ is the spatial coordinate. This integral was evaluated using the ray tracing algorithm. The term $B(u,E) = \exp[-\mu_{Al}(E)L_{Al}(u)]$ is the attenuation of the bow tie filter, where L_{Al} is the equivalent aluminum length of the bow tie filter and μ_{Al} is aluminum's total linear attenuation coefficient. Finally, the overall scaling factor c was introduced to control the CT dose level.

The log transformed projection data was then computed as

$$g(u) = -\log\left(\frac{\mathscr{P}[M(u)]+N_e}{M_0(u)}\right), \quad (2)$$

where $\mathscr{P}[\cdot]$ is the Poisson random number generator, N_e is the added electronic noise. $M_0(u)$ is the air norm computed using Eq. (1) without the phantom attenuation term.

For the PCD-CT cases, $W(E) = 1$, and $N_e = 0$. For the ith energy channel defined by lower and upper thresholds T_i and T_{i+1}, the response function was unity within this energy range, and zero otherwise.

We also compared with EID-CT. In this case, $W(E) = E$ and N_e is a Gaussian random number generator with mean 0 and standard deviation σ.

Once the projection data were calculated, we reconstructed CT images using the standard filtered backprojection (FBP) approach with a Shepp-Logen filter. CNR of each iodine contrast insert was measured with the background for each insert defined as the annular surrounding region with the same area.

2.2 Parameter selection

There are two major parameters affecting CNR. The first one is c related to the total photon number per pixel from the x-ray source, *i.e.* the total dose. We estimated the $\mathrm{CTDI}_{free\ air}$ at the isocenter, equivalent to CT Dose Index measured at isocenter in free air corresponding to different c values to ensure the study focuses on a range of doses covering clinically relevant situations.

As for the standard deviation of the electronic noise σ, we selected a square region at the center of the reconstructed CT images and determined the noise level in the CT images by calculating the standard deviation of that region. σ was adjusted to approximately match the image noise level at each dose level to what we observed in typical clinic CT images.

c	10	50	100	200	300
$\mathrm{CTDI}_{\text{free air}}$ [mGy]	3.3	16.5	32.9	65.8	98.7
Noise (PCD) [HU]	62	27	19	14	11
Noise (EID) [HU]	146	37	23	15	12

Table 1: $\mathrm{CTDI}_{\text{free air}}$ and the associated image noise levels.

2.3 Evaluation

In PCD-CT, the lowest energy threshold T_1 was set to 26 keV for all cases, which is the typical setup to eliminate electronic noise. In this study, we modeled the situation of 1 to 3 energy channels. The 1 channel situation utilized photons above T_1 in the the entire spectrum. For 2 channel, the threshold T_2 was used to separate the two channels. Similarly, in 3 channel, two thresholds T_2 and T_3 were used with $T_3 > T_2$. In each situation, we enumerated all possible energy thresholds in the list of energy thresholds from 28 to 120 keV with a 2 keV interval.

We focused on the study to investigate the variation of CNR in PCD-CT as we adjusted the energy thresholds, and compared it with that of EID-CT. In each PCD-CT case, we report the highest CNR among energy channels. When there are more than one energy channel, while it is possible to decompose the PCD-CT images to determine the iodine image, we focused on the PCD-CT images, as material decomposition is expected to increase noise level, likely reducing CNR.

3 Results

3.1 1-channel PCD-CT

The main difference between 1-channel PCD-CT and EID-CT is the energy weighting form and the electronic noise. Since the first threshold T_1 was used to eliminate the electronic noise in the PCD signal, the noise in the 1-channel PCD-CT image is inversely proportional to the CT dose. However, when c is large, the EID signal noise is still dominated by the Poisson noise; when the CT dose is decreased, the additional electronic noise becomes dominant, which is obvious from Table 1. The CNRs of various concentrations of iodine inserts were plotted in Figure 2. When the CT dose is relatively high (>30 mGy), the performance of EID-CT and 1-channel PCD-CT are close, especially for the iodine solutions with low concentrations. However, when the CT dose is low (<30 mGy), the PCD-CT outperforms the EID-CT in this task mainly due to PCD's capability of reducing electronic noise. Quantitatively, for the 3 mgI/mL iodine solution, the CNR obtained with 1-channel PCD-CT is 5% better than the CNR of EID-CT when $c = 300$, and about 1.78 times better when $c = 10$, and for 0.5 mgI/mL, the improvement of PCD is 14% and 46% when c=300 and 10 respectively.

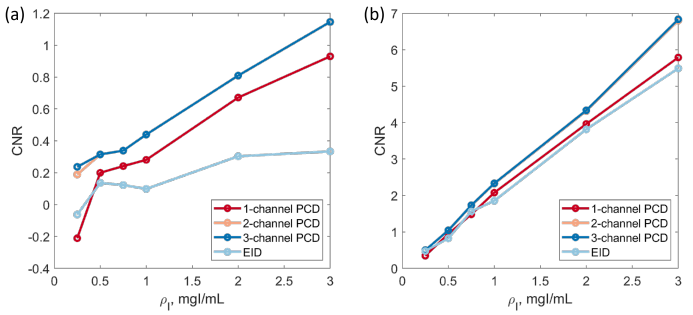


Figure 2: CNR as a function of iodine concentrations for low and high CT dose cases: (a) $c = 10$ and (b) $c = 300$.

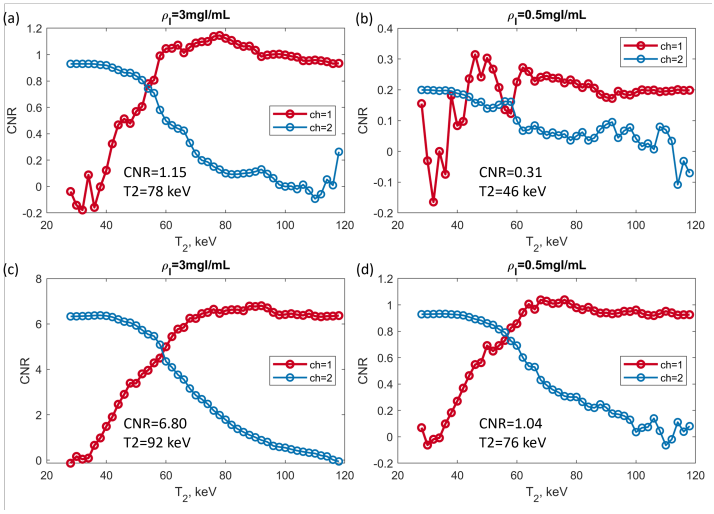


Figure 3: CNR curves of each energy channel in 2-channel PCD-CT as a function of energy threshold T_2 for selected iodine concentrations at two CT dose levels. Top row: $c = 10$, bottom row: $c = 300$.

3.2 2-channel PCD-CT

For the 2-channel PCD-CT scenario, the maximum CNR achievable is noticeably improved from the 1-channel case for all concentrations, as shown in Figure 2. Note that the maximum CNR of 2-channel PCD-CT was defined as the higher CNR between the two channels. Compared with the 1-channel case, although the 1-channel case used photons from the entire spectrum, yielding lower noise level, it deteriorates the contrast achieved from using photons at low energies.

To further investigate this issue, we plotted the CNR obtained within each energy channel for all cases with different thresholds in Figure 3. The 'X'-like shape of the curves is expected, since the CNR is determined by both contrast and noise. A narrow energy channel has a lower noise level, even if it offers high contrast. Nevertheless, the best CNR is usually from the first channel that covers the low energy range, when the noise no longer dominates. The maximum CNR of 3 mgI/mL iodine solution achieved with 2-channel PCD-CT was 23.9% larger than the EID-CT's CNR at $c = 300$, and about 243.1% larger at $c = 10$; For 0.5 mgI/mL, 26% and 131.2% improvement can be achieved with 2-channel PCD at c=300 and 10 respectively.

3.3 3-channel PCD-CT

Similarly, we plotted the maximum CNR among all the 3-channel for the 3-channel PCD-CT cases in Figure 2. Compared with the 2-channel PCD-CT cases, while the 3-channel PCD-CT has one more degree of freedom to further optimize the energy thresholds, the maximum CNRs for all concentrations were only slightly better than the ones of 2-channel PCD-CT.

Here again, we depicted the CNR maps as functions of two energy thresholds T_2 and T_3 under two CT doses for 0.5 and 3 mg/mL iodine solutions in Figure 4. One interesting discovery was that competition always occurs between the first two channels simply because the third channel cannot provide sufficient contrast. Here again, when $\rho_I = 3$ mgI/mL, the maximum CNR achieved with 3-channel PCD-CT was 24.5% higher than the EID-CT's CNR if $c = 300$, and about 243.8% higher if $c = 10$. When $\rho_I = 0.5$ mgI/mL, the maximum CNR achieved with 3-channel PCD-CT was 26.4% higher than the EID-CT's CNR if $c = 300$, and about 131.6% greater if $c = 10$.

3.4 PCD-CT and EID-CT image comparisons

Lastly, the CT images reconstructed with 1-3 channel PCD-CT and EID-CT are displayed in Figure 5 for a visual comparison. Note that for 2 and 3-channel PCD-CT, we display the CT image for the energy channel that maximized the CNR of the 3 mgI/mL insert. Clearly, the PCD-CT outperforms the EID-CT in this detectability task especially for low CT dose cases mainly due to the absence of electronic noise. The iodine inserts with 2 and 3 mgI/mL were clear visible from the CT images with $c = 10$, while not from the EID-CT image. The redundant photons at high energies impact the contrast, which is obvious by comparing the brightness of the inserts at high concentration levels (outer ring), i.e. Figure 5 (g) and (h). At a relatively high CT dose level (30 mGy), the CNR of iodine with ultra-low concentrations (< 3 mgI/mL) improves significantly.

4 Discussion

It is worth mentioning that the simulation method in this work did not consider the scattering of the photon in calculating the detector signals. Scattering is one of the major sources downgrading the reconstructed CT image quality. A common remedy is to use the anti-scatter grid to reduce the scatter photon signals. Therefore, conclusions derived from this work would still be valid for the actual clinical cases.

Additionally, the detector response function was assumed to be an ideal value of unity within the energy channel thresholds. This was an approximation of the detector responses in the real-world scenario. The actual detector response may impact the energy resolution. For example, a 40 keV photon could be counted when the threshold was set to 50

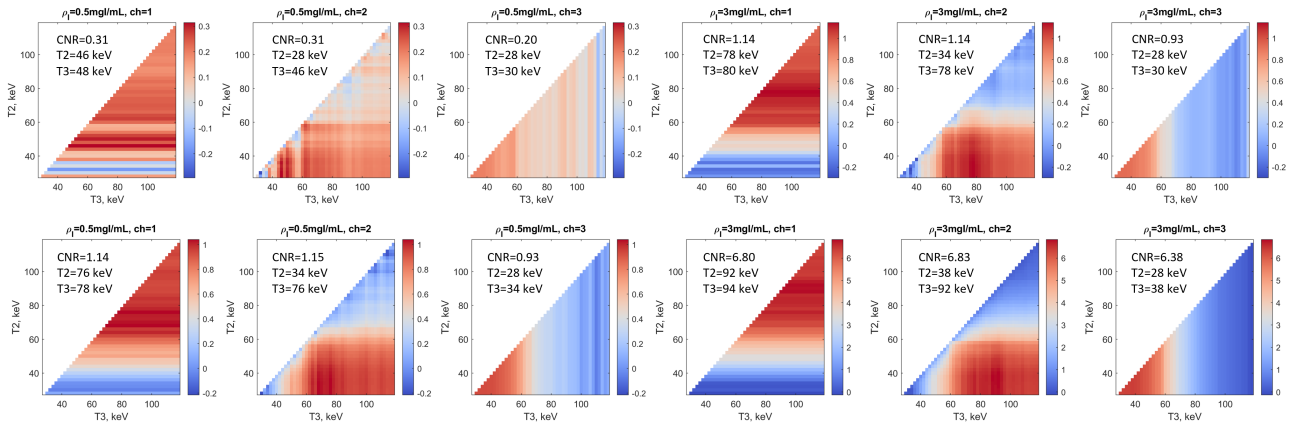


Figure 4: CNR map of 3 energy channels for selected concentrations of iodine for the 3-channel PCD-CT cases. Top row: $c = 10$. Bottom row: $c = 300$. The text insert in each subfigure displays the maximum CNR and its associated energy thresholds.

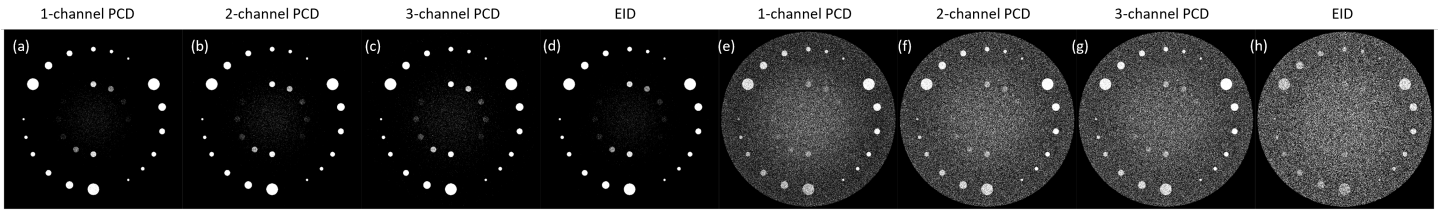


Figure 5: (a)-(d): CT images with $c = 300$, W/L:50/30 HU. (e)-(h): CT images with $c = 10$, W/L:200/30 HU.

keV. The inferior energy resolution lowers the contrast in the CT images. We have conducted an experimental study to measure the detector response function for different energy thresholds. It is our ongoing work to include this feature to our simulation pipeline to understand the impact of detector responses.

The fact that superior CNR was found in one of the energy channels of the multi-channel PCD-CT indicates that we need to abandon some photons to improve CNR. This means that there is room for further improving the setting like the x-ray spectrum and filtrations. A more interesting question is whether we can optimize the energy weights $W(E)$ on the detector as it counts photons to achieve a better CNR? There have been studies to investigate this question via simulation, although it is still a technical challenge to achieve a designed energy weighting by the circuits on the PCD. The simulation pipeline here can be employed to investigate this issue.

5 Conclusion

In this work, we developed a simulation pipeline to investigate the detectability of contrast agents with ultra-low concentrations. The iodine solutions of various concentrations were selected as the object, and 1/2/3-channel PCD and EID CTs were modeled to compute the CNR of these iodine solutions. Numerical results show that the PCD CT outperforms the EID CT in terms of CNR in general. Quantitatively, when the $\text{CTDI}_{free\ air}$ is 3.3 mGy, the 1-channel PCD-CT gains 46% higher CNR than the EID for 0.5% mgI.mL solution, and 178.9% for 3 mgI/mL. When the $\text{CTDI}_{free\ air}$ is 98.7 mGy, the CNR of 1-channel PCD-CT is close to that of EID. For 0.5 and 3 mgI/mL solutions, the CNR in the 1-channel PCD-CT is 14% and 5% greater than the EID CNRs. The 2/3-channel PCD-CT is able to improve the CNR about 26% and 24% for 0.5 and 3 mgI/mL solution respectively. The multi-channel PCD provides extra degrees of freedom in the energy dimension for finding the optimal channels.

References

[1] S. Leng, M. Bruesewitz, S. Tao, et al. “Photon-counting Detector CT: System Design and Clinical Applications of an Emerging Technology”. *RadioGraphics* 39.3 (2019), pp. 729–743. DOI: 10.1148/rg.2019180115.

[2] M. Danielsson, M. Persson, and M. Sjölin. “Photon-counting x-ray detectors for CT”. *Physics in Medicine &amp Biology* 66.3 (Jan. 2021), 03TR01. DOI: 10.1088/1361-6560/abc5a5.

[3] X. Hu, Y. Zhong, Y. Lai, et al. “Small animal photon counting cone-beam CT on a preclinical radiation research platform to improve radiation dose calculation accuracy”. *Physics in Medicine Biology* 67.19 (Sept. 2022), p. 195004. DOI: 10.1088/1361-6560/ac9176.

[4] L. Caschera, A. Lazzara, L. Piergallini, et al. “Contrast agents in diagnostic imaging: Present and future”. *Pharmacological research* 110 (2016), pp. 65–75.

[5] M. Andreucci, R. Solomon, and A. Tasanarong. “Side Effects of Radiographic Contrast Media: Pathogenesis, Risk Factors, and Prevention”. *BioMed Research International* 741018 (2014). DOI: 10.1155/2014/741018.

[6] K. Yang, X. Li, X. George Xu, et al. “Direct and fast measurement of CT beam filter profiles with simultaneous geometrical calibration”. *Medical Physics* 44.1 (2017), pp. 57–70. DOI: https://doi.org/10.1002/mp.12024.

Unraveling Interstitial Fluid Pressure: Analyzing Drug Transport through Cross Voxel Exchange Model and Dynamic Contrast Enhanced MRI

Yeyoung Janny Kim[1,2], Noha Sinno[2], Michael Milosevic[2,3,4] and Catherine Coolens[1,2,4]

[1]Department of Medical Biophysics, University of Toronto, Toronto, ON, Canada
[2]Princess Margaret Cancer Centre, University Health Network, Toronto, ON, Canada
[3]Institute of Medical Science, University of Toronto, Toronto, ON, Canada
[4]Department of Radiation Oncology, University of Toronto, Toronto, ON, Canada

Abstract Cancer treatment encounters challenges due to tumour heterogeneity, leading to unpredictable patient outcomes. Interstitial fluid pressure (IFP) is reported to play a role in drug resistance in solid tumors. To examine the relationship between IFP, and drug transport, the novel Cross-Voxel eXchange Model (CVXM) was applied. CVXM created various image-derived transport parameter maps, such as extravasation, convection, and diffusion maps. The fitted results were compared to the Tofts model. The CVXM convection map was converted to a gradient pressure map, which was further compared with the actual IFP measurements obtained using a microtip pressure transducer. Characteristics of IFP gradient, where the velocity flow is elevated near the periphery and decreased at the tumour center, were both found from CVXM convection maps and actual IFP measurements. CVXM provides unparallel information in understanding the pharmacokinetics of drugs in the tumour microenvironment, which will help us move closer towards personalized cancer treatment.

1 Introduction

Despite immense progress in cancer treatment, drug resistance in cancer patients remains a challenge attributed to tumour heterogeneity[1]. Among various factors that contribute to tumour heterogeneity, studies have increasingly pointed towards interstitial fluid pressure (IFP) as a main cause of drug resistance in cancer patients[2,3,4,5,6]. As solid tumours grow, they accumulate IFP at the center due to defective vasculature and lymphatic functions. This pressure gradient induces fluid convection in the surrounding normal tissue, creating a barrier obstructing the entry of large therapeutic molecules like immunotherapeutic monoclonal antibodies, nanoparticles, and radionuclides[6,7,8]. To comprehend the interplay between IFP and drug resistance, a thorough investigation into their transport mechanisms is needed.

Dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI) is a non-invasive and non-ionizing imaging technique that offers quantitative analyses into the biophysical properties of drug (tracer) transport within the tumor microenvironment. Conventionally, the widely used two-compartmental Tofts model focused solely on intra-voxel exchange, representing extravasation. Augmenting from the Tofts model (TM), Sinno et al. developed the novel Cross-Voxel eXchange Model (CVXM), allowing for the quantification of various pharmacokinetic parameters of tracer transport between blood vessels and tissue, such as diffusion, convection, and extravasation[9,10,11,12]. It is hypothesized that the image-derived transport parametric maps from DCE-MRI and CVXM, will help explore the intricate relationships between drug delivery, IFP, and the tumor microenvironment.

2 Materials and Methods

2.1. Tumour model

The ME180 human cervical carcinoma xenograft model was used for in vivo studies as it closely resembles clinical tumours and as IFP of ME180 has been extensively investigated[13,14]. ME180 tumours were initiated by inoculating the mice intramuscularly with $1x10^6$ cells. The intramuscular donor tumour fragments were then used for orthotopic implantation in mouse cervix. All in vivo experimental protocols in the study were approved by the University Health Network Institutional Animal Care and Use Committee.

2.2. DCE-MRI and IFP measurement

The mice (n=2) underwent imaging when the tumour volume reached 800 cm^3 using a 7T pre-clinical scanner (Biospec 70/30 USR, Bruker) located at the STTARR facility. DCE-MRI was conducted using Gd-DTPA 1mmol/mL (Gadovist, Bayer) was used as a contrast agent. 40 microliter saline was administered to flush 15 microliter of Gd-DTPA via tail vein injection in a bolus. T1-weighted images were acquired (TR=109ms, TE=1.737ms, α=45°) with a time resolution of 14 seconds for a total scan time of 28 minutes. The spatial resolution was $0.225x0.225x0.8mm^3$. The T10 maps were acquired using the variable TR technique for every mouse before contrast administration. IFP measurements were taken immediately after DCE-MRI using a Millar microtip pressure transducer with a probe diameter of 0.33mm (SPR-1000, Millar). The microtip probe was placed in a 21-gauge needle which assists the probe to be inserted into the tumour. Measurements were taken under ultrasound guidance (Vevo 2100, VisualSonics) for probe visualization. Preliminary results were done with a higher sample size (n=6) to test reproducibility of IFP measurement with the microtip pressure transducer (data not shown).

2.3. Model fitting

Elliptical region-of-interest (ROI) was drawn for each tumour using MATLAB assuming a radially symmetric and

elliptical tumour. Concentration of the contrast agent per voxel was calculated from the signal intensity using the non-linear relationship described in Landis et al.[15]. The dynamic change in concentration across timepoints was processed to generate the pharmacokinetic parameters using both the TM[16] and CVXM. CVXM was performed as described in Sinno et al.[12]. The velocity map derived from CVXM was used to identify the calculated IFP using Darcy's law: $v = -K\nabla P$, where v is velocity, K is the hydraulic conductivity coefficient and P is pressure. $7x10^{-7}$ $cm^2/mmHg{\cdot}s$ was used for the value of K[17].

3 Results

3.1. Observation of tracer uptake

The average change in concentration of Gd-DTPA over time is shown in Figure 1. It was observed that the tumour periphery achieved maximum uptake of tracer faster than the tumour center. At 7 minutes, the mean concentration at tumour periphery reached maximum, while the concentration at the center is approximately 2-fold lower relative to the periphery. It indicates that the tumour periphery had more well-defined vasculature, allowing for the tracer to infiltrate easily. On the other hand, the tumour center showed more gradual uptake of the tracer. This indicates an obstruction likely due to a build-up of IFP at the tumour center, caused limited tracer enhancement.

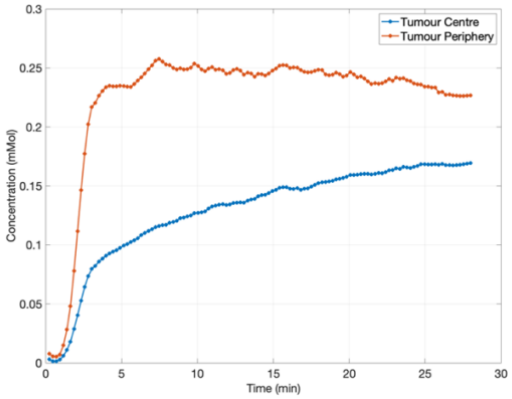


Figure 1. The mean concentration of the tracer at the tumour center and the periphery across timepoints. The plot shows the average of the samples.

3.2. Fitting results

Figure 2 displays various parametric maps of tracer transport derived from the TM and CVXM. Kext and Ktrans, representing tumour extravasation showed the same pattern in CVXM and TM. It showed high values near the periphery than the center, which agrees with the concentration uptake curve (Fig 1). Small values of Kext were observed in the areas with poor vasculature, indicating lack of blood flow at the tumour center. The value for Ve was chosen at a constant value of 0.1 for CVXM. Tofts' Ve shows heterogeneous distribution, values ranging from 0 to 3.5 on average. The mean diffusivity from CVXM was 230 ± 2.8 $\mu m/s^2$.Velocity was measured using CVXM. It was illustrated across all samples that the tumour periphery had higher velocity than the tumour center. The maximum velocity was determined to be 8.76 μm/s and 6.68 μm/s in mouse 1 and mouse 2 respectively, which fall in a common range of tracer velocity previously reported in the literature. From CVXM velocity map, null velocities were observed at the tumour center, which is expected since IFP gradient at the center is zero[2].

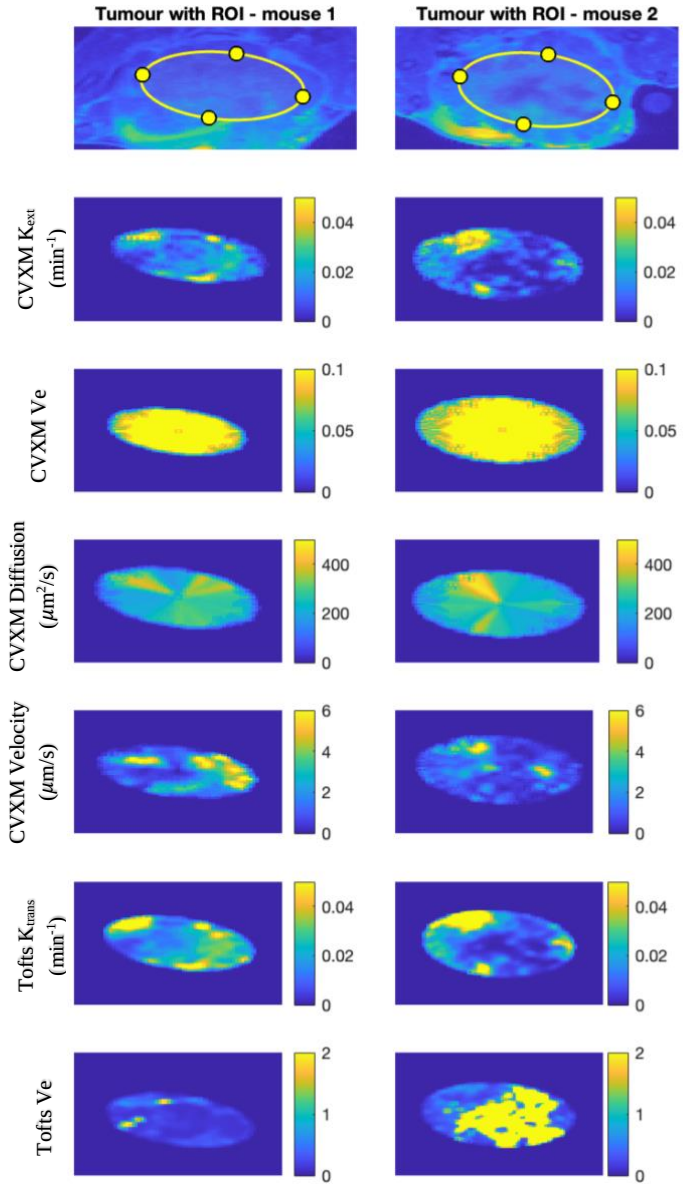


Figure 2. DCE-MRI images of the tumours at the last timepoint with ROIs drawn with yellow outline followed by the transport parametric maps.

3.3. IFP measurement

Using the microtip pressure transducer, the ranges of actual IFP were determined to be from 3.5 to 24mmHg for tumour center. Pressure measurements in the tumour periphery ranged from 0.3 to 2.3 mmHg. The mean of actual IFP was measured to be approximately 11.8 ± 6.9mmHg and 1.12 ± 0.8mmHg in the tumour center and periphery, respectively. The actual IFP values of the tumours are

summarized in Table 1. The ranges and means of IFP measured, agreed with the results from Lunt et al.[13]. IFP measurements were performed under ultrasound guidance. Figure 3 is an ultrasound of the tumour with the pressure probe inserted into the tumour. Ultrasound imaging allowed visualization of the tumour boundary. The gradient pressure map is shown in Figure 4, which was generated from the CVXM velocity map through Darcy's law. Most regions ranged from 0 to -50mmHg in gradient pressure, while some locations showed more drastic changes in pressures at the periphery.

Table 1. Summarized actual IFP values of the tumours measured using the microtip pressure transducer. The last row shows the mean of the samples (n=2).

	Center IFP (mmHg)	Periphery IFP (mmHg)
Mouse 1	11±7.3	1.6±1.0
Mouse 2	12.7±7.1	0.81±0.6
Mean	11.8±6.9	1.12±0.8

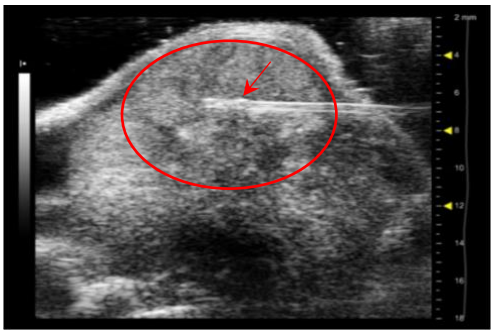


Figure 3. Ultrasound image of a mouse cervical tumour with the microtip pressure transducer inserted at the tumour center. Tumour is shown with the red ellipse and the probe is indicated with the red arrow.

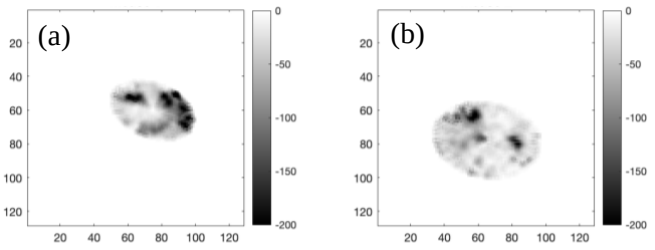


Figure 4. Pressure gradient maps of (a) mouse 1 and (b) mouse 2. Darker regions indicate rapid changes in pressure.

4 Discussion

Using CVXM, extravasation, convection and diffusion were quantified in ME180 tumours. The TM is unable to quantitate cross-voxel exchange parameters, while CVXM can describe these dynamic changes. Extravasation showed higher values at the periphery and lower values at the tumour center. This distribution pattern was observed in both CVXM Kext and Tofts Ktrans maps, and the fitted results are in alignment with the tracer uptake seen in the mean tracer concentration curve. Ve in CVXM was kept at a constant 0.1. The range of Ve should be from 0 to 1, however, the TM showed values greater than 3.5 in some areas. This characteristic originated from combinatory effect of diffusion and large gradient of Kext in some areas. These were unphysiological values, indicating a limitation in the TM. Sinno et al.[12] explains this effect of Tofts Ve further.

Convection was represented by the CVXM velocity map. It accurately showed increased flow rate at the periphery and smaller flow rate at the center. The gradient pressure map obtained also showed the same pattern. The periphery exhibited a more rapid tracer flow, which was governed by the high-pressure build-up at the center. The pattern in the velocity map and the pressure gradient map concurs with the reports of IFP gradient. IFP in tumours have been reported to have relatively homogeneous center until it reaches the tumour periphery near the normal tissue[2]. IFP is expected to drop sharply at the boundary near the normal tissue. This characteristic was further verified with the microtip pressure transducer, where a drop in pressure was observed from center to periphery.

CVXM's ability to study cross-voxel transports allow quantification of diffusion in the tumour. For a mean value of 230 μm/s^2, the diffused distance of tracer is about 5.5 voxels during the 28-minute scan time ($\sqrt{r^2} = \sqrt{4Dt}$). This was not an enough time to observe the complete action of diffusion reaching the center of the tumour. In addition, the tumour size chosen in the study was bigger than the tumours previously studied in other papers[18,19]. It can be speculated from the mean concentration (Fig 1) that that the primary means of transport at the tumour center was diffusion, rather than extravasation. With poor vasculature, the tracer is likely to be transported via diffusion at a slower rate than via extravasation. Large tumour size also increases the potential for tumour necrosis. It was observed that some voxels at the tumour center did not reach maximum tracer enhancement, while the peak enhancement reached around 7 minutes post-injection for voxels at tumour periphery. This suggests the possibility of necrotic tumour center and tracers primarily getting delivered via diffusion. The bigger tumour was a trade-off against easier access to the tumour for IFP measurement using the pressure probe. For future study, the experiments will be repeated for higher sample size (n=9). The data are acquired and are currently being analyzed.

5 Conclusion

IFP influences drug resistance in solid tumors. DCE-MRI combined with CVXM, aids in exploring cross-voxel

mechanisms of drug transport, including extravasation, convection, and diffusion. The calculated extravasation using CVXM showed similar trend to Tofts Ktrans. CVXM provides additional parametric maps to what the TM offers. Convection map from CVXM accurately illustrated the IFP gradient, showing increased velocity flow at the periphery and decreased velocity at the center. This characteristic of IFP gradient was further confirmed by microtip pressure transducer measurements. Diffusion map was also generated using CVXM. The findings from CVXM can help understand the heterogeneity in tumours and gauge the aggressiveness of the tumour. This comprehension will pave the way for personalized treatments, such as immunotherapy or drug delivery through nanoparticles, customized to the unique characteristics of each patient's tumour microenvironment. This approach holds the potential to enhance the efficacy of individualized cancer therapies.

References

[1] Zhang, A., Miao, K., Sun, H., & Deng, C. X. (2022). Tumor heterogeneity reshapes the tumor microenvironment to influence drug resistance. In International Journal of Biological Sciences, 18(7), p. 3019–3033. https://doi.org/10.7150/ijbs.72534

[2] Jain, R. K. (1987). *Transport of Molecules in the Tumor Interstitium: A Review1* (CANCER RESEARCH, Vol. 47, pp. 3039–3051). http://aacrjournals.org/cancerres/article-pdf/47/12/3039/2428062/cr0470123039.pdf

[3] Netti, P. A., Baxter, L. T., & Boucher, Y. (1995). Time-dependent Behavior of Interstitial Fluid Pressure in Solid Tumors: Implications for Drug Delivery.

[4] Heldin, C. H., Rubin, K., Pietras, K., & Östman, A. (2004). High interstitial fluid pressure—An obstacle in cancer therapy. *Nature Reviews Cancer*, *4*(10), 806–813. https://doi.org/10.1038/nrc1456

[5] Milosevic, M., Fyles, A., Hedley, D., & Hill, R. (2004). The human tumor microenvironment: Invasive (needle) measurement of oxygen and interstitial fluid pressure. *Seminars in Radiation Oncology*, *14*(3), 249–258. https://doi.org/10.1016/j.semradonc.2004.04.006

[6] Stapleton, S., Milosevic, M., Tannock, I. F., Allen, C., & Jaffray, D. A. (2015). The intra-tumoral relationship between microcirculation, interstitial fluid pressure and liposome accumulation. *Journal of Controlled Release*, *211*,163–170. https://doi.org/10.1016/j.jconrel.2015.06.008

[7] Baronzio, G., Parmar, G., & Baronzio, M. (2015). Overview of methods for overcoming hindrance to drug delivery to tumors, with special attention to tumor interstitial fluid. *Frontiers in Oncology*, *5*(JUL). https://doi.org/10.3389/fonc.2015.00165

[8] Zaheer, J., Kim, H., Lee, Y. J., Kim, J. S., & Lim, S. M. (2019). Combination radioimmunotherapy strategies for solid tumors. *International Journal of Molecular Sciences*, *20*(22). https://doi.org/10.3390/ijms20225579

[9] Coolens, C., Driscoll, B., Foltz, W., Pellow, C., Menard, C., & Chung, C. (2016). Comparison of Voxel-Wise Tumor Perfusion Changes Measured with Dynamic Contrast-Enhanced (DCE) MRI and Volumetric DCE CT in Patients with Metastatic Brain Cancer Treated with Radiosurgery. *Tomography*, *2*(4), 325–333. https://doi.org/10.18383/j.tom.2016.00178

[10] Foltz, W., Driscoll, B., Lee, S. L., Nayak, K., Nallapareddy, N., Fatemi, A., Ménard, C., Coolens, C., & Chung, C. (2019). Phantom Validation of DCE-MRI Magnitude and Phase-Based Vascular Input Function Measurements. *Tomography*, *5*(1), 77–89. https://doi.org/10.18383/j.tom.2019.00001

[11] Coolens, C., Driscoll, B., Foltz, W., Svistoun, I., Sinno, N., & Chung, C. (2019). Unified platform for multimodal voxel-based analysis to evaluate tumour perfusion and diffusion characteristics before and after radiation treatment evaluated in metastatic brain cancer. *The British Journal of Radiology*, *92*(1096), 20170461. https://doi.org/10.1259/bjr.20170461

[12] Sinno, N., Taylor, E., Milosevic, M., Jaffray, D. A., & Coolens, C. (2021). Incorporating cross-voxel exchange into the analysis of dynamic contrast-enhanced imaging data: Theory, simulations and experimental results. *Physics in Medicine and Biology*, *66*(20). https://doi.org/10.1088/1361-6560/ac2205

[13] Lunt, S. J., Kalliomaki, T. M. K., Brown, A., Yang, V. X., Milosevic, M., & Hill, R. P. (2008). Interstitial fluid pressure, vascularity and metastasis in ectopic, orthotopic and spontaneous tumours. *BMC Cancer*, *8*, 1–14. https://doi.org/10.1186/1471-2407-8-2

[14] Milosevic, M. F., Fyles, A. W., Wong, R., Pintilie, M., Kavanagh, M. C., Levin, W., Manchul, L. A., Keane, T. J., & Hill, R. P. (1998). Interstitial fluid pressure in cervical carcinoma: Within tumor heterogeneity, and relation to oxygen tension. Cancer, 82(12), 2418–2426. https://doi.org/10.1002/(SICI)1097-0142(19980615)82:12<2418::AID-CNCR16>3.0.CO;2-S

[15] Landis, C. S., Li, X., Telang, F. W., Coderre, J. A., Micca, P. L., Rooney, W. D., Latour, L. L., Vétek, G., Pályka, I., & Springer, C. S. (2000). Determination of the MRI contrast agent concentration time course in vivo following bolus injection: Effect of equilibrium transcytolemmal water exchange. Magnetic Resonance in Medicine, 44(4), 563–574. https://doi.org/10.1002/1522-2594(200010)44:4<563::AID-MRM10>3.0.CO;2-#

[16] Tofts, P. S., Brix, G., Buckley, D. L., Evelhoch, J. L., Henderson, E., Knopp, M. V., Larsson, H. B. W., Lee, T. Y., Mayr, N. A., Parker, G. J. M., Port, R. E., Taylor, J., & Weisskoff, R. M. (1999). Estimating kinetic parameters from dynamic contrast-enhanced T1- weighted MRI of a diffusable tracer: Standardized quantities and symbols. Journal of Magnetic Resonance Imaging, 10(3), 223–232. https://doi.org/10.1002/(SICI)1522-2586(199909)10:3<223::AID-JMRI2>3.0.CO;2-S

[17] Milosevic, M., Lunt, S. J., Leung, E., Skliarenko, J., Shaw, P., Fyles, A., & Hill, R. P. (2008). Interstitial permeability and elasticity in human cervix cancer. *Microvascular Research*, *75*(3), 381–390. https://doi.org/10.1016/j.mvr.2007.11.003

[18] Hompland, T., Ellingsen, C., Øvrebø, K. M., & Rofstad, E. K. (2012). Interstitial fluid pressure and associated lymph node metastasis revealed in tumors by dynamic contrast-enhanced MRI. *Cancer Research*, *72*(19), 4899–4908. https://doi.org/10.1158/0008-5472.CAN-12-0903

[19] Hompland, T., Gulliksrud, K., Ellingsen, C., & Rofstad, E. K. (2013). Assessment of the interstitial fluid pressure of tumors by dynamic contrast-enhanced magnetic resonance imaging with contrast agents of different molecular weights. *Acta Oncologica*, *52*(3), 627–635. https://doi.org/10.3109/0284186X.2012.737931

Revealing head and neck tumors spatial heterogeneity through quantitative MRI protocol: Initial findings

DANDACHLY Walid[1], LEPORQ Benjamin[1], SAYAQUE Laura[1], ALLIGNET Benoit[1,2], PILLEUL Frank[1,3], HAMELIN Olivier[3], BOUYER Charlène[2,3], GREGOIRE Vincent[2], BEUF Olivier[1]

1.Univ Lyon, INSA-Lyon, Université Claude Bernard Lyon 1, UJM-Saint Etienne, CNRS, Inserm, CREATIS UMR 5220, U1294, F-69621, LYON, France ; 2. CRLCC Léon Bérard – Département de Radiothérapie ; 3. CRLCC Léon Bérard – Département de Radiologie

Abstract Nowadays, radiotherapy planning is done on CT images, which results in a low tissue contrast and lack of functional information about the tumor. However, the use of MRI in this context presents a promising avenue for improvement. MRI is known for its superior capability to provide clear visualizations of organs at risk and to offer valuable tumor characterization. Due to all of this, MRI is considered as a valuable tool in improving the accuracy and effectiveness of radiotherapy planning. In this study, a new MRI protocol was introduced that includes the quantitative mapping of 13 unique functional parameters. Through the exploitation of the IVIM theory, blood perfusion and oxygen extraction fraction will be mapped alongside other important parameters, which will allow us to reveal the heterogeneity inside the tumor volume. Our aim is to use the resultant maps to accurately characterize the tumor and identify its status.

1. Introduction

Cancer is one of the principal deaths causes in the world. Head and neck cancer is the seventh most common cancer globally [1]. Due to the high anatomical complexity in this region, therapy treatment is particularly hard, and the need for a high-precision imaging modality that can characterize and gives accurate information about the biological state of the tissues arises.

Currently, radiotherapy planning is based on CT images which has low tissue contrast, and does not provide any functional information about the tumor. MRI has the potential to improve radiotherapy planning thanks to its unique ability to clearly visualize the tissues of the organs at risk, and to characterize the tumor.

It is known that tumors are very heterogeneous, even the same type of tumor can differ significantly between patients, which adds to the complexity of treatment [2]. Therefore, there exist populations of cells within the tumor with distinct biological structures and behaviors. For instance, tumor hypoxia is considered as a tough point for treatment. Low oxygen concentration plays an important role in various brain pathologies. Identifying the hypoxic regions in the tumor could help to predict the outcome of the radiotherapy, and to better modulate the dose deposition since hypoxic regions require two to three times the radiation dose than well-oxygenated regions [3].

Diffusion-weighted imaging (DWI) is an essential MR imaging method in cancer studies, it is based on measuring the Brownian motion of water molecules. The diffusion motion of the water molecules is responsible for the image intensity weighting, which can give us valuable information about tumor cellularity, viability, and microstructural environment [4]. Since blood microcirculation within a voxel mimics the motion of water, DWI will be sensitive to both water diffusion and blood microcirculation. For high b-values, water diffusion is the main contributor to signal decay, but for low b-values, both water diffusion and blood microcirculation will contribute. This is known as the Intravoxel Incoherent Motion theory (IVIM) [5].

Blood microcirculation in tumors is very valuable information in the oncology field. Tumor usually experiences irregular blood flow, blood perfusion is very spatially and temporally heterogeneous [6]. The heterogeneity of blood flow in the tumor volume will block the access of therapeutic agents to tumors, and results in the creation of distinct microenvironments within the tumor. The heterogeneity inside the tumor volume will increase tumor resistance and reduce the effectiveness of various therapies.

Another important parameter is the oxygen extraction fraction, OEF is a measure of the quantity of oxygen extracted by tissues from the bloodstream. It is mainly used in the neuroimaging field to monitor brain activity [7]. The brain consumes oxygen constantly to maintain its function, and the OEF represents the ability of the brain to extract oxygen from the blood.

In MRI, the main method used to measure the OEF is qBOLD, quantitative Blood Oxygenation Level Dependent. It is a non-invasive technique founded on the fact that any change in neural activity will result in a

change in blood flow and oxygen concentration in the brain. BOLD is based on the difference in magnetic susceptibility between oxygenated hemoglobin which is diamagnetic, and deoxygenated hemoglobin which is paramagnetic. This difference results in a local change of the magnetic field, which affects the signal in T2*-weighted imaging sequences.

The present work evaluates the feasibility of a multiparametric quantitative MRI acquisition and processing pipeline to reconstruct multiparametric maps including an oxygenation measurement to map head and neck tumor biology.

2. Materials and Methods

The designed protocol is dedicated for head and neck radiotherapy planning, and it consists of the following acquisitions:

- T_2, acquired using a 2D multiple echoes sequence, with a scan duration of 5 minutes 20 seconds.
- Chemical shift encoded, acquired using a flash 3D sequence, with a scan duration of 1 minute 21 seconds.
- Diffusion weighted, acquired using a WIP body diffusion MOCCO (990), with a scan duration of 7 minutes 51 seconds.

All the acquisitions were done on a Siemens Vida 3 tesla MR system (Siemens Healthineers, Erlangen, Germany) at the comprehensive cancer center Léon Bérard, with a frequency of one to two patients per week. All of the enrolled patients are diagnosed with a primary squamous cell carcinoma of the hypopharynx, larynx, oropharynx, or undifferentiated nasopharyngeal carcinoma. The patients were fully informed about the study, and a written consent was obtained from them according to the declaration of Helsinki. An immobilizer mask was used with each patient to reduce the motion and to simulate the treatment position, also a knee support was used to provide more comfort.

The delineation of the organs at risk and tumor volume was done on the CT images by two radiations oncologists. The following volumes were delineated: the organs at risk including the spinal cord, the parotid, the optic nerve, the brain stem, and the chiasma. The tumor volume including the gross tumor volume GTV, the low and high risk primary clinical target volume CTV. A landmark registration was performed on the GTV so it can be used with the MR acquisitions and the quantitative maps later.

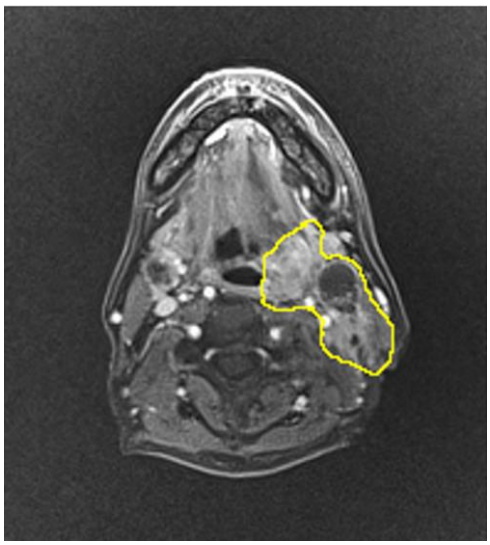


Figure 1: T1 StarVIBE image with the borders of the gross tumor volume indicated in yellow.

The R_2 mapping was obtained using a 2D TSE multi-echoes sequence, with 10 echoes ranging from 8 ms to 80 ms and having an interval of 8 ms. The relaxation rate can be calculated using the following equation.

$$S(TE) = M_0 \, exp\left(-TE/T2\right).$$

With TE being the echo time and M_0 the initial magnitude of the magnetization.

To estimate the blood volume, we first need to estimate the perfusion fraction, this can be done using the IVIM model (figure 2), with the following equation:

$$\frac{S}{S_0} = f * exp - bD * + (1 - f) * exp - bD.$$

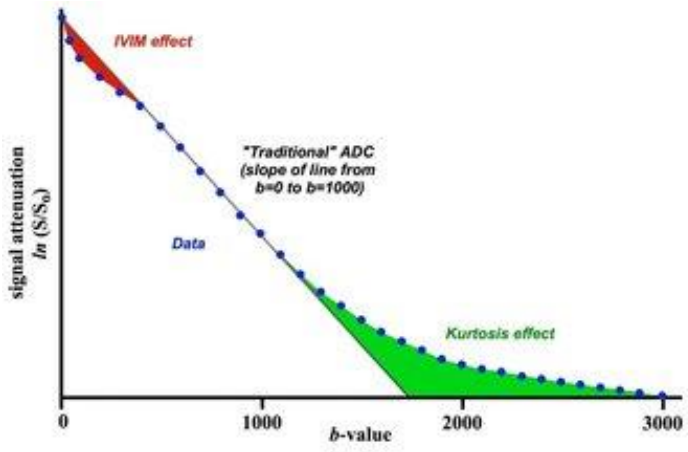


Figure 2: Decay curve of the MRI signal in diffusion-weighted images with the IVIM model (image source: https://mriquestions.com/ivim.html).

Where S_0 is the initial signal intensity, the b-value represents the magnitude and the period of the applied diffusion (i.e. the diffusion weighting in $s.mm^{-2}$) , S is the signal intensity for a given b-value, D is the molecular Gaussian diffusion related to the Brownian motion, D* is the pseudo-diffusion coefficient related to blood flow capillaries [8,9].

Now having all the necessary parameters, the estimation of the oxygen extraction fraction is straightforward.

3. Results

Using an in-house application developed in MATLAB R2023b (Matlab, Natick, MA, USA), 13 quantitative parametric maps were computed from the gross tumor volume of each patient including: blood volume, diffusion coefficient, pseudo-diffusion coefficient, fat only, water only, PDFF, M0, T_2, T_2', R_2', perfusion fraction, relative oxygen extraction fraction and venous oxygen saturation (Figure 3).

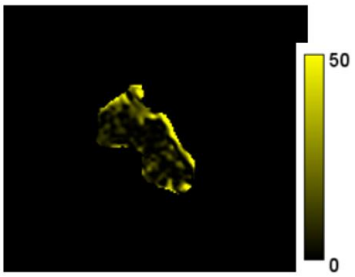


Figure 3-1: PDFF.

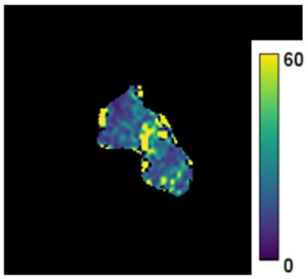


Figure 3-2: T_2.*

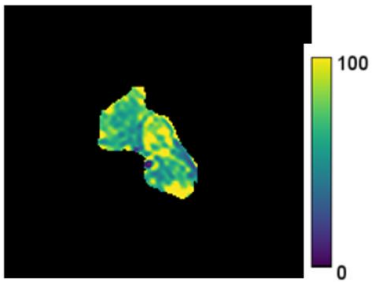


Figure 3-3: T_2.

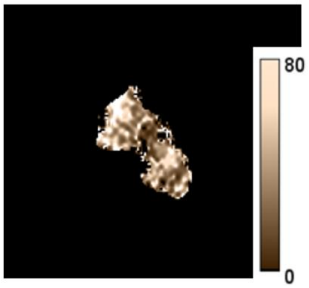


Figure 3-4: R_2`.

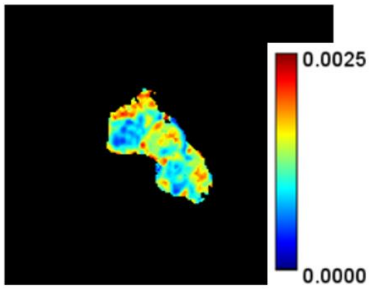


Figure 3-5: Pseudo diffusion coefficient.

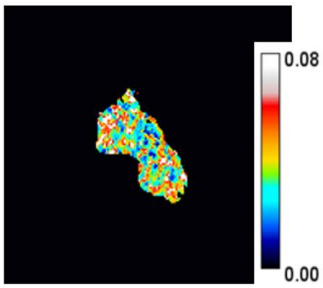


Figure 3-6: Diffusion coefficient.

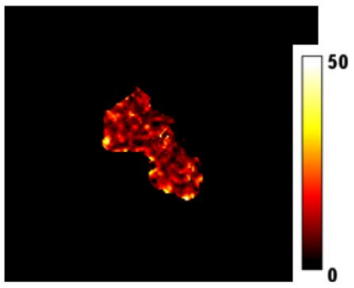


Figure 3-7: Blood volume.

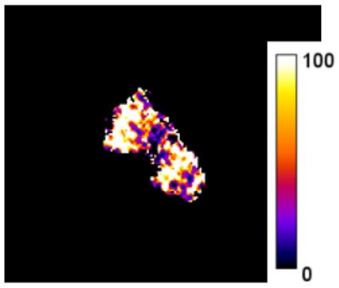


Figure 3-8: rOEF.

4. Discussion

Each map provides very important information about the tumor and it can help determine the tumor cellularity, viability, and microstructural environment, fat content and oxygeneation. For instance, the oxygen extraction fraction map reflects the hypoxic and normoxic regions inside the tumor which is a crucial information to determine the tumor's resistance to radiotherapy. These maps will can be used to accurately characterize the tumor spatial heterogeneity and personalize the therapy based on each patient case using habitat-imaging approach.

5. Conclusion

This study demonstrates the feasibility of multiparametric MRI in radiotherapy planning context, and suggests additional interests in the exploitation of quantitative MRI to improve the treatment course by spatially mapping of the tumor biology.

Acknowledgment

This work was funded by project HERESP (ANR-22-CE19-0027-01) and performed within the framework of LabEx PRIMES (ANR-11-LABX-0063) and SIRIC LyriCAN+ (INCa INSERM DGOS 12563) grants.

References

[1] Bray, F. F. J. S., et al. "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries (vol 68, pg 394, 2018)." *Ca-a Cancer Journal for Clinicians* 70.4 (2020): 313-313.

[2] Dagogo-Jack, Ibiayi, and Alice T. Shaw. "Tumour heterogeneity and resistance to cancer therapies." *Nature reviews Clinical oncology* 15.2 (2018): 81-94.

[3] Gérard, Michael, et al. "Hypoxia imaging and adaptive radiotherapy: a state-of-the-art approach in the management of glioma." *Frontiers in medicine* 6 (2019): 117.

[4] White, Nathan S., et al. "Diffusion-weighted imaging in cancer: physical foundations and applications of restriction spectrum imaging." *Cancer research* 74.17 (2014): 4638-4652.

[5] Le Bihan, Denis. "What can we see with IVIM MRI?." *Neuroimage* 187 (2019): 56-67.

[6] Brown, E. B., Campbell, R. B., Tsuzuki, Y., Xu, L., Carmeliet, P., Fukumura, D., & Jain, R. K. (2001). In vivo measurement of gene expression, angiogenesis and physiological function in tumors using multiphoton laser scanning microscopy. *Nature medicine*, *7*(7), 864-868.

[7] Aanerud, Joel, et al. "Brain energy metabolism and blood flow differences in healthy aging." *Journal of Cerebral Blood Flow & Metabolism* 32.7 (2012): 1177-1187.

[8] De Bazelaire, Cedric MJ, Guillaume D. Duhamel, Neil M. Rofsky, and David C. Alsop. "MR imaging relaxation times of abdominal and pelvic tissues measured in vivo at 3.0 T: preliminary results." *Radiology* 230, no. 3 (2004): 652-659.

[9] B. Leporq, H. Saint-Jalmes, C. Rabrait, F. Pilleul, O. Guillaud, J. Dumortier, JY. Scoazec, O. Beuf. Optimization of intra-voxel incoherent motion imaging at 3.0 Tesla for fast liver examination. *J Magn Reson Imaging* 41(5): 1209–1217 (2015).

Segment anything model for head and neck tumor segmentation with CT, PET and MRI multi-modality images

Jintao Ren[1,2], Mathis Rasmussen[1,2], Jasper Nijkamp[1,2], Jesper Grau Eriksen[1,3] and Stine Korreman[1,2]

[1]Department of Clinical Medicine, Aarhus University, Aarhus, Denmark
[2]Danish Center for Particle Therapy, Aarhus University Hospital, Aarhus, Denmark
[3]Department of Experimental Clinical Oncology, Aarhus University Hospital, Aarhus, Denmark

Abstract Deep learning presents novel opportunities for the auto-segmentation of gross tumor volume (GTV) in head and neck cancer (HNC), yet fully automatic methods usually necessitate significant manual refinement. This study investigates the Segment Anything Model (SAM), recognized for requiring minimal human prompting and its zero-shot generalization ability across natural images. We specifically examine MedSAM, a version of SAM fine-tuned with large-scale public medical images. Despite its progress, the integration of multi-modality images (CT, PET, MRI) for effective GTV delineation remains a challenge. Focusing on SAM's application in HNC GTV segmentation, we assess its performance in both zero-shot and fine-tuned scenarios using single (CT-only) and fused multi-modality images. Our study demonstrates that fine-tuning SAM significantly enhances its segmentation accuracy, building upon the already effective zero-shot results achieved with bounding box prompts. These findings open a promising avenue for semi-automatic HNC GTV segmentation.

1 Introduction

In radiotherapy, delineating the gross tumor volume (GTV) for head and neck cancer (HNC) is a critical yet challenging component of treatment planning. This complexity arises due to the proximity of adjacent organs at risk (OARs) within a constrained space and the tumors' heterogeneous appearance. Manual GTV delineation is not only time-consuming and dependent on a high level of expertise but also requires the integration of multi-modality images such as CT, PET, and MRI. Furthermore, the process is prone to inter-observer variability, which can significantly impact treatment outcomes. Consequently, there is a pressing need for a fully or semi-automatic approach that can streamline the delineation process, enhance consistency, and improve clinical workflow.

Deep learning segmentation models are increasingly recognized for their ability to discern complex image features from extensive datasets of historical images and annotations. Despite their promise, the application of HNC GTV segmentation often necessitates significant manual adjustments for certain patients[1]. Simplifying the definition of a region of interest (ROI) through bounding boxes has been shown to markedly improve outcomes[2, 3]. Furthermore, incorporating 'human-in-the-loop' interactions not only enhances model accuracy but also its interpretability[4,5,6].

Nonetheless, a major hurdle for these models is their struggle with generalization, particularly when there is a significant divergence between training and testing datasets. These could be for example trained on oropharyngeal tumors and tested on larynx tumors, or inter-modality inferences. The Segment Anything Model (SAM) has emerged as a promising solution, demonstrating exceptional zero-shot learning capabilities, trained on over 1.1 billion segmentation masks and roughly 11 million natural images[7]. SAM supports user prompts via points, bounding boxes, and text, yet its successful deployment in medical imaging (e.g., X-ray, CT, MRI) is hampered by the substantial domain gap between natural and medical images[8].

To bridge the domain gap, recent efforts have focused on fine-tuning both public and private medical image datasets[9, 10]. A notable contribution by Ma et al.[11] introduced MedSAM, a version of the SAM fine-tuned with over 1.5 million medical image-mask pairs, demonstrating significant promise in HNC tumor segmentation using CT images. However, GTV delineation in most hospitals involves integrating multiple registered image modalities (e.g., CT, PET, MRI). Given that SAM was initially trained on natural images with red-green-blue (RGB) channels, and MedSAM was fine-tuned for single-modality scenarios, the challenge of fusing registered multi-modality images remains. Consequently, there may be reservations regarding the efficacy of MedSAM's zero-shot inference on local multi-modality datasets.

In this study, we explore MedSAM's capabilities in segmenting both primary tumors (GTV-T) and the involved lymph nodes (GTV-N) across single- and multi-modality images. We assess the model's performance in zero-shot and fine-tuned scenarios on our local dataset.

2 Materials and Methods

A total of 576 HNC patients from Aarhus University Hospital treated with radiotherapy in the period 2013-2020 were included in this study. Each patient had undergone treatment planning with FDG-PET/CT, T1-weighted (T1w), and T2-weighted (T2w) MRI. The T1w and T2w images were initially registered to the CT images using rigid and then deformable registration methods. The dataset was randomly divided into three subsets: 470 for training, 94 for validation, and 97 for testing. The clinical

treatment planning process involved referencing all image modalities to delineate both GTV-T and GTV-N, with these clinical delineations serving as the ground truth.

2.1 Single- and Multi-modality

Although MedSAM is trained on datasets from various modalities, like the original SAM, it processes inputs from only a single modality at a time. For inputs involving a single modality, we adopted the approach of replicating the CT images across three channels to mimic RGB inputs, a method we refer to as 'CT-only.'

For multi-modality fusion, to accommodate four image modalities while maintaining SAM's simplicity, we devised a unique strategy: CT images were assigned to the red channel, PET images to the green channel and a blend of T1w and T2w MRI images to the blue channel. To integrate the T1w and T2w images, we first normalized them to the same intensity range and then computed the average intensity for each corresponding pixel pair. This approach is termed 'multimodal'.

2.2 Preprocessing and bounding box prompt

For preprocessing, CT images were adjusted with a window level of 40 and a window width of 400 to highlight soft tissue. PET images were clipped to a range of 0-6 SUV[12], while MRI images were adjusted to include intensities from the 0th to the 99.5th percentile of their maximum values. Subsequently, all images from these different modalities were rescaled to a 0-255 range, aligning them with the typical intensity range of natural images. After combining these into RGB channels, each 2D slice was normalized to float values between 0 and 1, preparing them as input for SAM.

Regarding the segmentation label masks, we treated each GTV-T and GTV-N object within the 2D slices as distinct segmentation targets. For SAM's prompting mechanism, we utilized a bounding box strategy. On each 2D slice, we introduced a random perturbation margin ranging from 0 to 20 mm to box out each GTV instance.

2.3 Zero-shot and finetuning of MedSAM model

We utilized the lite version of MedSAM, which employs a tiny vision transformer (TinyViT) model distilled from the larger variant, as our pre-trained SAM model base[13]. To evaluate its zero-shot generalization ability on our local dataset, we performed inference on our test set without any additional training.

For fine-tuning, the pre-trained lite MedSAM model was further trained on our local dataset for up to 20 epochs. We saved the epochs of the model that achieved the lowest Dice Loss on the validation set to ensure optimal performance.

We followed MedSAM's training protocol, employing a hybrid loss function that combines Dice loss, cross-entropy, and IoU loss. The AdamW optimizer was used, with a batch size of 8 and an initial learning rate of 3e-4. This rate was scheduled to reduce by a factor of 0.9 upon hitting a performance plateau, with patience of 5 epochs. The training was executed on two NVIDIA A40 GPUs, each equipped with 40 GB of GPU memory.

For the comparative analysis, we utilized two fully automated configurations of nnUNet[14], processing all image modalities (four image modalities) with the identical data split as SAM finetune. The first configuration functioned without a bounding box (nnUNet w/o box), whereas the second configuration included a bounding box as an additional input channel for each 2D slice, aligning with the approach taken by MedSAM (nnUNet w/ box). Both nnUNet configurations were trained for 1000 epochs in 3D full resolution.

2.4 Evaluation

To evaluate segmentation performance, we computed the mean and one standard deviation for the Dice Similarity Coefficient (DSC) as the primary metric, complemented by the 95% Hausdorff Distance (HD95). These evaluations were conducted on the entire 3D volumes of the images on the test set, comparing the segmented outputs with the ground truth for both GTV-T and GTV-N. Additionally, we monitored the fine-tuning time required for each epoch and the inference time needed to predict each slice.

3 Results

In assessing segmentation performance, we evaluated MedSAM under both zero-shot and fine-tuned conditions, using either CT-only or fused multimodal inputs. We also compared these results against the fully automatic multi-modality nnUNet, both in its standard configuration and an enhanced setup incorporating a bounding box. The outcomes, including both DSC and HD95 metrics, are detailed in Figure 1.

When examining zero-shot outcomes, multimodal inputs slightly outperform CT-only for GTV-T (0.70 vs. 0.69), whereas CT-only surpasses multimodal for GTV-N (0.75 vs. 0.68). However, following fine-tuning, both CT-only and multimodal approaches see a significant boost in segmentation accuracy, as reflected in DSC scores. Specifically, for CT-only inputs, DSC scores increased from 0.69 to 0.82 for GTV-T and from 0.75 to 0.85 for GTV-N. Similarly, multimodal inputs saw DSC improvements from 0.7 to 0.82 for GTV-T and from 0.68 to 0.85 for GTV-N with fine-tuning.

A nnUNet model without a bounding box exhibits lower performance (0.68 for GTV-T and 0.63 for GTV-N) and higher variance, however, the inclusion of bounding boxes

of each target at each 2D slice in nnUNet markedly improves results (0.88 for GTV-T and 0.89 for GTV-N). The HD95 exhibits a trend similar to that observed with DSC.

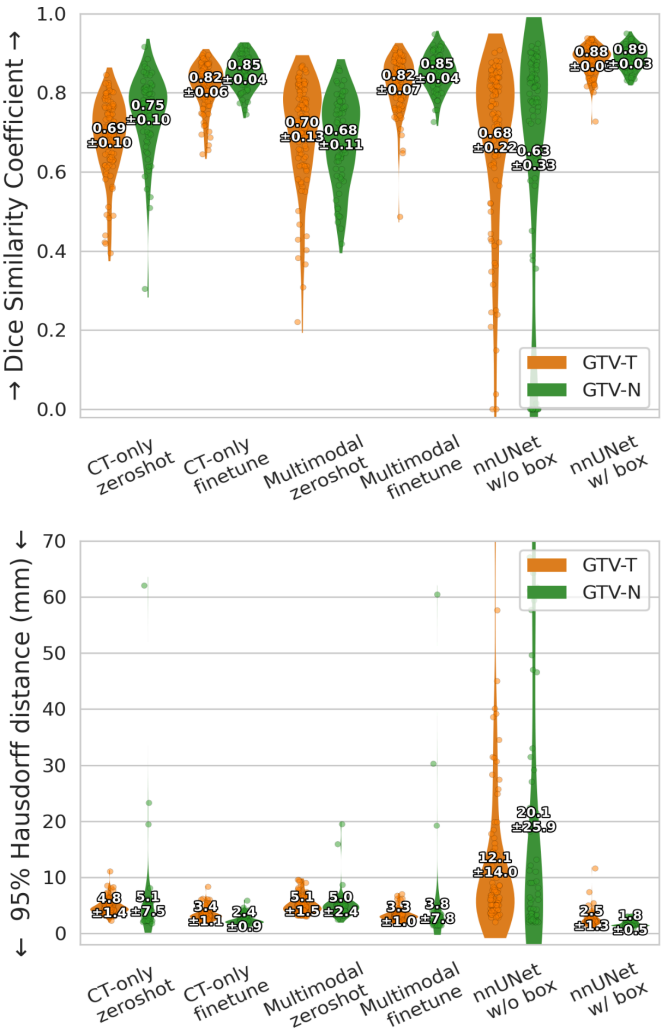


Figure 1: Comparative analysis of segmentation performance across groups for GTV-T and GTV-N, illustrated through violin plots using the Dice Similarity Coefficient (DSC) and the 95% Hausdorff Distance (HD95).

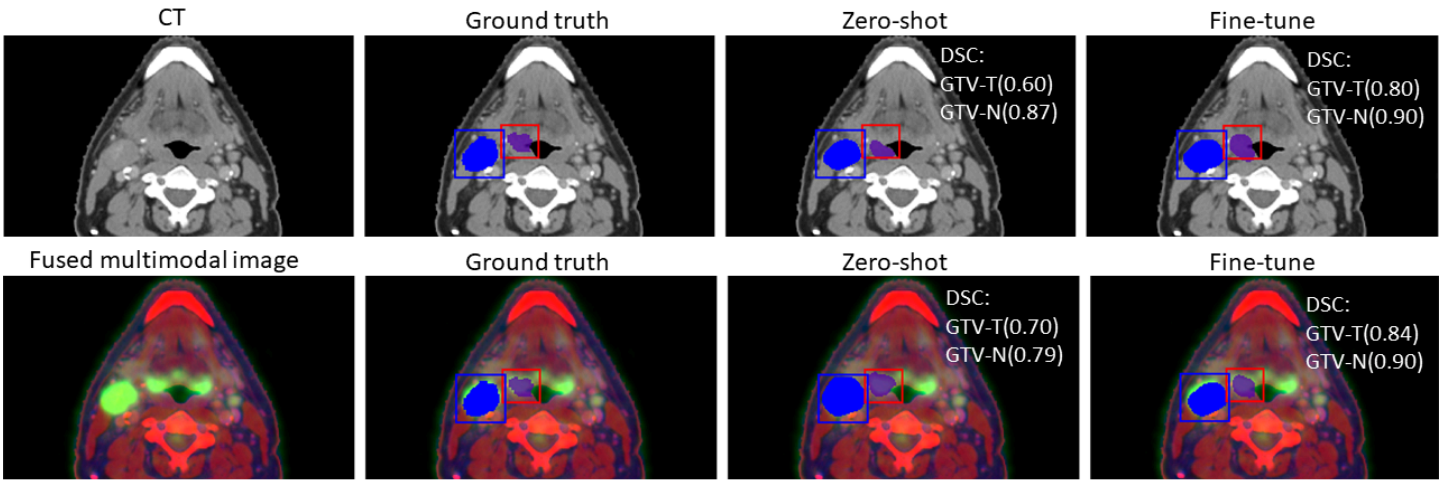
Figure 2: A case study illustration of CT-only and multimodal approaches on zero-shot and fine-tuned MedSAM for GTV-T and GTV-N segmentation.

The fine-tuning time for a lite MedSAM model using the training set (470 patients) averaged 11 minutes per epoch. We obtained the optimal fine-tuned models at epoch 8 for CT-only inputs and at epoch 4 for multimodal inputs. The prediction time for a target, given a bounding box on a 2D slice, was approximately 0.03 seconds, with no discernible difference between CT-only and multimodal inputs.

4 Discussion

In this investigation, we systematically explored the efficacy of the SAM model in segmenting GTV-T and GTV-N using local single- and multi-modality images. Even though the ground truth delineation referenced multimodality images, the CT-only input yielded comparable results for zero-shot inference with MedSAM on GTV-T. Interestingly, it achieved better results on GTV-N. The finetuning shows the effectiveness of enhancing segmentation accuracy and consistency for both groups.

Previous studies have demonstrated that fully automatic U-Net configurations, which combine multiple modalities (e.g., CT and PET), significantly outperform those using a single modality[15]. In our study, the CT-only approach with MedSAM achieved better segmentation accuracy compared to the fully automatic multi-modality nnUNet. This suggests that the inclusion of a bounding box for prompting effectively localizes the region of interest, even in zero-shot inference scenarios, as evidenced by the significant reduction in outliers highlighted by the HD95 metric. This benefit is also apparent in nnUNet's performance when utilizing bounding boxes. Conversely, the similar overall performance between CT-only and multimodal approaches in zero-shot inference could imply that the method used for fusing multimodal images may not be fully effective by SAM. This aligns with prior research emphasizing PET's crucial role in directing the network's focus to the relevant area[4,12]. Additionally, there are instances, where multimodal could enhance GTV-T boundary contrast, as depicted in Figure 2.



The significant improvement in performance following fine-tuning demonstrates the strong adaptation capability of SAM to new datasets within just a few epochs of training (8 for CT-only and 4 for multimodal). This is particularly notable in our proposed multimodal fusion strategy, considering that neither SAM nor MedSAM had previously encountered such image types. Despite their robust generalization capabilities, a gap remains between the fine-tuned models and the nnUNet when using the same bounding boxes. This gap may be attributed to nnUNet's powerful feature learning capability, stemming from its large model size.

These findings also spotlight a key limitation of fully automatic segmentation methods: their diminished focus when predicting GTV with full 3D image size. Interactions with models like nnUNet, while robust, can be more complex and time-intensive compared to simpler approaches like SAM. Predicting a comprehensive 3D head and neck image incorporating CT, PET, T1w, T2w, and an additional channel for human inputs can take up to 2 minutes when using nnUNet. In contrast, using a bounding box to prompt SAM for a single slice yielded instant responses (~0.03 seconds). Therefore, integrating SAM with U-Nets as a semi-automated annotation tool holds great promise for incorporating 'human-in-the-loop' practices in clinical settings. Such tools could not only increase the credibility and interpretability of AI models but also enable personalized treatment planning for optimized radiotherapy[4].

Our study has some limitations. Firstly, we contrasted the 2D MedSAM model with the 3D nnUNet, where the latter might benefit from contextual dependencies along the axial direction. However, given that SAM is designed for interactive use and nnUNet for fully automatic use, a direct comparison may not constitute a fair contest. The nnUNet, when equipped with a bounding box for each slice, sets a high benchmark for semi-automatic segmentation, whereas its performance without a bounding box represents a decent threshold for fully automatic segmentation. Secondly, while we selected the best-performing fine-tuned model based on its performance on the validation set to minimize the risk of overfitting, there remains a potential for catastrophic forgetting when applied to other datasets. Validating the model on a public HNC dataset could help mitigate this concern. Moreover, future studies exploring the integration of various prompting methods, including clicks, brushes (masks), or textual (language) inputs, in addition to bounding boxes, would be valuable.

5 Conclusion

In conclusion, our study demonstrated the potential of zero-shot and fine-tuning SAM for the segmentation of GTV in head and neck cancer. We found that fine-tuning significantly enhances segmentation accuracy for both CT-only and fused multimodal images, underscoring the model's robust adaptation capabilities. Our findings suggest promising avenues for enhancing semi-automatic segmentation tools in clinical settings using SAM.

References

[1] X. Luo et al., “Deep learning-based accurate delineation of primary gross tumor volume of nasopharyngeal carcinoma on heterogeneous magnetic resonance imaging: A large-scale and multi-center study,” Radiother. Oncol., vol. 180, p. 109480, Mar. 2023, doi: 10.1016/j.radonc.2023.109480.

[2] Y. Wang et al., “Comparison of deep learning networks for fully automated head and neck tumor delineation on multi-centric PET/CT images,” Radiat. Oncol., vol. 19, no. 1, p. 3, Jan. 2024, doi: 10.1186/s13014-023-02388-0.

[3] R. R. Outeiral et al., “Strategies for tackling the class imbalance problem of oropharyngeal primary tumor segmentation on magnetic resonance imaging,” Phys. Imaging Radiat. Oncol., vol. 23, pp. 144–149, Jul. 2022, doi: 10.1016/j.phro.2022.08.005.

[4] Z. Wei, J. Ren, S. S. Korreman, and J. Nijkamp, “Towards interactive deep-learning for tumour segmentation in head and neck cancer radiotherapy,” Phys. Imaging Radiat. Oncol., vol. 25, p. 100408, Jan. 2023, doi: 10.1016/j.phro.2022.12.005.

[5] M. E. Rasmussen, J. A. Nijkamp, J. G. Eriksen, and S. S. Korreman, “A simple single-cycle interactive strategy to improve deep learning-based segmentation of organs-at-risk in head-and-neck cancer,” Phys. Imaging Radiat. Oncol., vol. 26, p. 100426, Apr. 2023, doi: 10.1016/j.phro.2023.100426.

[6] A. G. Smith, D. Kutnár, I. R. Vogelius, S. Darkner, and J. Petersen, “Localise to segment: crop to improve organ at risk segmentation accuracy.” arXiv, Apr. 10, 2023. doi: 10.48550/arXiv.2304.04606.

[7] A. Kirillov et al., “Segment Anything.” arXiv, Apr. 05, 2023. doi: 10.48550/arXiv.2304.02643.

[8] M. A. Mazurowski, H. Dong, H. Gu, J. Yang, N. Konz, and Y. Zhang, “Segment anything model for medical image analysis: An experimental study,” Med. Image Anal., vol. 89, p. 102918, Oct. 2023, doi: 10.1016/j.media.2023.102918.

[9] J. Cheng et al., “SAM-Med2D.” arXiv, Aug. 30, 2023. doi: 10.48550/arXiv.2308.16184.

[10] J. Wu et al., “Medical SAM Adapter: Adapting Segment Anything Model for Medical Image Segmentation.” arXiv, Dec. 28, 2023. doi: 10.48550/arXiv.2304.12620.

[11] J. Ma, Y. He, F. Li, L. Han, C. You, and B. Wang, “Segment anything in medical images,” Nat. Commun., vol. 15, no. 1, Art. no. 1, Jan. 2024, doi: 10.1038/s41467-024-44824-z.

[12] J. Ren, B.-N. Huynh, A. R. Groendahl, O. Tomic, C. M. Futsaether, and S. S. Korreman, “PET Normalizations to Improve Deep Learning Auto-Segmentation of Head and Neck Tumors in 3D PET/CT,” in Head and Neck Tumor Segmentation and Outcome Prediction, V. Andrearczyk, V. Oreiller, M. Hatt, and A. Depeursinge, Eds., in Lecture Notes in Computer Science. Cham: Springer International Publishing, 2022, pp. 83–91. doi: 10.1007/978-3-030-98253-9_7.

[13] K. Wu et al., “TinyViT: Fast Pretraining Distillation for Small Vision Transformers.” arXiv, Jul. 21, 2022. doi: 10.48550/arXiv.2207.10666.

[14] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, “nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation,” Nat. Methods, vol. 18, no. 2, Art. no. 2, Feb. 2021, doi: 10.1038/s41592-020-01008-z.

[15] A. R. Groendahl et al., “A comparison of methods for fully automatic segmentation of tumors and involved nodes in PET/CT of head and neck cancers,” Phys. Med. Biol., vol. 66, no. 6, p. 065012, Mar. 2021, doi: 10.1088/1361-6560/abe553.

Test-time Adaptation for Medical Image Segmentation based on a Joint framework with an Adaptive Wavelet-VNet and a Refine Model (WaVNet-Refine)

Xiaoxue Qian[1], You Zhang[1]

[1]Department of Radiation Oncology, UT Southwestern Medical Center, Dallas, TX, 75390

Abstract For medical image segmentation, there often exists a domain gap between acquired training and testing datasets due to variations in imaging hardware/protocols. It leads to domain shift problems that may substantially degrade the performance of deep learning-based segmentation models. Due to the high cost of manual labeling and the need for privacy protection, it is often difficult to annotate new (target) domain data for model fine-tuning or re-access source domain training data to train domain generalization models. Therefore, test time adaptation using only target domain data is a more practical but challenging solution. In this study, we proposed an adaptive wavelet-VNet for test-time image segmentation adaptation. To improve the model's robustness to domain variations, we introduced the multi-level wavelet transform into an adaptive wavelet-VNet, where multiscale wavelet coefficients are concatenated to the VNet encoder to effectively extract domain-invariant features. Meanwhile, we employed a Refine model to correct the incomplete and noisy segmentation results caused by the domain shift. The adaptive wavelet-VNet and the Refine model were unified into a test-time adaptation framework (WaVNet-Refine). We evaluated WaVNet-Refine with a multi-domain liver dataset and showed its advantages over other state-of-the-art test-time adaptation methods.

1 Introduction

Deep neural networks have greatly improved the accuracy of medical image segmentation [1], while their performance heavily relies on the underlying data distribution. In real practice, medical images often contain domain shifts due to different imaging hardware or scanning protocols. When a model trained on a source-domain dataset was tested on another from unseen target domains, its performance typically declines due to the domain shifts. To tackle this problem, a few fine-tuning and domain adaptation methods were proposed [2, 3]. However, these methods usually require labeled data in the target domain, or need to re-access the source domain data for model generalization. The labeling costs and the data privacy concerns often render them in-practical. Test-time adaptation, which only uses the label-free, target-domain data during test time for model updates, presents a much more viable alternative. However, the label-free test-time adaptation is challenging to implement, as the unknown domain gap may render the model adaptation unreliable and inconsistent.

Thus, extracting effective, domain-invariant features is crucial for the test-time adaptation models to achieve robust segmentation results. Wavelets transforms [4] are known to be robust against variations in image intensity and quality, making them highly reliable for feature extraction in various medical imaging modalities and conditions. In this regard, we introduced the wavelet transform into the test-time adaptation model, which uses a series of filters to enhance domain-invariant information while suppressing domain-specific information. Specifically, the wavelet transform converts the images into low and high-frequency wavelet coefficients. The low-frequency coefficients preserve the general appearance of anatomies, while effectively reducing variations in their intensity and contrast. The high-frequency sub-bands, on the other hand, enhance anatomical and edge details. We embedded these wavelet coefficients into a VNet backbone to enhance the network's capability in learning domain-robust features for segmentation, denoted as WaVNet.

To allow WaVNet to adapt at the test time to further address domain differences, we introduced an adaptive module, which is a shallow convolutional module deployed on the front of WaVNet, to update the low level features to accommodate the testing sample. Since there is no label from the testing sample for supervised learning, we introduced a Refine model (Fig. 1), a segmentation correction model, to generate an unsupervised loss on WaVNet's output to update the shallow adaptive module. The Refine model is designed as a reconstruction network using inadequately-trained source-domain models, based on the observation that incorrect segmentations on unseen target domains may resemble the segmentations from inadequately trained source-domain models. During test-time adaptation, the input and output of Refine model are used to calculate an unsupervised loss to update the parameters of the adaptive module. Overall, we proposed a joint framework with adaptive WaVNet and a Refine model (WaVNet-Refine) for online test-time adaptation. We evaluated WaVNet-Refine on a multi-domain liver MR imaging set for liver segmentation, and compared its accuracy against several other state-of-the-art methods.

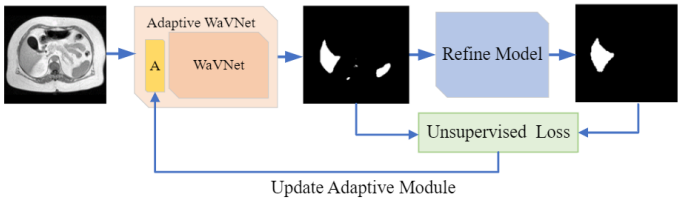


Figure 1: Overview of the WaVNet-Refine framework for test-time adaptation. A: the shallow low-level adaptive module to be updated during the test-time adaptation.

2 Materials and Methods

The overall WaVNet-Refine framework is shown in Fig.1, which comprises the adaptive WaVNet model and the subsequent Refine model. Details of the adaptive WaVNet model and the Refine model are introduced in the following.

2.1 The Adaptive WaVNet

2.1.1 Wavelet transform

Wavelet transform, a multi-resolution geometrical analysis method, uses filters to statistically analyze spectral features across various scales and orientations. Fig. 2 illustrates the use of multi-level wavelet transform to extract different scale features from a 3D image. Given a 3D input image $X \in R^{N\times N\times C}$ (N = 256 and C = 64 in this study) and a wavelet decomposition level of n = 3, the first level discrete wavelet transform (DWT) generates one low-frequency sub-band and seven high-frequency sub-bands for seven orientations, which are concatenated to compose the scale-1 features. Subsequently, performing the second-level DWT on the low-frequency sub-band from scale-1 further generates scale-2 features, and the third-level DWT generates the scale-3 features from the low-frequency sub-band of scale-2.

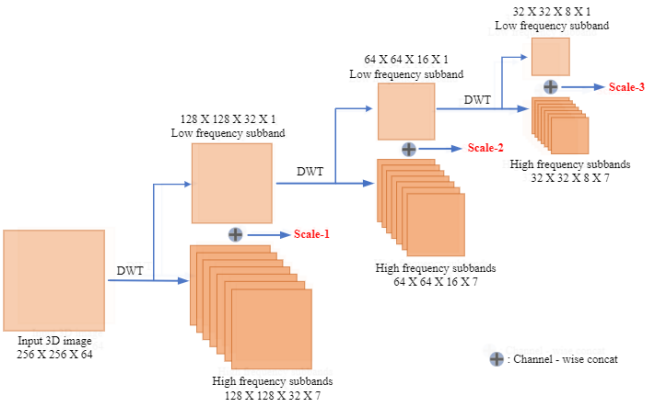


Figure 2: Multi-level wavelet transform on a 3D image.

The extracted features at different levels (scale-1, scale-2, scale-3) were integrated within the VNet architecture to promote the learning of domain-invariant features. From the perspective of signal filtering, wavelet transform and convolutional neural networks such as VNet share similar mathematical representations. The multi-level wavelet transform can be expressed as follows [5]:

$$F_{L,n+1} = (F_{L,n} * k_{L,n}) \downarrow 2 \quad (1)$$

$$F_{H,n+1} = (F_{L,n} * k_{H,n}) \downarrow 2 \quad (2)$$

where $k_{L,n}$ represents the scaling function, acting as a low-pass filter in the frequency domain; and $k_{H,n}$ denotes the wavelet function, acting as a high-pass filter. n denotes the wavelet level, and $F_{L,0}$ equals the input 3D image X .

On the other hand, a typical VNet [6] comprises multiple convolutional, down-sampling, and up-sampling layers. The convolution and down-sampling can be represented mathematically as follows:

$$F_{l+1} = (F_l * K) \downarrow p \quad (3)$$

where K denotes the convolutional filters, symbol $\downarrow$ represents the down-sampling operator, and p denotes the stride of down-sampling. When p<1, the down-sampling turns into up-sampling. l denotes the network layer, and F_0 equals X, the input 3D image .

From Eqs. 1-3, it can be observed that the wavelet transform and VNet share similar filtering operations on images, but from the perspectives of frequency (spectral) and spatial domains, respectively. Compared with the convolutional filtering of the conventional VNet, the wavelet transform extracts additional domain-invariant information through the spectral domain filtering, which can be combined with the VNet to train models that are more robust to domain variations. Correspondingly, we concatenated the three scale features extracted by the wavelet transform into the 2nd, 3rd, and 4th layers of the VNet encoder to fuse their information to generate the WaVNet, which learns hierarchical spectral and spatial features together to enhance the domain invariance of the model.

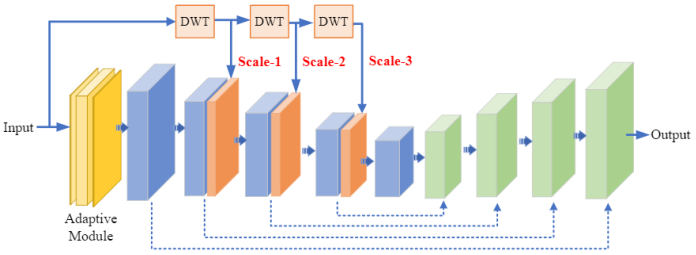


Figure 3: Details of the adaptive WaVNet, where the three-scale features extracted by the wavelet transform are concatenated with the VNet encoder to promote hierarchical spectral and spatial feature learning.

2.1.2. The adaptive module

Although incorporating the wavelet transform into VNet promotes domain-invariant learning, it is insufficient to address all domain gaps. It has been found that domain differences in medical images mainly appear in low-level features while high-level structural characteristics usually stay consistent across domains. Thus the low-level feature layers of the WaVNet can be further adapted online to address the domain gaps observed in the testing sample. Correspondingly, we developed an adaptive module for test-time adaptation. It is a shallow convolutional unit of three layers positioned at the front of the WaVNet, rendering an adaptive WaVNet. The detailed network structure is shown in Fig. 3. During test-time adaptation, we only updated the parameters of the adaptive module with an unsupervised loss generated by a Refine model, while keeping all the other network parameter frozen.

2.2 The Refine model

To update the adaptive module of WaVNet online, we need to generate a loss function for backpropagation. Due to the lack of 'ground-truth' labels from the testing samples, supervised training cannot be performed. In light of this, we developed a Refine model, serving as a segmentation corrector/evaluator, to evaluate the output segmentations from WaVNet and generate an unsupervised loss to update the adaptive module.

We used under-trained source models to generate incomplete and noisy segmentation data to train a Refine model, which helps to evaluate/fix incorrect segmentations. It is motivated by the observation that when source domain models were tested on target domain samples with distribution shifts, the segmentation results often showed

poor regional consistency, inaccurate edges, misclassifications, and noise, similar to those obtained from inadequately-trained source models. To generate under-trained source models, we used early stopping to derive a variation of WaVNet models, which were used to generate incomplete and noisy segmentation results from the source domain images. Using these incomplete/noisy segmentations as inputs, we trained the Refine model using their corresponding 'ground-truth' labels. We instantiated the Refine model as a reconstruction network, using an auto-encoder. The mean squared error (MSE) loss function was used:

$$L_{ref} = L_{MSE}(R_c(y'), y); y' = V_s(X) \quad (4)$$

where X and y denote the source images and labels, respectively. V_S represents the under-trained source models, and R_C denotes the Refine model. During training, by encoding and decoding the inputs, the Refine model gradually learned how to repair incomplete segmentations to match the labels.

3 Evaluation and Results

3.1 Datasets

We used the CHAOS liver segmentation dataset [7] to evaluate WaVNet-Refine. The CHAOS dataset includes two subsets of different MR sequences: T1 DUAL (40 cases) and T2 SPIR (20 cases). The T1 DUAL images are further divided into 20 in-phase (IP) and 20 out-phase (OP) images, which have the same repetition times but different gradient echo times and can be considered as different domains. The three-domain data were arranged into six comparison groups (IP (source) to OP (target), OP to IP, IP to T2, OP to T2, T2 to IP, and T2 to OP), each with 20 source samples and 20 target samples. For source domains, we used 15 randomly-selected samples for training and 5 for validation. For target domains, we randomly-selected 15 samples for supervised target domain model training (for comparison with our method) and 5 for adaptation and testing. All images were uniformly resampled to (1.5, 1.5, 1.5) mm³, cropped to (256, 256, 64), and normalized to [0, 1].

3.2 Implementation details and comparison methods

We used the original VNet [6] as the baseline model and applied a three-level DWT with the db1 wavelet [4] to construct the WaVNet. The channel number of the adaptive module was set to (1, 64, 1). The Refine model was implemented as a four-layer VNet with a sigmoid activation function at the end. The adaptive WaVNet was trained for 2000 epochs as the source model. We also trained five additional WaVNet models for (20, 40, 60, 80, 100) epochs to generate five under-trained models, and used them to output 100 incomplete segmentations (20 for each) as the training samples of the Refine model. The Refine model was trained for 1000 epochs. For online test-time adaptation, we fixed all WaVNet-Refine parameters except for those of the shallow adaptive module, and separately updated the network for two iterations using each single test sample. All the trainings are based on a batch size of 1, with the SGD optimizer and a learning rate of 0.0005.

We compared seven methods to evaluate the efficacy of our method: **1. No adaptation (NA)**: we trained a VNet model using the source domain's training samples for 2000 epochs, and tested on the 5 testing sets of the target domain. **2. From scratch (FS)**: we trained a VNet model using the target domain's 15 training cases for 2000 epochs, and tested on the 5 testing sets. **3. Test time adaptation: tent (TN)** [8]: We trained a VNet model on the source domain for 2000 epochs, performed test-time adaptation on the target domain's 5 testing sets using the entropy minimization loss for 10 epochs. **4. Test time adaptation: regional nuclear norm + contour regularization (RN)** [9]: Similar to TN, but used regional nuclear-norm and contour regularization losses for test-time adaptation. **5. WaVNet (WV):** Similar as WaVNet-Refine, without the Refine model or test-time adaptation. **6. VNet-Refine (VR)**: Similar to WaVNet-Refine, but without the wavelet transform. **7. WaVNet-Refine (WR)**: the framework proposed in this study. We use two metrics: the Dice coefficient (Dice) and the 95th percentile Hausdorff distance (HD95) to evaluate the segmentation results.

3.3 Segmentation results comparison

The Dice and HD95 results for liver segmentation, obtained using seven different methods across three domains' datasets, are detailed in Tables 1 and 2.

Method	*IP to OP*	*OP to IP*	*IP to T2*	*OP to T2*	*T2 to IP*	*T2 to OP*	*Mean*
NA	79.77	65.33	49.78	62.46	65.12	63.57	64.34
FS	88.40	90.19	90.23	90.23	90.19	88.40	89.61
TN	84.90	76.61	60.91	64.87	77.51	64.10	71.48
RN	84.49	77.94	65.45	71.54	74.89	68.93	73.87
WV	82.96	70.89	67.93	69.04	69.72	66.82	71.23
VR	86.56	82.45	65.50	73.68	75.43	70.32	75.65
WR	87.17	83.90	78.41	79.69	79.96	74.64	80.63

Table 1: Average Dice (%) of the seven methods on the liver dataset, each result of different methods/testing groups was averaged over 5 testing/adaptation cases.

Method	*IP to OP*	*OP to IP*	*IP to T2*	*OP to T2*	*T2 to IP*	*T2 to OP*	*Mean*
NA	26.33	65.11	80.31	78.35	67.93	45.42	60.58
FS	23.07	11.31	6.884	5.472	11.31	23.07	13.52
TN	24.98	49.39	78.58	62.95	38.25	37.41	48.59
RN	24.06	33.25	63.41	42.91	35.58	37.55	39.46
WV	13.26	35.84	24.88	20.65	29.35	28.02	25.33
VR	11.52	12.60	32.79	18.37	25.06	28.44	21.46
WR	9.01	11.36	25.01	18.00	19.82	15.89	16.51

Table 2: Average HD95 of the seven methods on the liver dataset, each result of different methods/testing groups was averaged over 5 testing/adaptation cases.

Serving as the no-adaptation benchmark, NA in Table 1 yielded the poorest results among all, showing the impacts

of domain shifts. With direct access to the target domain's images and labels for supervised training, the FS method achieved the best results. However, obtaining labeled data from the target domain can be challenging, especially when the domain shift is not made aware, or the domain shift happens frequently along different directions. Using the wavelet-incorporated VNet, WV offered better segmentation results than NA, while still suffering from the domain shifts and the lack of test-time adaptation. The four methods (TN, RN, VR, and WR) all outperformed the NA method, showing the benefits of test-time adaptation. However, compared with VR, TN and RN were based on statistical loss functions and did not use the Refine model. The Refine model helped VR to repair and evaluate the segmentation results using prior information for more effective test-time adaptation and achieve more accurate segmentation. Comparing VR with WR, the wavelet transform helped WR to capture more domain-robust features and made it the overall most accurate test-time adaptation methods.

Compared to the DICE coefficient shown in Table 1, the HD95 metric in Table 2 is more sensitive to detailed contour edge inaccuracy. The existing test-time domain adaptation methods, TN and RN, showed much inferior results compared with VR and WR for HD95. WR achieved a similar mean HD95 as the supervised FS method.

The qualitative segmentation results were shown in Fig. 4. The subpanels (a) and (b) showed the segmentations on two different slices using various methods, respectively. The WaVNet-Refine segmentation results are closer to the two target-domain supervised learning methods and the 'ground-truth', showing the effectiveness of combining domain-robust feature learning with Refine model-guided test-time domain adaptation.

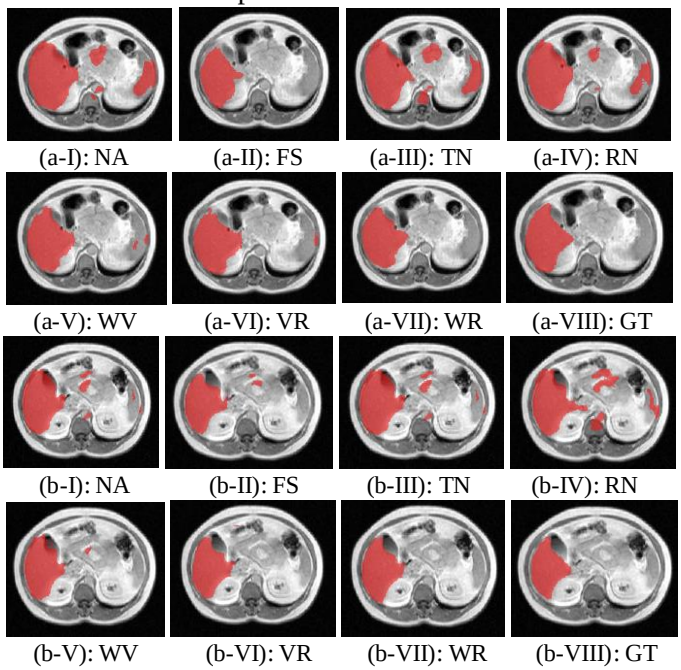


Figure 4: Visual comparison of segmentation results for two different slices (a, b), using the seven methods. (I) no adaptation (NA), (II) from scratch (FS), (III) tent (TN), (IV) regional nuclear norm + contour regularization (RN), (V) WaVNet (WV), (VI) VNet-Refine (VR), (VII) WaVNet-Refine (WR), and (VIII) Ground truth (GT).

4 Conclusion

We proposed a novel online test-time adaptation method, WaVNet-Refine, to address the challenges of domain shifts in medical image segmentation. WaVNet incorporates wavelet transforms to enhance domain-robust feature learning for segmentation, and the Refine model can further rectify and evaluate the segmentations from WaVNet to generate an unsupervised loss, which is used to fine-tune WaVNet during test time via a low-level feature adaptation module. WaVNet-Refine achieves substantial accuracy enhancement as compared with the other test-time adaptation methods, allowing more robust cross-domain segmentation. The test-time adaptation potential of WaVNet-Refine could benefit from additional domain-invariant learning, as well as other segmentation fine-tuning tools including atlas-based and prior contour-based learning. Additional evaluations on more anatomical sites and datasets are warranted.

References

[1] J. M. J. Valanarasu, V. A. Sindagi, I. Hacihaliloglu, et al. "Kiu-net: Overcomplete convolutional architectures for biomedical image and volumetric segmentation". *IEEE Transactions on Medical Imaging* 41.4 (2021), pp. 965–976. DOI: 10.1109/TMI.2021.3130469.

[2] J. Wang, C. Lan, C. Liu, et al. "Generalizing to unseen domains: A survey on domain generalization". *IEEE Transactions on Knowl- edge and Data Engineering* (2022), pp. 8052–8072. DOI: 10.1109/TKDE.2022.3178128.

[3] Q. Liu, Q. Dou, L. Yu, et al. "MS-Net: multi-site network for improving prostate segmentation with heterogeneous MRI data". *IEEE transactions on medical imaging* 39.9 (2020), pp. 2713–2724. DOI: 10.1109/TMI.2020.2974574.

[4] S. G. Mallat. "A theory for multiresolution signal decomposition: the wavelet representation". *IEEE transactions on pattern analysis and machine intelligence* 11.7 (1989), pp. 674–693. DOI: 10.1109/34.192463.

[5] S. Fujieda, K. Takayama, and T. Hachisuka. "Wavelet convolutional neural networks". *arXiv preprint arXiv:1805.08620* (2018). DOI: 10.48550/arXiv.1805.08620.

[6] F. Milletari, N. Navab, and S.-A. Ahmadi. "V-net: Fully convolutional neural networks for volumetric medical image segmentation". *2016 fourth international conference on 3D vision (3DV)*. Ieee. 2016, pp. 565–571. DOI: 10.1109/3DV.2016.79.

[7] A. E. Kavur, N. S. Gezer, M. Barış¸, et al. "CHAOS challenge-combined (CT-MR) healthy abdominal organ segmentation". *Medi- cal Image Analysis* 69 (2021), p. 101950. DOI: 10.1016/j.media.2020.101950.

[8] D. Wang, E. Shelhamer, S. Liu, et al. "Tent: Fully test-time adaptation by entropy minimization". *arXiv preprint arXiv:2006.10726* (2020). DOI: 10.48550/arXiv.2006.10726.

[9] M. Hu, T. Song, Y. Gu, et al. "Fully test-time adaptation for image segmentation". *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24*. 2021, pp. 251–260. DOI: 10.1007/978-3-030-87199-4_24.

MAGIC: Multitask Adversarial Generator of Images and Contours from CBCT for Adaptive Radiotherapy

Adam Szmul[1,2,3], Kamalram Thippu Jayaprakash[4], Rajesh Jena[4], Andrew Hoole[5], Catarina Veiga[2,3], Yipeng Hu[1,2,3], and Jamie R. McClelland[1,2,3]

[1]Wellcome / EPSRC Centre for Interventional and Surgical Sciences (WEISS), University College London, London, UK
[2]Centre for Medical Image Computing (CMIC), University College London, London, UK
[3]Department of Medical Physics and Biomedical Engineering, University College London, London UK
[4]Department of Oncology, Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK
[5]Department of Medical Physics, Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK

Abstract Adaptive radiotherapy requires images of planning CT-like quality and segmentations of structures for treatment plan evaluation and its adaptation. We propose an unsupervised framework which can jointly generate synthetic CTs and segment Organs at Risk, without requiring any segmentations on CBCT during training or inference. The presented results show feasibility of such an approach with improvement in CBCT images in terms of L2 histograms distance, mean absolute error and normalized cross correlation with respect to virtual CTs. The proposed framework generates segmentations that require further editing to be clinically acceptable but shows great promise considering the relatively small training dataset used. This proof-of-concept study has potentials for clinical uses, such as being an assistive tool in current manual segmentation, and future development as plan verification tool or a fully automated replanning.

1 Introduction

Adaptive radiotherapy requires images with comparable quality to planning Computed Tomography (pCT) for accurate treatment plan verification and its adaptation [1]. Daily acquired Cone Beam CT (CBCT) are of too low quality to allow for this, as they commonly contain reconstruction artefacts and have poor Hounsfield Units [HU] preservation [2]. Furthermore, organs at risk (OAR) and target volumes would need to be delineated from CBCTs for quantitatively verifying and adapting treatment plans. Manual delineation of those structures is very laborious and is routinely done only at the treatment planning stage on a pCT scan. Currently, when a patient's treatment plan is no longer valid, a new pCT needs to be acquired and segmented to prepare a new treatment plan. That process is not only time and resource consuming but may also introduce delays in a patient's treatment delivery [3]. A more efficient approach would be to convert CBCTs to synthetic CTs (synCT) and automatically delineate the structures on CBCT, allowing for offline or online adaption of the treatment plan.

Machine learning methods have already been shown to be capable of providing high quality synCTs, which can be used for dose calculation and structure segmentation [4]. Delineating structures from CBCTs remains challenging. While routinely delineated structures on pCT scans can be employed to train CT-dedicated segmentation methods, the absence of comparable delineations for CBCTs poses a challenge in the development of supervised machine learning methods.

In this work we propose a framework for joint synthetic CT (synCT) generation and structure segmentation from daily CBCT, which learns to perform both tasks simultaneously. The framework does not require the pCTs and CBCTs to be paired, and only requires segmentations on the pCTs and does not require segmentations on the CBCTs.

2 Materials and Methods

Retrospective data from Cambridge University Hospitals NHS Foundation Trust were used in our study. The Addenbrookes Secure Therapeutic Radiology Image Dataset (ASTRiD), IRAS Number 312143 database protocol provided the overarching ethics approval for the pCT and/or CBCT images. The CBCTs suitable for use in this study require full body field of view (FOV), allowing for potential use in dose calculation and treatment adaptation. We identified 38 patients whose CBCT scans could be included in this work, 29 of these were used for training and 9 for testing. On average there were 20 CBCT scans per patient. The pCTs from 144 patients were used, along with their corresponding structures. That dataset consisted of patients for which CBCTs were included in our experiment, and those where CBCTs were not suitable for the study. We included 4 structures in this study: bladder, rectum, left and right femur bones. To ensure consistency in the superior-inferior anatomical coverage between CBCTs and pCTs, we sampled slices from pCTs starting from the bottom of the rectum (3 slices below the rectum) up to above the bladder (3 slices above the bladder). This superior-inferior anatomical region corresponded to the FOV coverage of the CBCTs. The CBCT intensities were linearly scaled to best match the global intensity profiles of the pCTs.

We based our approach on 2D cycle-consistent Generative Adversarial Networks (cycleGAN) [5]. We adapted our previous framework that can generate high quality synthetic CTs from clinical CBCTs utilizing modified cycleGANs [6]. We introduced a two head generator, which consisted of a shared encoder and partially shared decoder. The decoder splits into separate image and segmentation decoders. The image and segmentation paths in the decoder shared 3 upsampling stages and then the rest of the decoder's layers (5) were image or segmentation specific.

We refer to it as mid split of the decoder. The proposed framework uses global residual learning, where the image generators predict only corrections to be applied to the input images [5]. Additional structure consistency was employed between the input and output images of the generators to enhance anatomical correspondence between the input images and their synthetic versions.

As the segmentations are only available for pCTs, the CT discriminator was adapted to evaluate not only images but to take image and corresponding segmentations as an input, creating a multichannel tensor. We also included a cycle consistency segmentation loss in the form of weighted Dice and Cross Entropy between the original pCT segmentations and the segmentations generated for synthetic CBCTs (when real CT were converted to synthetic CBCTs), A schematic diagram showing the proposed framework is illustrated in Fig. 1. The proposed framework requires unpaired CBCTs and CTs (with segmentations for CTs only) at the training stage, and just CBCTs at the inference stage.

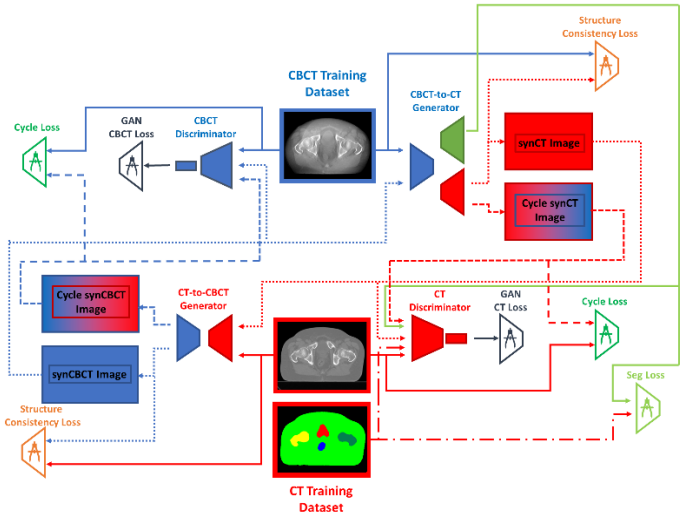


Fig.1 A schematic diagram of the framework.

The framework was first trained for 200 epochs with the weight for the cycle-segmentation loss set to 0, essentially training only image generation. For CT discriminator, which takes a multichannel tensor as an input, we propagated synCT to all channels. We used Adam optimizer with an initial learning rate of 0.0002 for the first 100 epochs, which was then linearly reduced to 0 over the following 100 epochs. Subsequently, the encoder and image decoder for CT generator were frozen to allow for segmentation decoder and CT discriminator to be trained for another 200 epochs. For that step the input of the CT discriminator was a multichannel tensor of synCT and segmentations. The same optimization and learning rate scheduler were used.

The output segmentations were post-processed with morphological operations, including hole filling and closing operations. The left and right femur bones were separated based on their spatial location.

To evaluate the synCTs, we created virtual CTs (vCTs) for the testing dataset by registering the pCTs to the CBCTs using NiftyReg [7] with locally normalized cross correlation as the similarity measure. To ensure a meaningful evaluation in terms of Hounsfield Units [HU], we performed intensity correction of the each CBCT image to a corresponding vCT, adjusting their mean value and standard deviation to match those of vCT. This operation was performed individually for each CBCT scan, in contrast to intensity correction performed globally during training. No intensity correction was applied to synCTs. The image evaluation was conducted based on mean absolute error (MAE) and normalized cross correlation (NCC). We also compared histograms of the images calculating the L2 norm between them.

Although the CBCTs used in this study had a large enough FOV to include all of the patients' anatomy, we discovered that the CBCTs contained reconstruction artefacts near the edge of the patient (see Fig. 3). This meant the CBCT scans did not provide reliable information on the anatomy close to the surface of the patient. Interestingly, the synCTs seemed to better resemble the pCTs in these regions and did not reproduce the inaccurate patient surface from the CBCTs. So that our evaluation was not negatively impacted by this, we automatically generated evaluations masks from the CBCTs using thresholding followed by morphological operations, and only performed the evaluation within these masks.

An experienced radiation oncologist (KTJ, over 10 years of experience) rated segmentations generated by our framework on its clinical usability on a scale of: clinically acceptable, requiring minor corrections, requiring major corrections, better to start from scratch. Femur bone segmentations, although they were split into left and right, were given a single score.

3 Results

The proposed framework generated synCTs that showed improved agreement with the vCTs compared to the CBCTs for all the metrics used. The histograms of all testing CBCT, vCT and synCT are shown in Fig.2, overlayed with plots approximating their shapes.

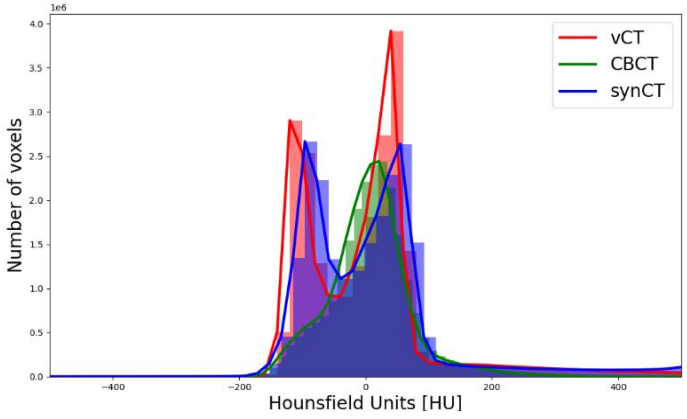


Fig.2 Histograms showing intensity distributions.

It can be noticed that the CBCT histogram has a distinctly different shape than the vCT histogram, with the synCT histogram more closely resembling the vCT histogram. The L2 norm between the vCT and the CBCT histograms was 2868 and was just 678 between the vCT and synCT.
The MAE between the vCT and CBCT (after intensity normalization) was 69.2 (std of 9.9) HU, and this was reduced to 50.7 (std of 7.9) HU between the vCT and synCT. The NCC improved from 0.77 (std of 0.04) for CBCT to 0.84 (std of 0.07) for synCT.

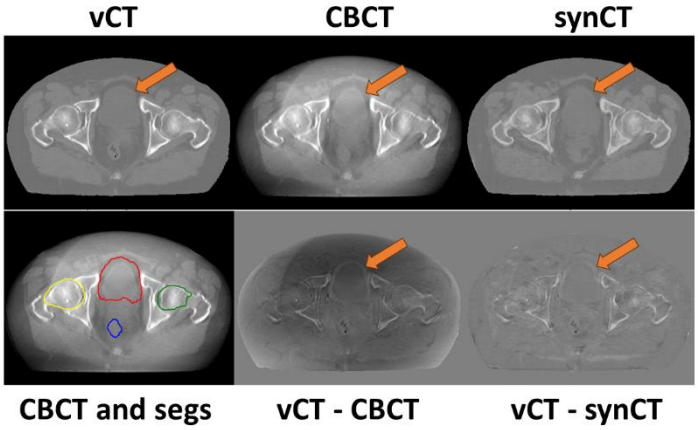


Fig.3 A comparison between the vCT, CBCT and synCT.

A visual comparison of the vCT, CBCT and synCT is shown in Fig. 3. The difference image between the vCT and CBCT, and the vCT and synCT, are also shown in the figure. It can be noticed that while the vCT shows generally good anatomical agreement with the CBCT, in some areas, such as the bladder, (as shown in Fig.3) there is a noticeable misalignment of the anatomy. That is highlighted with orange arrows in the figure. An example of the segmentations generated by our method overlayed over CBCT are also shown.

Tab.2 Qualitative evaluation of the segmentations.

	Acceptable	Minor edits	Major edits	Start from scratch
Bladder	0	1	3	5
Rectum	0	5	3	1
Femur	0	4	4	1

The qualitative evaluation is summarized in Tab. 2. The bladder was the worst performing structure, with over half of the cases (5) being scored as better to start from scratch. However, the segmented bladder volume was always roughly in the correct spatial position, usually being slightly over segmented. The rectum and femur bones both had scores almost equally distributed between requiring minor (rectum: 5, femur: 4) and major (rectum: 3, femur: 4) edits. For both structures only one case was scored as better to start from scratch.

4 Discussion

The presented results for image evaluation show improvement in the image quality from the CBCT level closer to the vCT. It is important to notice that the vCT is not a perfect representation of the ground truth image (i.e. a pCT paired with the CBCT). The registration of the pCT to the CBCT was capable of compensating for most of the differences, however it was not able to fully recover all anatomical differences, particularly for bladder and other highly deformable organs whose shape might change considerably day to day. At the same time, structures like the bladder are the most important, as their correct delineation is essential for treatment plan verification and adaptation. Therefore, some level of mismatch between the synCT and vCT is expected, and the vCT approach itself is not suitable for generating images for adaptive radiotherapy purposes.
One limitation of our study is that it performed only qualitative evaluation of the segmentations. In the future this could potentially be complemented by quantitative evaluation on a subset of cases. However, generating high quality manual segmentations on CBCT is time consuming and very challenging due to the poor image quality. Furthermore, the qualitative evaluation used in our study is closer to clinical practice, where segmentations would need to be reviewed and if necessary edited for clinical use.
The rectum and femur segmentations generated by our method already show some potential for clinical use, but we acknowledge that further improvements are required, especially to the bladder segmentations, for the method to be adopted clinically. However, we think the approach has great promise as it does not require manual segmentations on the CBCTs for training, so can potentially use far larger training datasets than were used in this study, which we would expect to lead to improved results.
One reason for the limited size of the training dataset in this study is that we discovered after collecting the data that the CBCT scans for many of the patients had a small FOV that excluded a significant amount of the patient's anatomy, and hence were not suitable for use. Even the CBCTs with a larger FOV that were used had reconstruction artefacts near the body surface due to the truncated data. Our method actually produced synCTs and body segmentations that plausibly resembled how we would expect the anatomy to appear in these regions (based on how it appears in the pCTs), but it is impossible to determine how accurate these were, hence we excluded this region from the image evaluation and did not evaluate the body segmentations. This highlights an important point for any methods that attempts to generate synCTs and auto-segmentations from CBCTs for plan verification and adaption: most CBCTs acquired clinically are not intended for this use, and if they are to be used for this in the future then it is important to acquire them with a large FOV that accurately reconstructs complete the patients' anatomy.

Training a GAN to jointly generate images and segmentations is a novel approach (as far as we are aware) – therefore there is no well-established training strategy. While a lot of research has been done training generative models, such as GANs and cycleGANs [8], or on how to best train a supervised segmentation model [9], a joint approach is yet to be investigated.

We observed that a shared decoder was more challenging to train, as one task usually dominated the training, leading to deterioration in the performance of the other. We investigated different splits between the image and segmentation decoders, going from completely separate decoders for each task, to a shared decoder that only splits in the layer just before the outputs were generated. When using independent decoders, we observed that the generated segmentations were not well aligned with the synCT. We found that a mid-level split in the decoder provided some level of information exchange between the two tasks and gave a well-balanced compromise which led to the best results. On the other hand, we found the training could be unstable when training both tasks simultaneously, therefore we decided to train the encoder and image generation decoder first, and only then train the segmentation decoder. Further work is required to explore other training strategies and establish the optimal approach - this includes the possibility of using a validation dataset to detect over- or under-fitting of either task facilitating better stopping criteria (but this requires a larger development dataset than we had available here).

Another limitation of our study is the lack of comparison to other methods. However, to the best of our knowledge, there are no other published methods that can produce a synCT image and segmentations using just the data we used in this study. The existing studies were mostly relying on propagating contours from the pCTs, which we have already identified as not suitable for highly deformable organs such as the bladder. Those propagated contours could be used directly or facilitated training of a supervised CBCT segmentation method [10]. An alternative approach would be to employ a cascade, when one method would be dedicated for generating only synCT, which would then be segmented using a segmentation tool trained on pCTs [11]. Both approaches do not address the problem simultaneously and introduce a lot of uncertainty along the way, starting from challenges with accurate registration of the structures between pCT and CBCT, through the quality of synCT generation which would heavily influence the quality of segmentations. The evaluation of our method against such approaches will be the focus of future work.

A final limitation of the method we presented here is that it is 2D based and so processes each slice independently. We experimented with 3D networks but found their increased memory footprint greatly limited the depth of the networks we could employ, which led to poor results. In the future we will investigate the use of '2.5D' networks, that take multiple adjacent slices as input and generate the outputs corresponding to the central slice.

5 Conclusion

In this study we presented a proof-of-concept method for jointly generating synthetic CT and OAR segmentation from CBCT. As the method uses an unsupervised approach it does not require paired CBCTs and pCTs and does not require segmentations on the CBCTs. We demonstrated our method on a relatively small dataset and showed that such an approach is possible and can provide promising results.

Acknowledgement

CV was supported by the Royal Academy of Engineering under the Research Fellowship scheme (RF\201718\17140). AS and JRM acknowledge support from the Wellcome/EPSRC Centre for Interventional and Surgical Sciences (WEISS) (203145/Z/16/Z). The authors would like to acknowledge Ana-Roberta Nita from School of Clinical Medicine, University of Cambridge, Cambridge, UK, for their help with identifying patients to be included in the study.

References

[1] Glide-Hurst, Carri K., et al. "Adaptive radiation therapy (ART) strategies and technical considerations: a state-of-the-ART review from NRG oncology. "*International Journal of Radiation Oncology* Biology* Physics* 109.4 (2021): 1054-1075. DOI: 10.1016/j.ijrobp.2020.10.021

[2] Kong, Vickie C., Andrea Marshall, and Hon Biu Chan. "Cone beam computed tomography: the challenges and strategies in its application for dose accumulation." *Journal of Medical Imaging and Radiation Sciences* 47.1 (2016): 92-97. DOI: 10.1016/j.jmir.2015.09.012

[3] Mushonga, Melinda, et al. "Unanticipated Radiation Re-planning for Stage III Non-Small Cell Lung Cancer." *Advances in Radiation Oncology* (2023): 01275. DOI:10.1016/j.adro.2023.101275

[4] Spadea, Maria Francesca, et al. "Deep learning based synthetic-CT generation in radiotherapy and PET: a review." *Medical physics* 48.11 (2021): 6537-6566. DOI: 10.1002/mp.15150

[5] Szmul, Adam, et al. "Deep learning based synthetic CT from cone beam CT generation for abdominal paediatric radiotherapy." *Physics in Medicine & Biology* 68.10 (2023): 105006. DOI: 10.1088/1361-6560/acc921

[6] Zhu, Jun-Yan, et al. "Unpaired image-to-image translation using cycle-consistent adversarial networks." *Proceedings of the IEEE international conference on computer vision*. 2017. DOI: 10.48550/arXiv.1703.10593

[7] Modat, Marc, et al. "Fast free-form deformation using graphics processing units." *Computer methods and programs in biomedicine* 98.3 (2010): 278-284. DOI: 10.1016/j.cmpb.2009.09.002

[8] Salimans, Tim, et al. "Improved techniques for training gans." *Advances in neural information processing systems* 29 (2016). DOI: 10.48550/arXiv.1606.03498

[9] Alzubaidi, Laith, et al. "Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions." *Journal of big Data* 8 (2021): 1-74. DOI: 10.1186/s40537-021-00444-8

[10] Liang, Xiao, et al. "Deep learning based direct segmentation assisted by deformable image registration for cone-beam CT based auto-segmentation for adaptive radiotherapy." *Physics in Medicine & Biology* 68.4 (2023): 045012. DOI: 10.1088/1361-6560/acb4d7

[11] Dai, Xianjin, et al. "Synthetic CT-aided multiorgan segmentation for CBCT-guided adaptive pancreatic radiotherapy." *Medical Physics* 48.11 (2021): 7063-7073. DOI: 10.1002/mp.15264

Improving Clinical Target Volume Segmentation throughout a Gastric Cancer Radiotherapy Clinical Trial

Phillip Chlap[1,2], Mark Lee[1,2], Trevor Leong[4], Matthew Field[1,2], Jason Dowling[1,5], Hang Min[1,5], Julie Chu[4], Jennifer Tan[4], Phillip K Tran[4], Tomas Kron[4], Annette Haworth[3], Martin A Ebert[6,7], Shalini K Vinod[1,2] and Lois Holloway[1,2,3]

[1]University of New South Wales and the Ingham Institute for Applied Medical Research, Sydney, Australia
[2] Liverpool and Macarthur Cancer Therapy Centres, South-Western Sydney Local Health District, Sydney, Australia
[3]Institute for Medical Physics, School of Physics, The University of Sydney, Sydney, Australia
[4]Peter MacCallum Cancer Centre and Sir Peter MacCallum Department of Oncology, University of Melbourne, Melbourne, Australia
[5]Australian e-Health Research Centre, CSIRO, Brisbane, Australia
[6]Sir Charles Gairdner Hospital and University of Western Australia, Perth, Australia
[7]School of Physics, Mathematics and Computing, University of Western Australia, Perth, Australia

Abstract Radiotherapy clinical trials can benefit from an accurate segmentation model to help ensure consistency of contour definitions across patients treated in the trial, however, only a small dataset is often available initially as trial recruitment begins and deep learning segmentation models typically require large datasets for training. In this work we investigate the effect of retraining a segmentation model as more data becomes available throughout a trial. Models were trained for Clinical Target Volume (CTV) segmentation in the TOPGEAR (gastric cancer) clinical trial. There were 5 iterations of model training with datasets containing only 10 cases in the initial model, then increasing in increments to 43 in the last iteration. Each model was tested on a hold-out set of 50 cases, where the Dice Similarity Coefficient (DSC), Surface DSC (sDSC), Mean Distance to Agreement (MDA) and Hausdorff Distance were computed between the segmentation model and the manual contour. The DSC improved from 0.8 to 0.82 between the initial model and the last model. Improvement was also observed for other metrics suggesting the proposed methodology for retraining a segmentation model would be useful for prospective clinical trials.

1 Introduction

Current state-of-the-art models for the segmentation of structures in medical images commonly utilize deep learning. Relatively large datasets are typically required to achieve high quality results, however, they can be difficult to obtain especially in the medical domain. These datasets are often obtained from retrospective clinical datasets which may already have segmentations available for model training.

For radiotherapy clinical trials, ensuring that contour definitions across different patients and participating centers are consistent with the trial protocol is important for patient safety and outcome analysis [1]. Training an auto-segmentation model to predict an estimate of the contour that conforms to trial protocol is one approach to providing automated contour quality assurance (QA) [2].

If the clinical trial is successful, this model may be used to ensure future patients receive treatment consistent with the trial protocol. To train such a model the entire dataset could be used to train a model following the completion of the trial.
However such a model may also be used during the trial to provide feedback to centers and allow contours to be corrected prior to patients being treated and reduce burden on manual quality assurance resources. In this scenario, it may not be feasible to prepare a sufficiently large training dataset for use in the trial as the structure to be segmented may be novel and defined as part of the clinical trial itself.

The TOPGEAR clinical trial defines a novel Clinical Target Volume (CTV) for treatment of gastric cancer patients [3]. In this work, we explore a methodology for training a CTV segmentation model for TOPGEAR by simulating the initial phase of the trial. Although the trial has now completed recruiting patients, the clinical trial dataset is used to demonstrate how this process would be applied to a prospective clinical trial to help determine how much data is required to train a useful segmentation model.

2 Materials and Methods

Australasian Gastro-Intestinal Trials Group (AGITG) AG0407GR/TROG 08.08 "TOPGEAR" (NCT01924819) is a phase II/III trial for patients with resectable gastric cancer. Patients were recruited and randomized to either a chemoradiation arm or chemotherapy only arm, both prior to resection. The CTV defined in the trial protocol to be used to deliver radiation to patients in the chemoradiation arm requires expertise to contour. The trial protocol describes which anatomy to include or exclude from the CTV which is defined on the patient's Computed Tomography (CT) scan. To help ensure consistent definition throughout the trial, manual contour quality assurance was conducted on the first 5 patients treated at each participating center followed by every third patient thereafter. By the time the trial closed to recruitment, 174 patients from the chemoradiation arm had a manual review of the treatment plan conducted, including of the CTV.

To help reduce the resources required for manual review of the TOPGEAR CTV contours, we defined a methodology for training an auto-segmentation model able to predict a valid CTV structure on an unseen case. Rather than utilize the entire available 174 patient dataset upfront to train a CTV segmentation model (which may also be useful to help

QA future treatments intended to conform to TOPGEAR's protocol), we used this opportunity to define and evaluate a methodology for training a model throughout the course of a clinical trial with the goal that the model could be used with increasing confidence and accuracy as the trial progresses.

Of the 174 cases, 50 were randomly selected as a hold-out set to be used for validation of models trained. The remaining cases were split into groups of 10 and used for training. This split was performed randomly to help generalize the methodology to clinical trials more broadly and reduce bias in the order in which patients were treated in TOPGEAR (where some centers treated more patients at the start of the trial than others).

The initial set of 10 cases had the CTV defined by 5 Radiation Oncologists on each case. This set was deemed feasible to obtain for the first 10 patients in a clinical trial or prior to a clinical trial. While having several observers on this data was useful for determining the amount of inter-observer variability present in the data, these structures were combined into a single structure using the STAPLE method for this work [4].

An additional 4 sets of 10 cases from the trial dataset were then used to complete a total of 5 iterations of model training. Furthermore, upon review of the trial cases together with 2 experienced Radiation Oncologists (who were directly involved in developing the trial protocol), it was determined that not all cases that were reviewed during the trial were acceptable due to significant artifacts in the imaging (e.g. poorly reconstructed 4DCT) or use of the wrong primary dataset (e.g. use of the average intensity projection from a 4DCT resulting in blurry outlines of structures from motion artifact). Therefore, these cases were either corrected or excluded when correction wasn't possible (due to issues such as poor image quality).

Another notable property of the initial set of 10 cases was that there were no contrast enhanced cases included in this set. In the iterations that followed, contrast enhanced cases were included (approx. 25%), to help improve segmentation accuracy on those scans.
Table 1 defines the subsets of data used at each iteration of CTV segmentation model training.

To train the segmentation model, the 2D nnUNet was used [5]. A total of 5 models were trained over 5 iterations with each model being trained using data from previous iterations as well as additional data. The trained models were then applied to each case in the hold-out set. The Dice Similarity Coefficient (DSC), Mean Distance to Agreement (MDA), Hausdorff Distance (HD) and Surface DSC (sDSC) were computed between the auto-segmentation and the manual segmentation. A one-sided t-test was also performed to determine the significance of the improvement between iterations based on the p-value (significant when $p<0.05$).

Table 1: The number of cases which required correction (Corr) or were excluded (Excl) in each subset of 10 cases. The total dataset size column is the final number of cases included for training at each iteration.

Subset	Cases	Corr	Excl	Used for training	Total Dataset size
Initial	10	-	-	10	10
Iteration 1	10	0	3	7	17
Iteration 2	10	6	3	7	24
Iteration 3	10	3	0	10	34
Iteration 4	10	5	1	9	43

3 Results

The mean of each of the metrics improved with increasing iterations of the model (except for HD and MDA in iteration 2). Figure 1 plots the mean value of each metric calculated using the hold-out set. The significance of the improvement from each iteration to the next as well as between each iteration and the initial model is also provided.

The mean DSC improved from 0.80 to 0.82 from the initial to the final model. A significant improvement was seen between each model based on DSC, except in iteration 2. The sDSC showed similar improvement based on the mean (from 0.57 to 0.60 between initial and final models) and significance (all significant except iteration 2 and 4).

The surface distance metrics showed overall improvement in mean from the initial to the final model (from 39.6mm to 35.6mm for HD and 5.3mm to 4.7 for MDA), however, the overall significance was lower between iterations than for DSC and sDSC.

Figure 2 visualizes the manual contour and the segmentations from the initial and final model for an example case from the hold-out set.

4 Discussion

Throughout the 5 iterations of model training performed, the overall model performance improved. The abdominal anatomy is a relatively challenging area for auto-segmentation due to large anatomical differences between patient anatomy and the presence of gas pockets. By including contrast enhanced images, the iterations following the initial model, accuracy improved on these scans in the hold-out set.

With 5 iterations of model training, it is difficult to conclude whether the segmentation quality is beginning to plateau

and how many more training cases would help improve model performance to a specific level. However, the methodology does demonstrate overall model improvement at each iteration indicating that this approach may be useful for use in a prospective clinical trial.

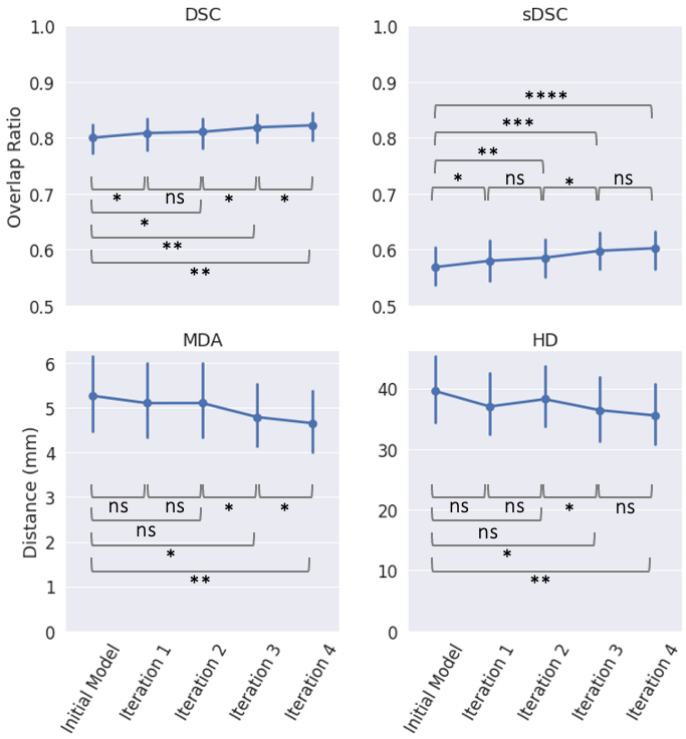


Figure 1: Improvement of the mean Dice Similarity Coefficient (DSC), Surface DSC (sDSC), Mean Distance to Agreement (MDA) and Hausdorff Distance (HD) with standard deviation error bars across models trained (blue). Improvement between each iteration to the next and each iteration to the initial model are also indicated (black) where ns: not significant, *: p<0.05, **: p<0.01, ***: p<0.001, ****: p<0.0001.

Since a significant improvement was not determined between each iteration of model training, a hard stopping criterion is difficult to define. Further iterations may contain data which will benefit the overall performance of the model, even if the previous iteration didn't improve significantly. Further investigation is required to determine the target value of a metric to ensure the model is useful for the task of contour QA. Since uncertainty is present in the data, which can be determined through inter-observer variation, metrics assessing the segmentation quality can only be expected to reach this level as a maximum.

The use of active learning may further benefit model accuracy while decreasing manual review resources required [6]. This would involve automatically selecting the cases in the next training iteration that would most benefit model performance. Like this, cases which are most likely to significantly improve model performance can be included in further iterations of model training while others are excluded.

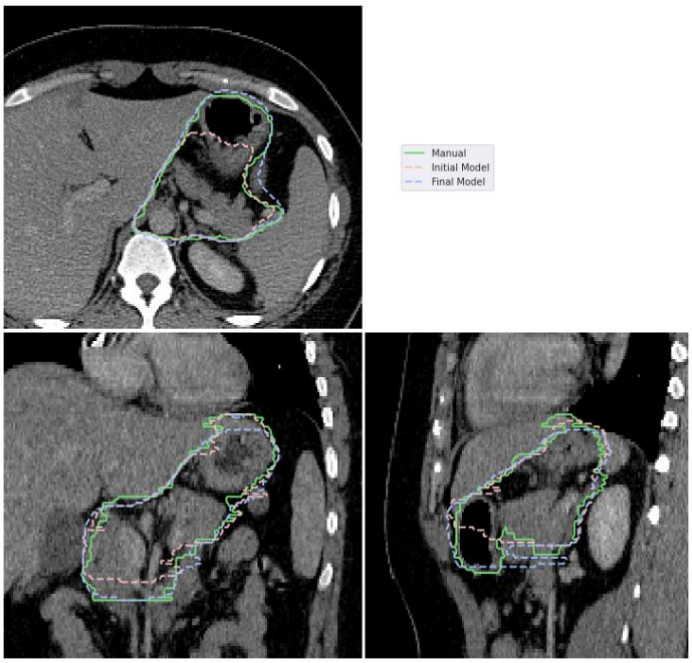


Figure 2: Cross section of an example case from hold-out set visualizing the manual CTV contour (green), the initial model's segmentation (red) and the final model's segmentation (blue). The initial DSC for this case was 0.84 which improved to 0.88 for the final model.

5 Conclusion

The methodology for CTV segmentation throughout the TOPGEAR clinical trial demonstrated improvement of segmentation accuracy across models trained. Further research is required to determine the benefit of training beyond 5 iterations as well as experimentation with active learning approaches to maximize model performance given the manual review resources available. Different segmentation approaches such as the 3D nnUNet will also be explored. Finally, these models will be applied to the task of contour quality assurance within the TOPGEAR trial.

References

[1] L. J. Peters *et al.*, "Critical impact of radiotherapy protocol compliance and quality in the treatment of advanced head and neck cancer: Results from TROG 02.02," *Journal of Clinical Oncology*, vol. 28, no. 18, pp. 2996–3001, Jun. 2010, doi: 10.1200/JCO.2009.27.4498.

[2] H. Min *et al.*, "Automatic radiotherapy delineation quality assurance on prostate MRI with deep learning in a multicentre clinical trial," *Phys Med Biol*, vol. 66, no. 19, p. 195008, 2021, doi: 10.1088/1361-6560/ac25d5.

[3] T. Leong *et al.*, "TOPGEAR: A randomised phase III trial of perioperative ECF chemotherapy versus preoperative chemoradiation plus perioperative ECF chemotherapy for resectable gastric cancer (an international, intergroup trial of the AGITG/TROG/EORTC/NCIC CTG)," *BMC Cancer*, vol. 15, no. 1, 2015, doi: 10.1186/s12885-015-1529-x.

[4] S. K. Warfield, K. H. Zou, and W. M. Wells, "Simultaneous truth and performance level estimation (STAPLE): An algorithm for the validation of image segmentation," *IEEE Trans Med Imaging*, vol. 23, no. 7, pp. 903–921, Jul. 2004, doi: 10.1109/TMI.2004.828354.

[5] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation," *Nat Methods*, vol. 18, no. 2, pp. 203–211, 2021, doi: 10.1038/s41592-020-01008-z.

[6] S. Budd, E. C. Robinson, and B. Kainz, "A survey on active learning and human-in-the-loop deep learning for medical image analysis," *Medical Image Analysis*, vol. 71. Elsevier B.V., p. 102062, 2021. doi: 10.1016/j.media.2021.102062.

OAR-Weighted Dice Score: A spatially aware, radiosensitivity aware metric for target structure contour quality assessment

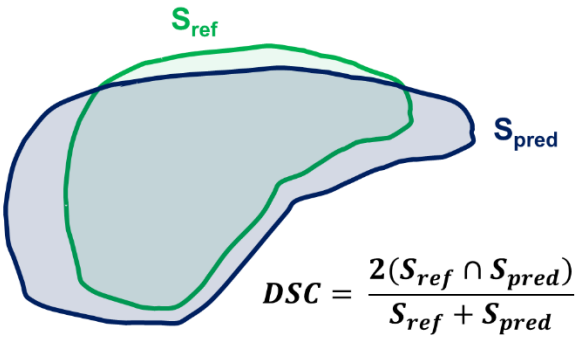
Lucas McCullum[1,2], Kareem A. Wahid[2,3], Barbara Marquez[1,4], Clifton D. Fuller[2]

[1]UT MD Anderson Cancer Center UTHealth Houston Graduate School of Biomedical Sciences, Houston, USA
[2]Department of Radiation Oncology, The University of Texas MD Anderson Cancer Center, Houston, USA
[3]Department of Imaging Physics, The University of Texas MD Anderson Cancer Center, Houston, USA
[4]Department of Radiation Physics, The University of Texas MD Anderson Cancer Center, Houston, USA

Abstract The Dice Similarity Coefficient (DSC) is the current *de facto* standard to determine agreement between a reference segmentation and one generated by manual / auto-contouring approaches. This metric is useful for non-spatially important images; however, radiation therapy requires consideration of nearby Organs-at-Risk (OARs) and their radiosensitivity which are currently unaccounted for with the traditional DSC. In this work, we introduce the OAR-DSC which accounts for nearby OARs and their radiosensitivity when computing the DSC. We illustrate the importance of this through cases where two proposed contours have similar DSC, but lower OAR-DSC when one contour expands closer to the surrounding OARs. This work is important because the OAR-DSC may be used by deep learning auto-contouring algorithms in a radiation therapy specific loss function, thereby progressing on the current disregard for the importance of these differences on the final radiation dose plan generation, delivery, and risks of patient toxicity.

1 Introduction

Automated structure contouring, also known as auto-contouring, is a method to automatically extract desired regions in an image. Typically, ground truth contours are made manually by in-field experts and compared to the predictions from an auto-contouring algorithm. One standard *de facto* metric of measuring the agreement between the ground truth and predicted segmentations in the current era of deep learning is the Dice Score, or more formally the Dice Similarity Coefficient (DSC) [1,2], as described in Figure 1.

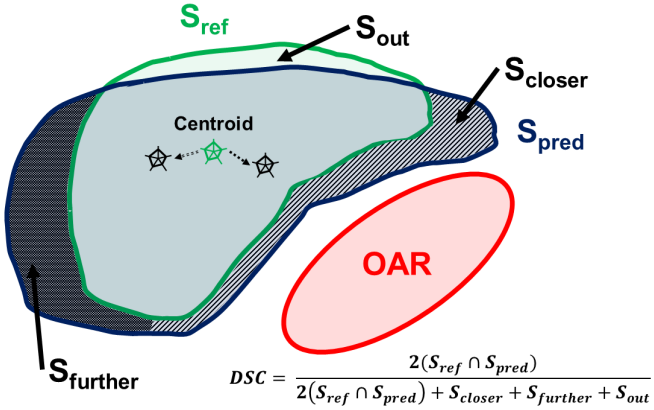


The DSC has subsequently extended to and now dominates the relevant literature in the field of radiation oncology to compare ground truth segmentations of tumors and surrounding normal tissue sensitive to unintended radiation, also known as Organs-at-Risk (OAR). Since deep learning models are often trained using a negative / positive feedback system where lower DSC is penalized and higher DSC is reinforced, equivalent DSC will produce the same feedback during model training, regardless of where the errors occurred. This is because the original formulation of DSC does not account for spatial information since it did not need to for identifying performance of auto-segmentation of objects in images.

Therefore, relating this concern back to the field of radiation oncology, two predicted tumor contours with the same DSC will both be equally encouraged even if one contour disagrees closer to the OAR, resulting in a potentially increased probability of toxicity to the OAR given the segmentation is accepted for radiation dose plan creation. Further, only one prior work to the author's knowledge has attempted weighting the DSC for radiotherapy applications, however the dose was considered which necessitates unmanageable computational resources at this time [3]. Hence, in this work we introduce the concept and formulation of the OAR-Weighted DSC (OAR-DSC) which accounts for spatial location of disagreements between ground truth and auto-contouring generated contours in relation to the surrounding OARs in 2D.

2 Materials and Methods

To account for surrounding OARs, the DSC can be rearranged and calculated as follows in Figure 2.



expression can be weighted to achieve the OAR-DSC as will be shown below.

This formulation is further described where S_{ref} is the reference ground truth segmentation, S_{pred} is the predicted auto-contouring, S_{closer} is the joint region of voxels not in S_{ref} whose inclusion brings the centroid of S_{ref} closer to the OAR, $S_{further}$ is the joint region of voxels not in S_{ref} whose inclusion brings the centroid of S_{ref} further from the OAR, and S_{out} is the remaining region in the reference ground truth segmentation not covered by the predicted auto-contouring. This also applies to potential voxels inside S_{ref} which overlap with the OAR; however, these will receive the maximum penalty weighting as will be shown below in the description of the OAR-DSC.

To achieve the OAR-DSC, a weighting factor, w, was applied to S_{closer} as a function of its distance to the OAR to artificially decrease the DSC. This weighting factor cannot be greater than 1 and no weighting factor is applied to $S_{further}$ to counteract the effects of S_{closer} since this could lead to artificially inflated DSC which may be misleading.

The weighting factor is determined with a simple parameterized exponential decay and scaled if in the presence of multiple OARs based on the maximum distance from the reference contour and the nearest distance to the OAR, d. A detailed derivation for a given OAR (i.e., OAR_i) is given:

$$OAR\text{-}DSC = \frac{2(S_{ref} \cap S_{pred})}{2(S_{ref} \cap S_{pred}) + \frac{S_{closer}}{w_{OAR^i}} + S_{further} + S_{out}}$$

Where:

$$w_{OAR^i} = \frac{1}{n_{S_{closer}}} \sum_{j=1}^{n_{S_{closer}}} e^{\alpha\left[\frac{d_{max,OAR^i}}{d_{max,OAR}}\left(\frac{d^j_{pred}}{d^j_{pred}+d^j_{ref}}\right)-1\right]}$$

In this representation, an iteration across all the voxels in S_{closer} is done and at each step, d_{max,OAR^i} is given as the maximum of all the nearest distances from the reference contour to the given OAR_i, $d_{max,OAR}$ is given as the maximum of all the nearest distances from the reference contour to the OAR across all considered OARs, d^j_{pred} is the nearest distance from the current voxel to S_{pred}, and d^j_{ref} is the nearest distance from the current voxel to S_{ref}. The α parameter is representative of the radiation sensitivity of the OAR, therefore increasing values of α will result in decreasing OAR-DSC for the same case. It is important to note that when α = 0, then w_{OAR^i} = 1, which implies that the OAR-DSC is simply equivalent to the DSC. The OAR-DSC can be computed for each OAR and averaged together to provide a comprehensive score to account for all OARs. A graphical summary of the steps required to calculate the OAR-DSC can be seen below in Figure 3.

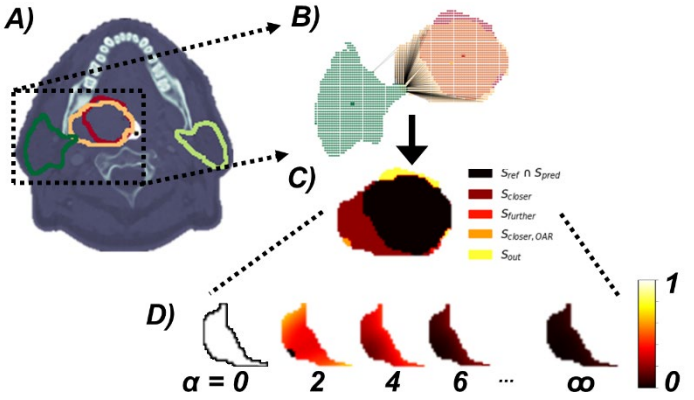
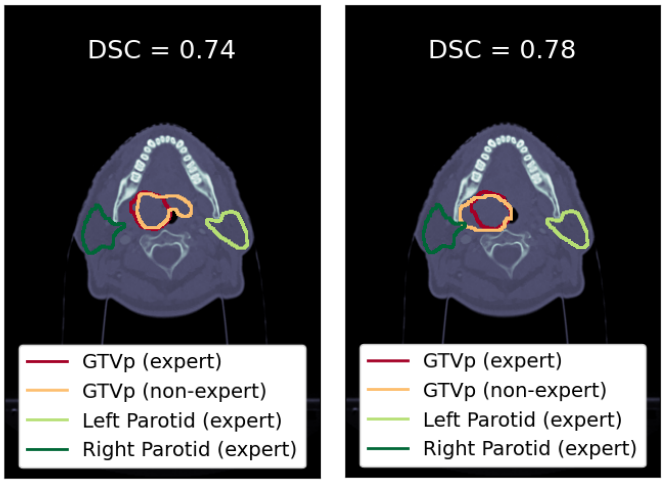


Figure 3: Detailed graphical summary of the derived maps for the computation of the OAR-DSC of the head and neck region. A) desired target and OAR contours are shown on an anatomical Computed Tomography (CT) image, B) the contours and their voxels are extracted and distances are computed to the OAR, C) voxels are classified into the relevant categories for the OAR-DSC, and D) weighted maps are computed for the voxels inside of S_{closer}. Note, the map for α = 0 is all 1 which is white and is represented with a black contour.

Example cases of the OAR-DSC applied in different anatomical sites was done by utilizing the open-source Contouring Collaborative for Consensus in Radiation Oncology (C3RO) segmentations for the head and neck, gynecological, and gastrointestinal sites [4]. Ground truth reference contours and OAR contours were given by the Simultaneous Truth and Performance Level Estimation (STAPLE) expert segmentations while simulated auto-contouring contours were given by select non-expert segmentation cases. The time required to compute the OAR-DSC compared to the DSC is negligible (~5 seconds for two OARs and six α values) due to the vectorization of the operations. The code for calculating the OAR-DSC given contours in Neuroimaging Informatics Technology Initiative (NIfTI) format with some examples from the C3RO segmentations is freely available at: https://github.com/Lucas-Mc/OAR-weighted_Dice-Score.git.

3 Results

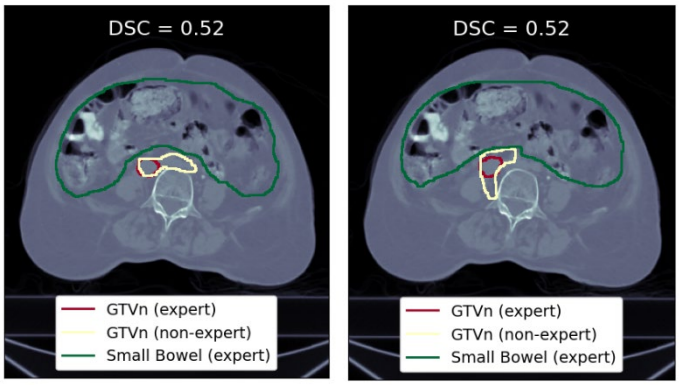
Applying the OAR-DSC to a few sample cases from C3RO illustrates the following results.



	α					
	0	2	4	6	8	10
OAR-DSC (left case)	0.74	0.71	0.67	0.64	0.61	0.58
OAR-DSC (right case)	0.78	0.68	0.56	0.47	0.42	0.40

Figure 4: Comparison of two head and neck cancer cases with similar DSC for the GTVp, yet different proximity to the OARs (left and right parotid). The cumulative OAR-DSC was calculated for each case at different α levels, demonstrating increased consideration for the OARs in the right case.

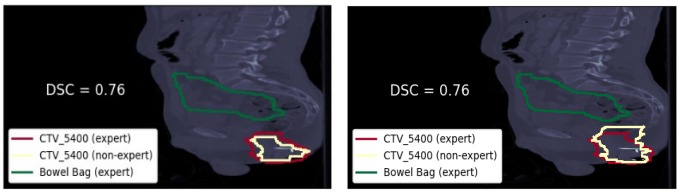
Figure 4 show a comparison for a head and neck cancer case of two contours with similar DSC scores, yet very different implications for the treatment planning, delivery, and control of toxicity due to the proximity to the OARs (left and right parotid). In fact, the predicted target contour for the primary Gross Tumor Volume (GTVp) which invades the left parotid OAR is rated with a higher unweighted DSC (α = 0), thus being preferred in a deep learning training session. Therefore, the OAR-DSC was calculated for each at different α levels, demonstrating increased consideration for the OARs in the right case as its OAR-DSC falls below the left case with increasing α values.



	α					
	0	2	4	6	8	10
OAR-DSC (left case)	0.52	0.50	0.48	0.45	0.42	0.39
OAR-DSC (right case)	0.52	0.40	0.29	0.21	0.15	0.11

Figure 5: Examples of OAR-DSC as it applies to the gynecological region.

Figure 5 shows a sample case for the gynecological region with the nodal Gross Tumor Volume (GTVn) set as the target volume and small bowel set as the OAR. Notice how the left case appears that it should have a lower OAR-DSC than the right case, however this is not the case since a majority of the non-matching voxels work to move the centroid of the reference contour further away from the OAR. The non-matching voxels also border the OAR more closely in the right case increasing the weighting factor and decreasing the OAR-DSC.



	α					
	0	2	4	6	8	10
OAR-DSC (left case)	0.76	0.76	0.76	0.76	0.76	0.76
OAR-DSC (right case)	0.76	0.71	0.65	0.59	0.53	0.48

Figure 6: Examples of OAR-DSC as it applies to the gastrointestinal system.

Figure 6 shows a sample case for the gynecological region with the Clinical Target Volume to receive 5400 cGy (CTV_5400) set as the target volume and bowel bag set as the OAR. Although both the left and right cases have contours with a DSC of 0.76, the contour in the right case extends closer to the OAR and thus has a decreasing OAR-DSC at higher α levels. The left case keeps the same OAR-

DSC throughout since it is entirely contained within the reference contour.

4 Discussion

Since the OAR-DSC inherently includes the DSC at α values of 0, this metric has demonstrable translatability to current clinical practice due to simple adjustments of α moving from traditional DSC to an OAR-weighted DSC. However, consideration should be given to standardize the α values and correlate them with validated radiation sensitivity outcomes, such as those provided by QUANTEC including mean and max dose constraints which could be translated to scalable α values [5].

The limitation of this approach is that care should be taken to account for all desired OARs during the determination of the OAR-DSC, otherwise the auto-contouring algorithm will be unaware of their presence and possibly suggest contours near the OARs. However, this is analogous to the current practice where no considerations are taken for OARs and even more considerate since at least some weighting is being done. Further, this formulation does not specify requirements for utilizing the Gross Tumor Volume (GTV), Clinical Target Volume (CTV), Planning Target Volume (PTV), or Planning organ at Risk Volume (PRV), however these could be included as the boundaries for the segmentations and used in the same way. This practice, however, should be standardized to ensure comparable OAR-DSC across anatomical sites and centers. It is possible to extend the formulations introduced here to add separate weightings to GTV and CTV or OAR and PRV if equal concern is not desired for the margins as compared to the definite volumes. The formulation also currently uses relative distances between the voxels and OARs to determine radiation sensitivity fraction, but future work could apply absolute distance-based weighting instead based on radiobiological concerns and setup uncertainty (i.e., less than 2 mm should receive the maximum loss weighting). A final limitation of the OAR-DSC is that the OAR segmentations would have to be generated first in a fully automated contouring workflow so that they can be considered when generating the target contours.

Future work includes investigating alternative metrics which could be modified to fit in this framework, such as the surface DSC, and how the OAR-DSC can be expanded to enable more computationally efficient workflows [6]. Additionally, a 3D implementation of the OAR-DSC would enable consideration of all surrounding OARs instead of just those which are in-plane.

5 Conclusion

We have introduced the OAR-DSC, a novel method to account for location-specific deviations between an auto-contouring contour and reference contour in relation to surrounding OARs in 2D. Thereby, cases where two proposed auto-contours have similar DSC, we have shown that the OAR-DSC differs between the two especially in cases where one contour expands closer to the surrounding OARs. This metric can be used by deep learning auto-contouring algorithms as a method to add a radiation therapy specific loss function to the generated contours, progressing on its current disregard for the critical influence of these differences on the final radiation dose plan and its risks to the probability of patient toxicity.

References

[1] Sørensen, T. (1948). "A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons". Kongelige Danske Videnskabernes Selskab. 5 (4): 1–34.

[2] Dice, Lee R. (1945). "Measures of the Amount of Ecologic Association Between Species". Ecology. 26 (3): 297–302. https://doi.org/10.2307/1932409

[3] Fried Z. (2020). “Application of a weighted Dice Similarity Coefficient (DSC) for Structure Comparison in Radiation Treatment Planning”. Master’s Thesis. Oregon State University.

[4] Wahid, K.A., Lin, D., Sahin, O. et al. Large scale crowdsourced radiotherapy segmentations across a variety of cancer anatomic sites. Sci Data 10, 161 (2023). https://doi.org/10.1038/s41597-023-02062-w

[5] Bentzen, S. M., Constine, L. S., Deasy, J. O., Eisbruch, A., Jackson, A., Marks, L. B., Ten Haken, R. K., & Yorke, E. D. (2010). Quantitative Analyses of Normal Tissue Effects in the Clinic (QUANTEC): an introduction to the scientific issues. International journal of radiation oncology, biology, physics, 76(3 Suppl), S3–S9. https://doi.org/10.1016/j.ijrobp.2009.09.040

[6] Nikolov, S., Blackwell, S., Zverovitch, A., Mendes, R., Livne, M., De Fauw, J., ... & Ronneberger, O. (2018). Deep learning to achieve clinically applicable segmentation of head and neck anatomy for radiotherapy. arXiv preprint arXiv:1809.04430. https://doi.org/10.48550/arXiv.1809.04430

Acknowledgments

Research support was provided by the NIH/NCI Cancer Center Support Grant (CCSG) Pilot Research Program Award from the UT MD Anderson CCSG Image-Driven/Biologically-Informed Therapy Program (IDBT) (PIs Brock/Fuller, P30CA016672-43S1). Lucas McCullum is support by an NIH Diversity Supplement (3R01CA257814-02S2). Kareem A. Wahid is supported by the NCI NRSA Image Guided Cancer Therapy Training Program (T32CA261856). Barbara Marquez is supported by the American Legion Auxiliary Fellowships in Cancer Research and funded by UTHealth Innovation for Cancer Prevention Research Training Program Pre-doctoral Fellowship (Cancer Prevention and Research Institute of Texas grant #RP210042).

Towards Automatic Identification of Missing Tissues using a Geometric-Learning Correspondence Model

Eliana Vasquez Osorio[1,2] and Edward Henderson[1]

[1]Faculty of Biology, Medicine and Health, University of Manchester, Manchester, UK
[2]The Christie NHS Foundation Trust Hospital, Manchester, UK

Abstract Missing tissue presents a big challenge for dose mapping, e.g., in the reirradiation setting. We propose a pipeline to identify missing tissue on intra-patient structure meshes using a previously trained geometric-learning correspondence model. For our application, we relied on the prediction discrepancies between forward and backward correspondences of the input meshes, quantified using a correspondence-based Inverse Consistency Error (*cICE*). We optimised the threshold applied to *cICE* to identify missing points in a dataset of 35 simulated mandible resections. Our identified threshold, 5.5 mm, produced a balanced accuracy score of 0.883 in the training data, using an ensemble approach. This pipeline produced plausible results for a real case where 25% of the mandible was removed after a surgical intervention. The pipeline, however, failed on a more extreme case where 50% of the mandible was removed. This is the first time geometric-learning modelling is proposed to identify missing points in corresponding anatomy.

1 Introduction

Anatomy of patients undergoing medical treatment often has large 'atypical' changes. In particular, for reirradiation cases in the head and neck region, mandible surgeries due to osteoradionecrosis [1] are common, resulting in large anatomical mismatches. The missing tissue between images presents challenges to registration algorithms, with few methods explicitly accounting for them [2]. In this study, we propose a pipeline to identify missing tissue on surface meshes of structures using a previously trained geometric-learning correspondence model. This step will enable robust contour-based registrations in particular for estimating cumulative dose metrics in the reirradiation setting.

2 Methods

We used a model previously trained to identify dense point correspondences for head and neck organs at risk [3]. The correspondence model was trained using an unsupervised approach, where both the geometry of the shapes as well as the imaging information of the planning CT scans were used during training to improve model performance. This model has been shown to be effective for identifying inter-patient correspondences for the brainstem, mandible, parotid glands, spinal cord and submandibular glands.
At inference time, the model receives two meshes representing the structures, i.e., the source and target meshes, and assigns a vertex on the target mesh to every vertex from the source mesh (defined by n vertices). More formally, the correspondence between the source and target vertices is represented as the vector $M^{s\to t}$, of n positions, where the correspondence for a given $i-th$ vertex is $M^{s\to t}(i)=j$, with j being the index of the corresponding vertex on the target mesh. Note that the model's predictions are not symmetric, as some target vertices may not be assigned to a source vertex.

2.1 Identification of missing tissue

We propose to identify missing tissue by levering the discrepancy between correspondences predicted by the model when the role of the input meshes are reversed. The first step is to quantify the discrepancies between the correspondence predictions. For this, we estimated the inverse consistency error based on the correspondences, *cICE* figure 1, using the following equation:

$$cICE_i = d(x(i), x(M^{t\to s}(M^{s\to t}(i))))$$

where $d()$ is the Euclidean distance between two points, $x(i)$ is the 3D coordinates of the $i-th$ source vertex, $M^{s\to t}$ and $M^{t\to s}$ are the correspondence vectors predicted by the model for the two directions, i.e., the *forward* (source to target) and *backward* (target to source) correspondences. In the ideal case, $M^{t\to s}(M^{s\to t}(i))$ is equal to i, rendering $cICE(i)$ to zero. As the correspondence model generates discrete predictions, noise is expected on the vertex assignment. Therefore, we applied a median filter to the *cICE* on the source mesh to mitigate for this effect.
Our assumption is that points with the largest discrepancies are missing on the target mesh. Therefore, we classify all

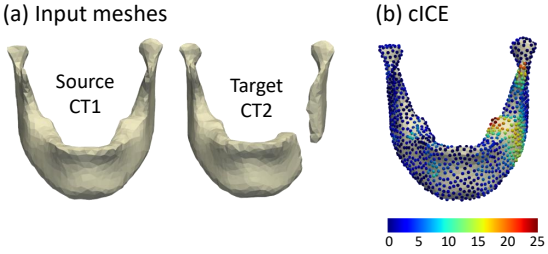


Figure 1: Example of input meshes and the derived *cICE*. Notice that large *cICE* values are located in regions of point mismatch between the source and target meshes.

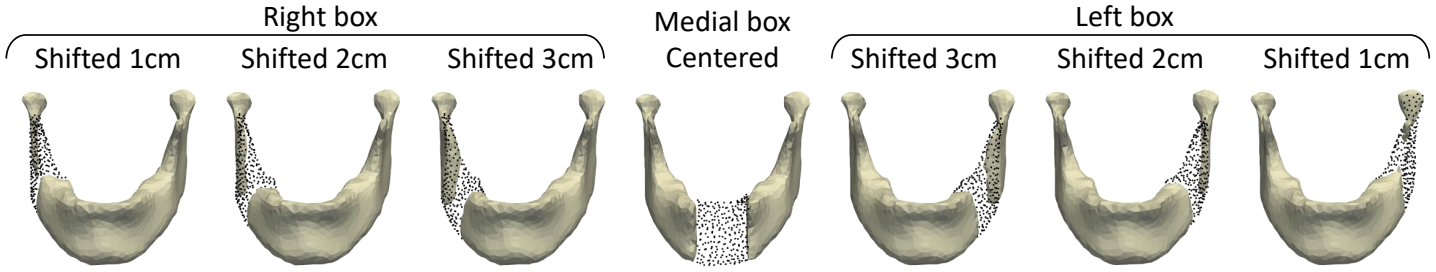


Figure 2: Simulated cuts for the CT2 mandible of a patient.

vertices in the source mesh with a *cICE* value above a given threshold t as missing in the target mesh. As we applied a median filter in the previous step, we again apply a median filter to 'dilate' the selection, in an operation similar to morphological opening filter of binary masks.

2.2 Pipeline optimisation

Mandible contours from 5 patients who underwent reirradiation in the head and neck region were used for this optimisation. Each patient had their mandible delineated on two CT scans: CT1 and CT2, where CT2 was the planning CT scan of the the most recent radiotherapy course. We framed our implementation to predict points present in the CT1 mandible and missing in the CT2 mandible meshes. We optimised two aspects: 1) we compared the performance of an ensemble of multiple correspondence models versus a single model, and 2) we optimised the value of the threshold t.

2.2.1 Ensemble vs single model

As using ensemble approaches have been shown to improve deep-learning performance **??**, we evaluated whether such an approach improved our predictions. For these, we used the predictions of five corresponding models previously generated as part of a 5-fold cross validation[3]. We used a simple ensemble approach: majority vote where only points identified as missing by at least three of these single models were kept.

2.2.2 Threshold optimisation

We conducted a simulation study to find the optimal value of the threshold t for which we generated simulated mandible resections as training data. We evaluated a range of 19 thresholds, ranging from 1 to 10 mm. We selected the best threshold as the one that maximised the prediction accuracy over all training data.

Training data generation We generated training data by artificially cutting the CT2 mandibles using 3cm-wide boxes positioned in different locations, see figure 2. We consistently applied this cutting procedure by using ROI algebra in the scripting interface in RayStation v12.0.100.0 (research) and exporting the generated cut mandibles as masks in MHD format. These masks were then processed to generate triangular meshes following the same pre-processing steps as presented previously [3]. In total, we generated 35 mesh pairs where between 11% and 28% of the mandible volume was removed.

Golden standard Notice that the ground truth is directly available for the CT2 mandibles via the cutting boxes. However, to evaluate the predictions made by our pipeline, these ground truth needed to be 'projected' to the CT1 mandible. Therefore, we used the forward correspondences for the complete CT2 mandibles to generate the golden standard for missing points on the CT1 mandible mesh: for every mesh pair, a given point was classified as missing if it corresponded to a point within the cutting box. We harmonised the golden standard across the five folds using a majority voting filter, i.e., a given point was classified as missing if it was identified as missing by at least 3 single models. See figure 3(a) for an example.

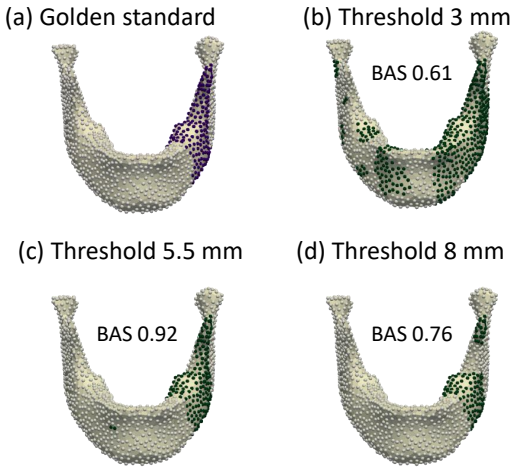


Figure 3: Golden standard for the mesh pair shown in figure 1(a), where the dark points have been identified as missing. (b-d) points identified as missing after applying the proposed pipeline for different thresholds and their balanced accuracy score (BAS). These thresholds are applied to the *cICE* presented in figure 1(b).

Prediction evaluation We evaluated the prediction of missing points for different thresholds against the golden standard using an accuracy score. As we expected to have

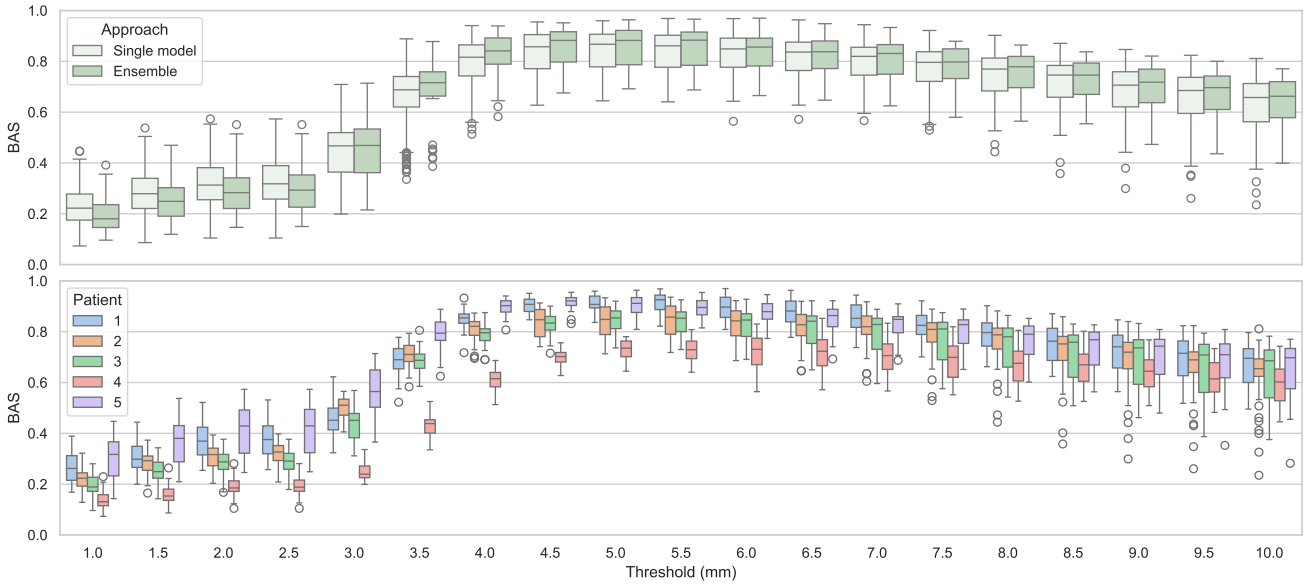


Figure 4: Balanced accuracy score (BAS) for all simulations and tested thresholds, discriminating by single model vs ensemble approach (top) and by patient (bottom).

classifications with imbalanced class distributions, e.g., most points are not missing in CT2, we used the balanced accuracy score (BAS). BAS has been demonstrated to be robust to imbalanced class distributions [4] and it is defined as the average of recall obtained on each class, figure 3. We used the implementation provided in scikit-learn v1.3.2.

2.2.3 Qualitative evaluation of generalisation

We applied the most optimal pipeline and threshold to two extra cases where the patients underwent mandible surgery in between the two radiotherapy courses due to osteoradionecrosis. These patients lost 25% and 50% of the mandible during the surgical intervention, respectively.

3 Results

Figure 4 shows the performance of the pipeline. BAS above 0.8 are observed for both approaches at thresholds between 4.0 and 7.0 mm, with the largest median being 0.867 for the single model approach at 5.0 mm and 0.883 for the ensemble approach at 5.5 mm. The performance of the classification pipeline was different for some patients, with results for patient 4 systematically lower for most thresholds. The pipeline performed best for patients 5 and 1, followed by patients 2 and 3.

Figure 5 shows the results of applying the ensemble pipeline with threshold 5.5 mm on two extra patients who underwent surgical intervention due to mandible osteoradionecrosis after their first radiotherapy course. The results on ORN patient 1 are plausible with the largest region identifying the most anterior part of the mandible. Two small regions, up in the mandible head and the coronoid process were also identified. The prediction for ORN patient 2 was not successful, as most of the points were identified as missing. However, the calculated *cICE* followed a similar pattern as other mesh pairs but with a much higher magnitude, compare figure 6 and 1.

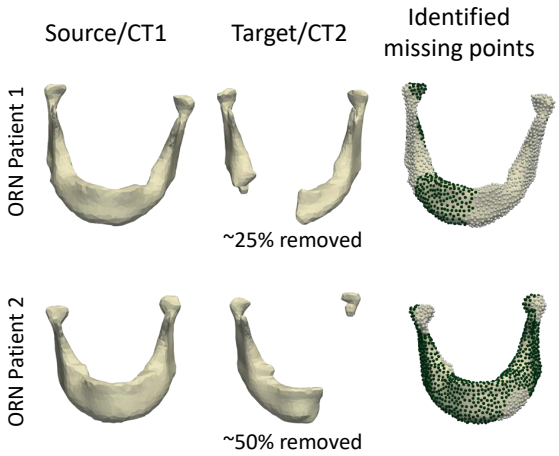


Figure 5: Performance of the ensemble pipeline using the found optimal threshold of 5.5 mm to data of two patients who underwent mandible surgery due to osteoradionecrosis after their first radiotherapy course.

4 Discussion

In this study we proposed, implemented and optimised a pipeline to identify missing tissue on intra-patient organ meshes based on a previously trained geometric-learning correspondence model, previously shown to be effective at identifying inter-patient correspondences for head and neck organs. For our application, we relied on the prediction discrepancies between the forward and backward correspondences quantified via correspondence-based inverse consistency error or *cICE*, where points with *cICE* above a given threshold were identified as missing. We optimised this

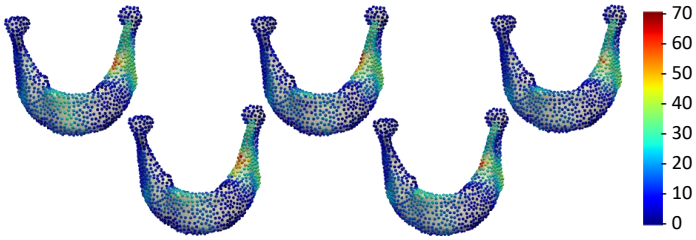


Figure 6: *cICE* for all folds for ORN patient 2. Note the pattern similarity with *cICE* presented in figure 1

threshold in a dataset of 35 simulated mandible resections by evaluating the balanced accuracy score (BAS) between the golden standard and the predictions. Our identified threshold, 5.5 mm, produced a BAS of 0.883 in the training data using an ensemble approach. This pipeline was then shown to produce plausible results for a real case of a patient where 25% of their mandible was missing after a surgical intervention. The pipeline, however, failed to identify missing points on a more extreme case, where 50% of the mandible was removed.

This is the first time that geometric-learning correspondence models are used to identify missing vertices on organ meshes. A recent related work using a deep-learning approach is the method proposed by Andresen and colleagues [5] where they use a convolutional neural network to jointly perform non-rigid registration and segmentation of non-corresponding anatomy. They applied their proposed network to inter-patient MR scans with artificially added stroke lesions. The reported Jaccard indices[1] ranging between 0.59 and 0.62 highlights the need to find better solutions for identifying missing tissues.

Comparing our method to more traditional approaches within radiotherapy, the work published by Vasquez Osorio and colleagues [6] is relevant. The authors used an adapted version of the thin-plate spline robust-point matching algorithm to register automatically segmented features between livers contoured on CT and MR. They demonstrated that outliers present in both datasets could be identified by using fuzzy correspondences, even for large numbers (such as 55% of input points). Similarly, the work by Myronenko and Song [7] has been shown to generate robust registration results in the presence of noise and missing points, which could then be indirectly used to identify outliers/missing points. However, a common disadvantage is the long computational time.

This pipeline was shown to produce plausible results for a case where 25% of their mandible was missing. However, the pipeline failed in the second, more extreme case, where 50% of the mandible was removed. Several factors can be related to this failure. First, the extreme case was out of distribution for the training dataset (11%-28% of the mandibles removed). Therefore, the 5.5 mm could be a very conservative value for this case, especially as the *cICE* reached 70 mm, see Figure 6. Defining an adaptive thresholds, informed by the difference in areas or volumes between the input meshes, and evaluating the pipeline in other 'softer' organs are areas of future development.

Another limitation was the use of an model originally trained for a different purpose. The correspondence model was trained with complete organs which always consisted of a single connected mesh. Designing a correspondence model that explicitly includes classification of missing points is also a promising avenue of research. In this case, the procedure described in section 2.2.2 could be used as data augmentation strategy. Additionally, using volumetric meshes (composed of tetrahedra) rather than surface meshes (composed of triangles) and incorporating imaging at inference time may also provide extra robustness.

Conclusion We have demonstrated how a correspondence model and a post-processing pipeline can be used to identify missing tissue on surface meshes of structures of the same patient using a previously trained geometric-learning correspondence model. Further work to enable its use on more extreme cases is required. We are paving the way to robustly utilise contour-based registrations for radiotherapy applications, such as cumulative dose estimations in the reirradiation setting.

[1] Overlap-based contour similarity measure similar to the Dice Similarity Coefficient

References

[1] K. Lang, T. Held, E. Meixner, et al. "Frequency of osteoradionecrosis of the lower jaw after radiotherapy of oral cancer patients correlated with dosimetric parameters and other risk factors". *Head amp; face medicine* 18.1 (Feb. 2022), p. 7. DOI: 10.1186/s13005-022-00311-8.

[2] S. Nithiananthan, S. Schafer, D. J. Mirota, et al. "Extra-dimensional Demons: A method for incorporating missing tissue in deformable image registration". *Medical Physics* 39.9 (2012), pp. 5718–5731. DOI: https://doi.org/10.1118/1.4747270.

[3] E. G. A. Henderson, M. van Herk, A. F. Green, et al. "Unsupervised Correspondence with Combined Geometric Learning and Imaging for Radiotherapy Applications". *Shape in Medical Imaging*. Cham: Springer Nature Switzerland, 2023, pp. 75–89. DOI: 10.1007/978-3-031-46914-5_7.

[4] K. H. Brodersen, C. S. Ong, K. E. Stephan, et al. "The Balanced Accuracy and Its Posterior Distribution". *2010 20th International Conference on Pattern Recognition*. 2010, pp. 3121–3124. DOI: 10.1109/ICPR.2010.764.

[5] J. Andresen, T. Kepp, J. Ehrhardt, et al. "Deep learning-based simultaneous registration and unsupervised non-correspondence segmentation of medical images with pathologies". *International Journal of Computer Assisted Radiology and Surgery* 17 (Mar. 2022). DOI: 10.1007/s11548-022-02577-4.

[6] E. Vasquez Osorio, M. Hoogeman, A. Mendez Romero, et al. "Accurate CT/MR vessel-guided nonrigid registration of largely deformed livers". *Medical Physics* 39.5 (2012), pp. 2463–2477. DOI: 10.1118/1.3701779.

[7] A. Myronenko and X. Song. "Point Set Registration: Coherent Point Drift". *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32.12 (2010), pp. 2262–2275. DOI: 10.1109/TPAMI.2010.46.

Fully Automatic Flap Segmentation through Deep Learning Methods to Refine Head and Neck Cancer Postoperative Radiotherapy, a GORTEC study

Juliette Thariat [1,9], Zacharia Mesbah [2], Youssef Chahir [3], Abir Fathallah [4], Vianney Bastit [5], Ingrid Breuskin [6], Arnaud Beddok [7], Alice Blache [8], Jean Bourhis [9], Mathieu Hatt [4] & Romain Modzelewski [2]

1 Department of Radiation Oncology, centre François-Baclesse, France, 2 LITIS - EA4108 - Quantif, Rouen, France, 3 Image Team GREYC-CNRS UMR, Caen, France, 4 LaTIM – INSERM - UMR 1101, France
5 Department of Head and Neck Surgical Oncology, Caen, France, 6 Department of Head and Neck Surgical Oncology - Gustave Roussy, Villejuif, France, 7 Department of Radiation Oncology, Institut Curie, France
8 Department of Radiation Oncology, Amiens, France, 9 Tours, GORTEC.

Abstract Postoperative radiotherapy (poRT) may negatively impact flap outcomes in head and neck cancer (HNC) patients. This study sought to develop a fully automatic method for flap segmentation, to facilitate future attempts at evaluating dose-volume effects on HNC flaps. 148 HNC patients' flaps were delineated by two experts to generate ground-truth segmentation masks. The nnUNet and MedSAM frameworks were used to train flap segmentation. The Dice Similarity Coefficient (DSC), Intersection Over Union (IOU) and Hausdorff Distance metrics were used to evaluate the neural network performances. The nnUNet outperformed MedSAM. It achieved median DSCs of 86.9% in the test set for soft tissue flap segmentation. Features resulting in poorer performances of automatic segmentation were pedicled flaps, moderate-to-severe artifacts (present in 42% of patient poRT CTs), and low-volume centers.

Keywords: Radiotherapy, Head and Neck Neoplasms; Surgical Flaps; Radiation Tolerance; Automatic Segmentation; Neural Networks

I Introduction

The definition of target volumes and organs-at-risk (OAR) in the postoperative setting is complex, the native anatomy being modified by tumor resection, substantial tissue defect and inflammation, edema, fibrosis, surgical clips. Reconstructive surgery with a flap aims at restoring the functional and esthetic impairment of tumor resection, but it adds complexity to the delineation task.

The surgical literature suggests that radiotherapy (poRT) damages flaps. To date, volumetric, textural and functional studies are lacking. Beyond multidisciplinary patient management, a fully automatic flap delineation is a prerequisite to the assessment of the causality of radiation-induced deterioration of flaps. However, flap delineation is a new, challenging and time-consuming task for radiation oncologists. Following guidelines for the delineation of common soft-tissue flaps [1], interobserver variability of flap delineation achieved a Dice Similarity Coefficient (DSC) of 0.70 [2], of similar magnitude as in international consensus studies of tumor and OAR delineation [3].

Our goal was therefore to develop a fully automatic segmentation of soft-tissue flaps, which has never been reported. If successful, this should provide a unique opportunity to assess current irradiation practice and characterize flap dose-volume effects in a future work.

2 Materials and Methods

The imaging dataset consisted of poRT planning DICOM-RT data from a GDPR-compliant GORTEC prospective study (*NCT03576417*). Out of 152 patients (33 centers) who had a flap on poRT planning CT between 2019 and 2021, 4 were excluded due to missing DICOM slices, leaving 148 CT and RT structure sets containing tumor volumes and organs. Ground truth flap contours were created according to flap delineation guidelines [1]. Contouring by two independent experts was blind to patient characteristics and outcomes.

Two deep learning frameworks were used. The nnU-Net ('no-new-Net') is a robust and self-adapting framework based on vanilla (i.e native, unmodified) 2D and 3D U-Nets [4]. It is the standard approach for automated segmentation of medical images. A more recent framework based on transformers, MedSAM "Segment Anything Model" fine-tuned using medical images was also used. This 2D approach requires as input a region of interest containing the object of interest. As the nnU-Net outperformed MedSAM (Supplementary materials) without requiring a user-defined input (i.e., flap region), we focus on the nnU-Net from here onwards. The nnU-Net framework trained an out-of-the-box generic U-Net [5] on our dataset [4], randomly split into training and test sets of 120 and 28 patients. Images were resampled to a voxel size of $1.17\times1.17\times2mm^3$ (median voxel size). Images and Ground Truth masks were resampled using 3rd-order spline and nearest-neighbor interpolation, respectively. Data augmentation consisted of all the nnU-Net random flips, affine transforms (rotation, translation and/or scaling to the volume) elastic deformation (deformation following a grid mesh) of CTs and segmentation maps, brightness/contrast modifications and image quality reduction. Parameters of the nnUnet framework (implementation and training; optimizer, decreasing learning rate, dropout rate, cross-validation setting, hyperparameters, etc) are provided (Suppl materials). Standard metrics were Dice Similarity Coefficient (DSC), Intersection Over Union (IOU), relative/absolute differences in mean dose between automatically & manually delineated flaps.

Importance of clinical and imaging characteristics on flap segmentation

The impact of flaps (x_1, …, x_n), and image characteristics were assessed with respect to the models' performance. The average DSC with the 5-fold models was computed for each patient (training and test sets), and all characteristics were extracted either from the DICOM or by visual Quality Control and computed in a correlation matrix. A XGBoost modeling was fit to these features to determine the function $\hat{f}$ that best approximates f such that:f(x1, …, xn) = DSC with a maximum depth of XGBoost set to 15. A SHAP value graph showed feature importance of the XGBoost model.

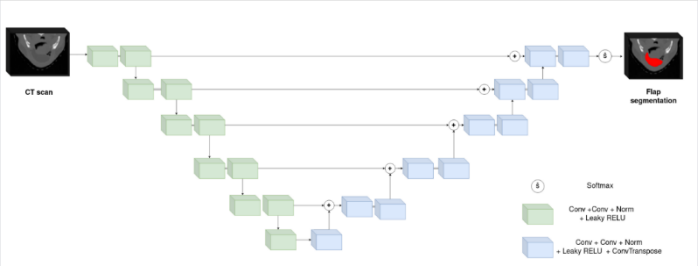


Figure 1: Flap UNet framework

3 Results

3.1. Population and flaps characteristics

Most (69.4%) flaps were in the oral cavity (Table 1). They consisted of soft tissues only (62.4%) with bone resection without (9.4%) or with bone reconstruction (28.2%). Free flaps represented 92.6% of cases, others were pedicled flaps. Median flap volume was 50.5 cm3 (range 3.9-199.8 cm^3). Metallic artifacts due to dental filling/crowns/implants or fixation plates were present in 71.8% of the cases, including 42.3% hampering delineation. Other parameters could vary between patients. A substantial proportion of the flaps were included in the high-dose radiotherapy region: median % of flap volume in the high-risk planning target volume (PTV) was 82.6% with many flaps having most high-risk PTV seemingly centered onto the flap (Figure 2D). Median % of flap volume in the low-risk PTV was 92.7%.

3.2. Performance of deep-learning segmentation

On the test set, the nnU-Net achieved a mean DSC of 74.0% (median 86.9%), mean IoU of 64.1% (median 76.9%) for soft tissue flap segmentation. Mean DSC by medSAM was 63% only, showed irrelevant results (Suppl materials) & was therefore abandoned. This lower performance could be attributed to the 2D nature of the method, & inability to correlate adjacent slice information leading to large flap delineation errors from one slice to the other. The prediction and ground truth of the five flaps that achieved a DSC ≤50% in at least two folds were examined visually. Suboptimal segmentation was either due to under-representation of some flaps (pedicled) or image characteristics (artefacts).

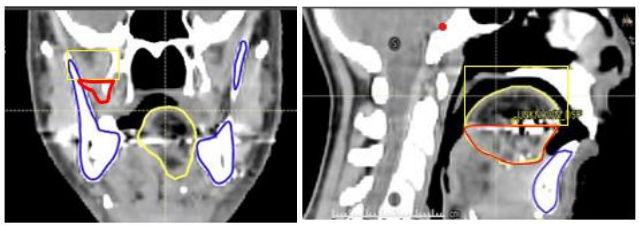


Figure 2: challenging segmentation task; flap (yellow) best seen delineated manually by using coronal (left) and sagittal (right) planes due to dental artefacts in axial planes

3.3. Quantitative impact of clinical or imaging characteristics of automatic segmentation

Table 1 shows the number of patients per center, high (>15), medium (2-14), low (1-2), flap location, whether the flap was pedicled or a free flap, whether bone was resected/not, & with/without bone reconstruction, X-Ray current and exposure (physical characteristics of the acquisition), voxel size & volume (voxel spacing, slice thickness), artifact severity (0 (no artifacts), 1 (slight), 2 (moderated) 3 (severe)), & the DSC of each patient on the 5 fold models. Of note, planning CT were extracted following dosimetry, i.e. Tomotherapy CT had been transformed into proprietary format of lower spatial resolution. The correlation matrix showed correlations between DSC & pedicled or free flap (-0.20) and artifact severity (none/slight versus moderate/severe) (-0.23). The DSC correlated with center volume. The XGBoost model fit to the data of the whole dataset to predict DSC reached a r^2 value of 0.944. SHAP values (Figure 2) indicating the importance of each feature in the prediction of the DSC showed strong negative impact of pedicled flaps, moderate-to-severe artifacts & low volume centers; positive impact of flap location in the oropharynx or oral cavity, high volume centers, & a bone flap on mean DSC of soft tissue flap segmentation.

Characteristics	**Number (%)**
Flap site: oral cavity / oropharynx / larynx - hypopharynx	102 (68.92) / 26 (17.57) / 20 (13.51)
Flap tissues (visual detection): Soft tissue flap without bone resection Soft tissue flap with bone resection wo reconstruction Soft tissue & bone flap (visual)	 92 (62.16) 14 (9.46) 42 (28.38)
Flap type (visual detection): Free flap / Pedicled flap	137 (92.57) / 11 (7.43)
Number of flaps per center : Mean, SD, (range), median	3, 8 (range:0-55), 1
Number of flaps per center: Low volume centers (1-2) / intermediate (3-15) / high (> 15)	20 (60.60) / 11 (33.33) / 2 (0.07)
Proportion of flaps over operated patients: mean, median	0.34, 0.27
Flap volume (cm3) (DICOM RT SS)	60.1, SD 42.7 (range: 3.9-199.8), median 50.7
High-dose PTV volume (cm^3) (DICOM RT)	211.4, SD 171.5 (range: 0-1274.7), median 175.6
Low-dose PTV volume (cm^3) (DICOM RT)	591.2 SD 260.4 (range:0-1163.8), median 591.2
% of flap volume in high-dose PTV (66 Gy) (DICOM fields)	68.02, SD 36.09 (range:0-100), median 82.86
% of flap volume in low-dose PTV (54.45 Gy) (DICOM fields)	72.06, SD 39.23 (range:0-100) median 96.96
Tracheostomy (visual detection)	33 (22.30)
Feeding tube (visual detection)	26 (17.57)
Semi-quantitative aspect of metallic artefacts (visual detection): Absent / Mild / Moderate / Severe	41 (27.70) / 44 (29.73) / 28 (18.92) / 35 (23.65)
Surgical clips in operative bed (visual)	144 (97.30)
Lead marbles on mask (visual)	130 (87.84)
Contrast agent injection (visual detection): None / IV (Arterial Phase) / IV (Portal Venous Phase)	106 (71.62) / 34 (22.97) / 8 (5.41)
CT acquisitions parameters in kV (DICOM tag 0018,0060)	139 (93.92%) at 120 (9 missing Hi-Art CT)
XRay Tube Current (DICOM tag	196, 124 (0-594) median

0018,1151)	165
Exposure in mA (DICOM tag 0018,1152)	132, SD 11 (0-69) median 96
Native slice thickness z (mm) (DICOM tag 0018,1050)	2.31, SD 0.45 (1.25-3), median 2
Native spacing x or y (mm) (DICOM tag 0028,0030)	1.28, SD 0.41 (0.59-2.73) median 1.17
Native Voxel volume (mm^3) (x*y*z)	4.39 SD 4.08 (0.69-22.43) median 2.75

Abbreviations: PTV: planning tumor volume, Gray Gy

Table 1: Patient, flap & image characteristics from DICOM RT data

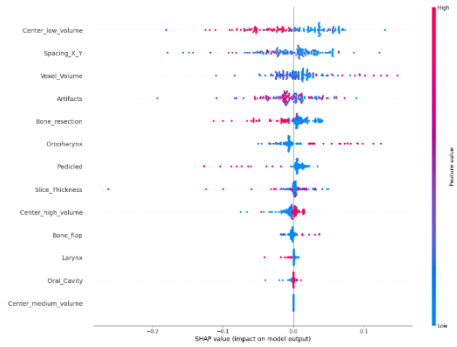


Figure 2: SHAP values of the XGboost DSC model
Each horizontal line represents a feature, each dot on these lines a single patient, & its position along the x-axis corresponds to the importance the feature had in the DSC prediction. Red dots on the right (positive) & blue dots on the left (negative) mean positive or negative impact on the predicted DSC.

4 Discussion

Although flaps are critical to improve functional outcomes, radiation oncologists pay little attention to flaps when planning poRT [6]. Deep mutual understanding of technical aspects of medical/surgical disciplines is critical to patient care beyond collaborative selection of treatment indications. Flap segmentation at the poRT planning step represents a great opportunity to address the effects of radiotherapy on flaps objectively, such as flap shrinkage by fatty atrophy, conversion from muscle to fat, fibrosis, & correlations of changes with dose distribution & functional toxicities. Flap segmentation can also help to improve our understanding of the patterns of failure in the presence of flaps.

Based on a multicentric flap database, we showed that current practice consisted in including the flap (median % of flap volume over 80%) in the high-dose region (i.e. region believed at high-risk for tumor colonization) and ≥96% in the radiotherapy field. Whether this constitutes a cognitive bias or relies on true knowledge of the patterns of failure in the presence of flaps was beyond the scope of this study but is hypothesis-generating.

Compared to a mean DSC of 0.70 for interobserver variability of manual flap delineation by experts in the field [2], our study showed good accuracy of the trained nnUNet method, achieving median and mean DSC of 0.87 and 0.74. Given the highly challenging nature of flaps segmentation compared to other segmentation tasks such as PET/CT tumor delineation in HNC where mean DSC is within 0.77-0.80 for the best algorithms, the accuracy of the flap nnUNET is more than satisfactory.

Our study also showed that dose difference between the automated & manually delineated flaps was too small to be clinically relevant. We will therefore use this fully automatic flap segmentation in our future works to assess segmentation performances, flap radiomics & dose-volume effects (volume, texture) longitudinally based on diagnostic & follow-up images. Finally, our seminal segmentation work could also serve as the basis for future clinical trials.

Limitations to this study include an imbalanced dataset where pedicled flaps represented a minority of the cases, an observation that however well represents current practice. Pedicled flaps are usually reserved to patients with significant comorbidities and who cannot stand prolonged anesthesia. Pedicled flaps were more segmentation error-prone. However, the pedicled is less exposed than the flap body, the latter being next or in the tumor bed. Thus, the fact that our method only segments the body is relevant to radiotherapy practice. We also observed a predominance of oral cavity flaps, which illustrates the critical functions of this anatomic site where substance loss resulting from tumor resection can substantially modify swallowing and speech. It also illustrates the preponderance of surgery in this tumor site, where oropharyngeal and laryngeal tumors can more often undergo definitive chemoradiation. Interestingly, the presence of a bone flap seemed to improve soft tissue flap segmentation despite the fact that bone flaps were underrepresented. We also showed the importance of image quality. Contrary to diagnostic images, RT planning often relied on non-contrast enhanced acquisitions in our dataset (often due to limited resources at the time of planning CT). Voxel size was chosen to fulfill the delineation and computational times requirements of the busy clinics, sometimes at the price of a poorer spatial resolution (often to limit the number of CT slices to delineate and the dose calculation time). In our dataset, it should be noted that for TomoTherapy cases, dosimetry CT were extracted for ground truth and automated delineation, however manual delineation was performed on a native CT in practice; a factor that might have altered the quality of the ground truth. This aspect emphasizes the good performances of our method and its robustness with respect to low spatial resolution and low contrast images. We used a robust deep learning method, which today is considered the state of the art for medical image segmentation. Although new models have been proposed (such as transformers), their advantage in limited size datasets such as in medical studies as in our case, is not yet confirmed, and our results demonstrated no medSAM added value.

Another concern, independently of the current segmentation method, was the suboptimal management of artefacts, despite broad access to minimization algorithms (MAR) softwares in routine CTs. We could not investigate the impact of MAR as use of MAR was not consistently indicated in the DICOM tags. More systematic & comprehensive tagging of DICOM fields of radiotherapy planning CTs should be encouraged to improve quality and standardization of radiotherapy CTs. Artefacts were very frequent & significantly altered both manual & automatic segmentation, leading to missegmented flaps with the automated algorithm and also altered manual delineation although one can rely on the three planes visualization to counteract artefacts dominating in some planes. It should be noted that removing artefacted images from training datasets would limit the generalization of the models. Therefore our choice was deliberate to deal with image quality as encountered in routine practice. By inputting a full head and neck CT scan to our inference pipeline, accurate segmentation was achieved.

Future investigations may explore other strategies, such as multiclass segmentation of OAR & flap, multitask architectures for flap classification, pre-processing of images for reducing artifacts & harmonize image properties, as well as pre-training on larger medical datasets.

Finally, we trained the first deep learning-based segmentation method dedicated to flap segmentation. This has potential major practical implications for poRT planning. We also observed that radiation oncologists included the flap in the high-dose target volumes. Our trained model should be a reliable too for longitudinal clinical research studies on radiotherapy, evaluating radiation-induced contributions to sequelae and tumor relapses, optimizing reconstructive surgery and radiotherapy combinations toward more customized postoperative radiotherapy.

6 Acknowledgments

With financial support from ITMO Cancer of Aviesan within the framework of the 2021-2030 Cancer Control Strategy, on funds administered by INSERM

To all GORTEC investigators and staff

You can also see external validation abstract by Abir Fathallah

References

[1] Le Guevelou J, Bastit V, Marcy PY, Lasne-Cardon A, Guzene L, Gerard M, et al. Flap delineation guidelines in postoperative head and neck radiation therapy for head and neck cancers. Radiother Oncol 2020;151:256–65. https://doi.org/10.1016/j.radonc.2020.08.025.

[2] Beddok A, Guzene L, Coutte A, Thomson D, Yom SS, Calugaru V, et al. International assessment of interobserver reproducibility of flap delineation in head and neck carcinoma. Acta Oncol 2022:1–8. https://doi.org/10.1080/0284186X.2022.2036367.

[3] Guzene L, Beddok A, Nioche C, Modzelewski R, Loiseau C, Salleron J, et al. Assessing Interobserver Variability in the Delineation of Structures in Radiation Oncology: A Systematic Review. Int J Radiat Oncol Biol Phys 2022:S0360-3016(22)03551-9. https://doi.org/10.1016/j.ijrobp.2022.11.021.

[4] Isensee F, Petersen J, Klein A, Zimmerer D, Jaeger PF, Kohl S, et al. nnU-Net: Self-adapting Framework for U-Net-Based Medical Image Segmentation 2018. https://doi.org/10.48550/ARXIV.1809.10486.

[5] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation 2015. https://doi.org/10.48550/ARXIV.1505.04597.

[6] Carsuzaa F, Lapeyre M, Gregoire V, Maingon P, Beddok A, Marcy P-Y, et al. Recommendations for postoperative radiotherapy in head & neck squamous cell carcinoma in the presence of flaps: A GORTEC internationally-reviewed HNCIG-endorsed consensus. Radiother Oncol 2021;160:140–7. https://doi.org/10.1016/j.radonc.2021.04.026.

A rat model of lymphocyte recirculation and the effect of irradiation

Chris Beekman[1], Natalia Carrasco-Rojas[2], Julia Withrow[2], Wesley Bolch[2], and Harald Paganetti[1]

[1]Department of Radiation Oncology, Mass General Hospital / Harvard Medical School, Boston, Massachusetts.
[2]Department of Biomedical Engineering, University of Florida, Gainesville, Florida.

Abstract Purpose: To create a 4-dimensional (4D) rat model of lymphocyte recirculation between the blood and other lymphatic compartments to for dynamic dose accumulation and effect assessment during irradiation. Methods: A graph representation of the lymphatic compartments was created, including 1. major organs divided into a blood and a lymphatic part; 2. the lymphatic network including lymph nodes (LNs) and Peyer's patches. This graph was converted into a physiologically based pharmacokinetic (PBPK) model as both a system of differential equations and a stochastic model. By overlaying the lymphocyte migration kinetics with a 3D model of rat anatomy, the dose could be dynamically accumulated under different irradiation scenarios. Results & Conclusions: Irradiation of blood compartments contributes negligibly to the total dose to the recirculating lymphocyte pool. Radiation-induced depletion of viable lymphocytes in the blood is caused by irradiation of secondary lymphatic organs and their subsequent redistribution among lymphatic compartments.

1 Introduction

Radiation-induced lymphopenia (RIL) has been studied extensively because of its adverse effect on treatment outcome [1]. Since the depletion is measured in the blood, it is often assumed that blood irradiation itself contributes significantly to this phenomenon. Consequently, modeling approaches have thus far been focused on the blood dose [2, 3]. However, it is known that lymphocytes continuously migrate between the blood and other lymphatic compartments. In particular, (naïve) lymphocytes enter lymph nodes (LNs) and Peyer's patches in the gut through high endothelial venules (HEVs). From there, the lymphocytes may migrate through a chain of LNs before returning to the blood via the thoracic duct or the right lymphatic duct. Thus, the routes and kinetics of lymphocyte recirculation depend on the architecture of the lymphatic network. In addition, lymphocytes can be retained in the microvasculature of organs by two different mechanisms. First, because of their size, lymphocytes may need to deform to squeeze through the capillary network [4]. Second, lymphocytes may roll along and attach to endothelial cells in post-capillary venules (PCVs), thus becoming 'marginalized' in the organ microvasculature [5]. Finally, lymphocytes released from open-ended arterioles in the red pulp of the spleen separate from the blood stream and are directed to the white pulp, where they reside for some time before eventually being released back into the blood circulation [6]. For all these reasons, the blood alone provides only a very limited view of the true dynamics of lymphocytes; a much deeper understanding of lymphocyte recirculation kinetics is needed to study causes and effects of the depletion of circulating lymphocytes.

The aim of this study is therefore to create a 4D model of lymphocyte recirculation in which we can track individual "lymphocyte parcels". Overlayed with a radiation dose rate, a lymphocyte dose distribution can be obtained, from which surviving fractions can be calculated. Finally, simulated lymphocyte concentrations in the blood can be monitored and potentially compared with measurements.

The rat provides a good starting point: it has a well-studied lymphatic anatomy and many experimental studies on lymphocyte kinetics have been performed in rat. Moreover, our simulation results can be potentially be validated by carefully designed experiments.

2 Materials and Methods

As a first step in the construction of a 4D model of lymphocyte recirculation in rat, a physiologically based pharmacokinetic (PBPK) model of lymphocyte flow was constructed. In this compartmental model, blood circulation was considered the distributor of lymphocytes through the body. We extended the blood compartments of the lungs, liver, and intestines with a 'lymphatic' compartment, denoting marginalized lymphocytes. The blood compartment of the spleen (red pulp) was similarly equipped with a lymphatic compartment (white pulp) (Fig 1a). We represented this model as a directed graph whose edges represent the anatomical connections between compartments (nodes). This allowed for rapid and insightful architecture modification. A graphical representation of the LN connectivity was also incorporated, guided by anatomical reports from literature [7]. A total of 38 LN compartments were added, each representing either a single LN or a group of LNs (Fig1b). A rate variable k_{ij} was associated to each graph edge specifying the exchange rate from compartment i to j. From the graph we extract the rate matrix $\boldsymbol{K}$ as the adjacency matrix weighted by the rate variables. This rate matrix specifies the kinetics of lymphocyte recirculation.

Both a method based on ordinary differential equations (ODE) and a stochastic method were developed to simulate the kinetics of lymphocyte migration for a given $\boldsymbol{K}$. In the ODE method, the following system is solved:

$$\frac{dy_i}{dt} = \sum_j K_{ji}\, y_j - K_{ij} y_i$$

where y_i is the number of lymphocytes in compartment i. In the stochastic method, a walk over the graph is simulated for a large number of particles with each particle representing some amount of lymphocytes. The rate matrix is converted in a transition matrix $\boldsymbol{T}$, denoting the jumping

probabilities between compartments in an incremental time step Δt:

$$T_{ij} = \left[K_{ij} + I_{ij} \cdot \left(1 - \sum_j K_{ij} \right) \right] \Delta t$$

where $\boldsymbol{I}$ is the identity matrix. The rate matrix determines the mean transit time of each compartment by $MTT_i = 1/\sum_j K_{ij}$. However, where the ODE approach only models exponential transit time distributions, the stochastic approach allows us to choose these distributions ourselves.

Outflux rates from blood compartments were obtained from Brown et al. [8] as the ratio of the outflow and the blood volume: $k_{ij} = Q_{ij}/BV_i$. However, fluxes between lymphatic compartments are not well established. Nevertheless, experimental data on the kinetics of lymphocyte migration in laboratory animals based on the injection of radiolabeled lymphocytes [3, 4] and thoracic duct cannulation studies [11] are available and can be used to fit the unknown exchange rates. Previously, Ganusov & Auerbach fitted migration rates for the rat [12]. We have extended this effort by 1. fitting a PBPK-based model in which blood circulation distributes lymphocytes among vascular compartments and the fitted parameters correspond to measurable quantities; 2. by explicitly incorporating the lymphatic architecture of the rat; 3. by performing the fitting based on Bayesian methods that account for the multiplicative measurement noise. With as many as 38 lymph nodes (LNs) included, we did not fit a separate outflux rate for each, but assumed that both the number of lymphocytes in each LN and the influx from the blood through HEVs were proportional to the LN volumes, which were obtained from literature. With these assumptions we can fit a single parameter γ for all LNs and set:

$$k_{out}(LN_i) = f_i * \gamma$$

Where f_i is a LN-specific factor explicitly calculated from the LN volume and its connectivity in the graph. We fitted the model using the fast ODE representation as this requires many evaluations. The "probabilistic inference on noisy time series" (pints) python package was used, and the stability of the solution space explored using MCMC sampling methods [13]. The fitting was performed with respect to the experimental data by Smith & Ford [9]. Once fitted, the $\boldsymbol{K}$ matrix can be imported into the stochastic representation of the model and appropriate transit time distributions can be selected. We then validated our fit on another experimental study describing the temporal appearance of radiolabeled lymphocytes collected during thoracic duct cannulation [11].

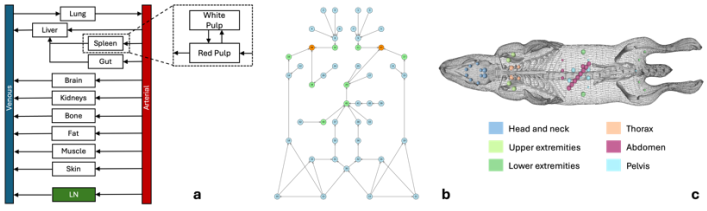


*Figure 1: **a)** PBPK model of lymphocyte migration. **b)** Graphical structure of rat LNs, which is completely contained in the LN box from **a)**. **c)** 3D distribution of LNs in a rat phantom.*

Next, the effect of irradiation was investigated. To determine which LNs are in a given radiation field, their physical location in the rat must be considered. Therefore, again guided by anatomical descriptions [7], the LNs were positioned in a 3D model of a rat phantom (Fig 1c). Additionally, 10 Peyer's patches were uniformly distributed over the surface of the small intestines. Hypothetical treatment plans were then generated for the rat phantom for a variety of sites using an analytical pencil beam method (Fig 2).

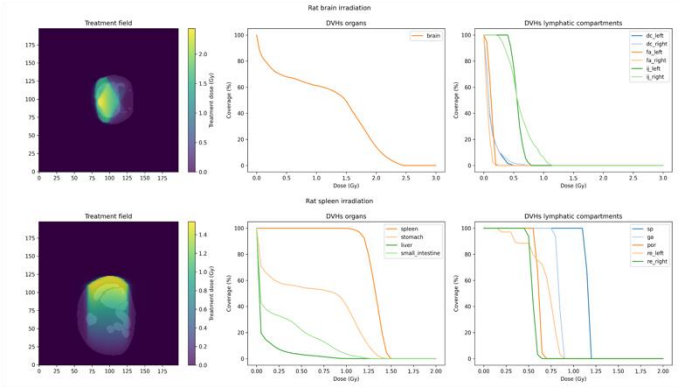


Figure 2: Dose distributions and DVHs for photon irradiation of the rat brain (top) and spleen (bottom).

Using the stochastic model to simulate a large number (500000) of individual 'lymphocyte parcels', each parcel accumulates dose as it travels through the compartments, thereby encountering various dose rate values. Hence, the accumulated dose depends on both the spatiotemporal path of each parcel and the dynamics of the radiation delivery, resulting in a heterogeneous lymphocyte dose distribution. To get an idea of the impact of the radiation on the viability of the lymphocytes, the cell viability was determined probabilistically with a chance $p_k = e^{-\alpha D_k}$, where D_k is the accumulated dose in lymphocyte parcel k during one fraction of the irradiation regimen and α is a radiosensitivity parameter. Then the spatiotemporal redistribution of viable cells after irradiation was determined using the lymphocyte recirculation model.

3 Results

A good fit to the experimental data was found (Fig 3), and different MCMC runs consistently found the same optimum. All secondary lymphatic organs (spleen, (mesenteric) lymph nodes and Peyer's patches accumulate a large amount of lymphocytes, while only a few percent (2.6%) of the total lymphocyte pool circulate in the blood. The marginalized pools of lymphocytes slowly traversing the lung and liver are comparable in number as the lymphocyte pool in the blood.

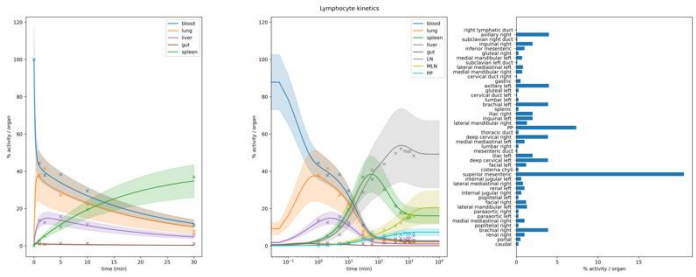


Figure 3: Fitting result for short- and long-term migration kinetics of lymphocytes. On the right, the percentage of recovered activity across different lymph nodes when equilibrium is reached.

The *in-silico* cannulation study provided further validation of our obtained fit. Even though the exact profile of lymphocyte appearance did not precisely match, the peak of lymphocyte recovery is obtained after about 12 hours, consistent with experimental data (Fig 4).

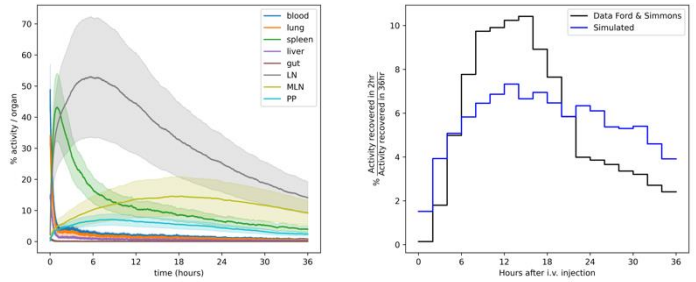


Figure 4: Simulated thoracic duct cannulation. Left: the resulting kinetics of lymphocyte depletion in various compartments. Right: collection of lymphocytes over time compared to experimental data.

We took a closer look at what happens in the main organs playing a role in lymphocyte migration aside from LNs and Peyer's patches. We denote by f the probability that a lymphocyte separates from the blood stream in an organ to become marginalized in the case of non-lymphatic organs, or to cross from the red pulp to the white pulp in the spleen. *MTT* denotes the mean time particles spend in this lymphatic compartment. The values are listed in Table 1. Interestingly, almost all lymphocytes entering the red pulp of the spleen cross over to the white pulp and spend a long time there, indicating the important role of the spleen in immune regulation.

	Lung	Liver	Gut	Spleen
f	0.42	0.44	0.22	0.88
MTT	24 s	59 s	9 s	1.85 h

Table 1: Fraction f of lymphocytes in organ blood going to the corresponding lymphatic component, and the mean transit time (MTT) in that component.

Lymphocyte dose distributions were calculated for 4 sites (brain, right lung, small intestines, and spleen) for hypothetical irradiation scenarios in which radiation was delivered by one or more photon beams delivering a fractional dose of approximately 2 Gy in 20 seconds. Figures 5 and 6 show the resulting lymphocyte dose profiles and DVHs for brain and spleen irradiation, respectively. It is clear that the lymphatic compartments contribute much more to the total lymphocyte dose than the blood compartment, whose contribution is negligible. This is not entirely surprising given the small fraction of recirculating lymphocytes found in the blood at equilibrium. Table 2 lists the expected fraction of lymphocytes still viable after one and 20 fractions of dose.

While most of the radiation damage may not occur in the blood, the decline of viable cells in other lymphatic compartments will, through the process of lymphocyte recirculation, over time become apparent in the blood. In Figure 7, the dynamics of the decline of viable cells after splenic irradiation – initially most dramatic in the spleen itself – was plotted over time. 18 hours after irradiation, a new equilibrium is almost reached, in which the density of viable lymphocytes in each compartment attains its new value, equal to its original value multiplied by the expected fraction of lymphocytes that remain viable after the irradiation.

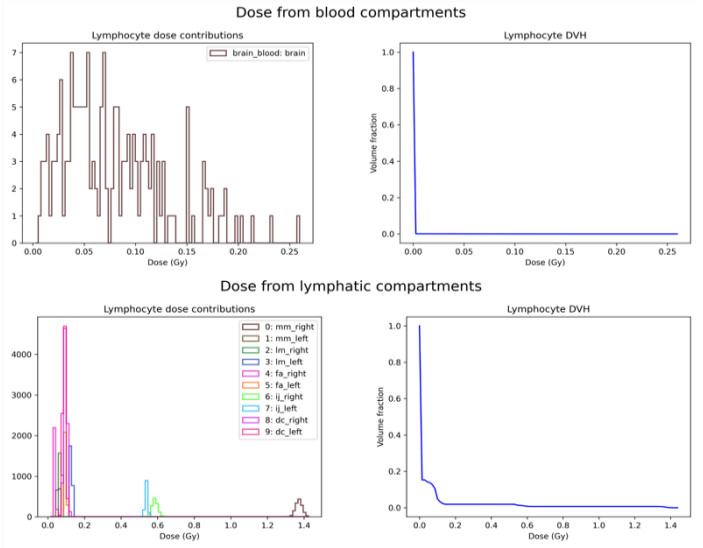


Figure 5: Lymphocyte dose for brain irradiation.

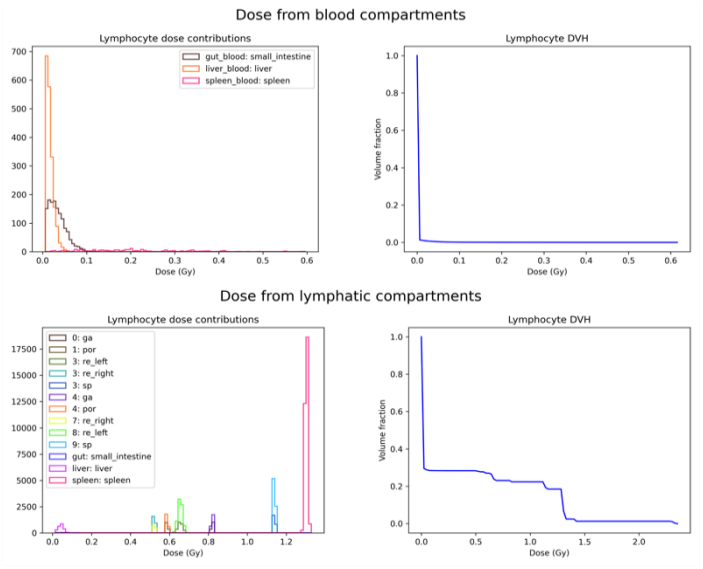


Figure 6: Lymphocyte dose for spleen irradiation.

	Fraction dose (Gy)	Viable fraction	Viable fraction (20×)
Brain	2.0	0.994	0.88
Right lung	1.75	0.988	0.79
Small intestine	1.6	0.954	0.39
Spleen	1.5	0.928	0.22

Table 2: Maximum treatment dose and the expected viable fraction of lymphocytes after one irradiation fraction and after 20. An α of 0.25 Gy^{-1} was assumed.

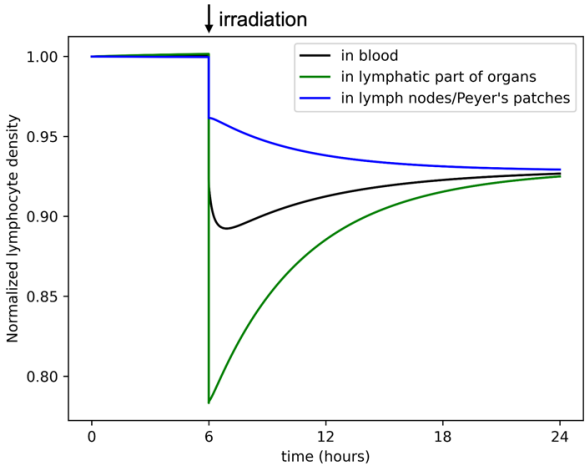


Figure 7: Kinetics of redistribution of viable lymphocytes between compartments remaining after splenic irradiation.

4 Discussion

We have developed a new framework to simulate lymphocyte recirculation in the rat. By combining the kinetics of lymphocyte migration with a 3D model of rat anatomy, including LNs and Peyer's patches, a 4D model of lymphocyte recirculation was obtained that can be used for lymphocyte dose calculations.

Blood dose contributes marginally to overall lymphocyte depletion. Instead, it was consistently observed that irradiation of secondary lymphatic organs such as LNs, Peyer's patches and the spleen resulted in large lymphocyte doses and marked reductions of viable lymphocytes throughout the compartments, including the blood.

Having developed a rat model, we acknowledge that most experimental data on radiation effects are performed in mice, while clinical reality mostly considers humans. While it is highly likely that our findings are qualitatively transferable, it is not obvious how to quantitatively scale the model. While this may be feasible for mice, which are anatomically similar to rats [14], creating a human recirculation model may be more involved.

While we have considered all lymphocyte on equal footing, in reality different subsets may have different recirculation properties. In particular, T cells have been shown to recirculate more rapidly than B cells [10]. Furthermore, antigen-experienced cells do not cross HEVs to enter lymph nodes but rather infiltrate and drain from non-lymphatic tissue [5]. Thus, the simulated pathways mostly apply to lymphocytes with a naïve or central memory phenotype.

5 Conclusion

Our results suggest that blood dose is unlikely to contribute significantly to the lymphocyte depletion often observed after radiotherapy. Rather, the dose to secondary lymphoid organs appears to be the determining factor.

References

[1] B. P. Venkatesulu, S. Mallick, S. H. Lin, and S. Krishnan, "A systematic review of the influence of radiation-induced lymphopenia on survival outcomes in solid tumors," *Crit. Rev. Oncol. Hematol.*, vol. 123, no. July 2017, pp. 42–51, 2018, doi: 10.1016/j.critrevonc.2018.01.003.

[2] S. Yovino, L. Kleinberg, S. A. Grossman, M. Narayanan, and E. Ford, "The etiology of treatment-related lymphopenia in patients with malignant gliomas: Modeling radiation dose to circulating lymphocytes explains clinical observations and suggests methods of modifying the impact of radiation on immune cells," *Cancer Invest.*, vol. 31, no. 2, pp. 140–144, 2013, doi: 10.3109/07357907.2012.762780.

[3] C. Beekman *et al.*, "A stochastic model of blood flow to calculate blood dose during radiotherapy," *Phys. Med. Biol.*, vol. 68, no. 22, p. 225007, Nov. 2023, doi: 10.1088/1361-6560/ad02d6.

[4] G. P. Downey, D. E. Doherty, B. Schwab, E. L. Elson, P. M. Henson, and G. S. Worthen, "Retention of leukocytes in capillaries: Role of cell size and deformability," *J. Appl. Physiol.*, vol. 69, no. 5, pp. 1767–1778, 1990, doi: 10.1152/jappl.1990.69.5.1767.

[5] A. N. Kogan and U. H. von Andrian, "Lymphocyte Trafficking," *Microcirculation*, pp. 449–482, 2008, doi: 10.1016/B978-0-12-374530-9.00012-7.

[6] R. E. Mebius and G. Kraal, "Structure and function of the spleen," *Nat. Rev. Immunol.*, vol. 5, no. 8, pp. 606–616, 2005, doi: 10.1038/nri1669.

[7] N. L. Tilney, "Patterns of lymphatic drainage in the adult laboratory rat.," *J. Anat.*, vol. 109, no. Pt 3, pp. 369–83, 1971.

[8] R. P. Brown, M. D. Delp, S. L. Lindstedt, L. R. Rhomberg, and R. P. Beliles, "Physiological Parameter Values for Physiologically Based Pharmacokinetic Models," *Toxicol. Ind. Health*, vol. 13, no. 4, pp. 407–484, Jul. 1997, doi: 10.1177/074823379701300401.

[9] M. E. Smith and W. L. Ford, "The recirculating lymphocyte pool of the rat: a systematic description of the migratory behaviour of recirculating lymphocytes.," *Immunology*, vol. 49, no. 1, pp. 83–94, 1983.

[10] J. Westermann, Z. Puskas, and R. Pabst, "Blood Transit and Recirculation Kinetics of Lymphocyte Subsets in Normal Rats," *Scand. J. Immunol.*, vol. 28, no. 2, pp. 203–210, 1988, doi: 10.1111/j.1365-3083.1988.tb02432.x.

[11] W. L. Ford and S. J. Simmonds, "THE TEMPO OF LYMPHOCYTE RECIRCULATION FROM BLOOD TO LYMPH IN THE RAT," *Cell Prolif.*, vol. 5, no. 2, pp. 175–189, Mar. 1972, doi: 10.1111/j.1365-2184.1972.tb01014.x.

[12] V. V. Ganusov and J. Auerbach, "Mathematical Modeling Reveals Kinetics of Lymphocyte Recirculation in the Whole Organism," *PLoS Comput. Biol.*, vol. 10, no. 5, 2014, doi: 10.1371/journal.pcbi.1003586.

[13] M. Clerx *et al.*, "Probabilistic Inference on Noisy Time Series (PINTS)," *J. Open Res. Softw.*, vol. 7, pp. 1–6, 2019, doi: 10.5334/jors.252.

[14] Y. Kawashima, M. Sugimura, Y. C. Hwang, and N. Kudo, "THE LYMPH SYSTEM IN MICE," *Jpn. J. Vet. Res.*, vol. 12, pp. 69–78, 1964.

An AI-based approach for modeling the synergy between radiotherapy and immunotherapy

Hao Peng[1,2,#], Casey Moore[3], Yuanyuan Zhang[1], Debabrata Saha[1], Steve Jiang[1,2], Robert Timmerman[1]

1Departments of Radiation Oncology, University of Texas Southwestern Medical Center, Dallas, Texas
2 Medical Artificial Intelligence and Automation Laboratory, University of Texas Southwestern Medical Center, Dallas, Texas
3 Departments of Immunology, University of Texas Southwestern Medical Center, Dallas, Texas

Abstract PULSAR (Personalized, Ultra-fractionated Stereotactic Adaptive Radiotherapy) is designed to administer tumoricidal doses in a pulsed mode with extended intervals, spanning weeks or months. This approach leverages longer intervals to adapt the treatment plan based on tumor changes and enhance immune-modulated effects. We seek to elucidate the potential synergy between combined PULSAR and PD-L1 blockade immunotherapy using experimental data from a Lewis Lung Carcinoma syngeneic murine cancer model. Employing a Long Short-Term Memory (LSTM) Recurrent Neural Network (RNN) model, we simulate the treatment response, treating irradiation and anti-PD-L1 as external stimuli occurring in a temporal sequence. Our findings demonstrate that: 1) The model effectively simulates tumor growth by integrating various parameters such as timing and dose, and 2) The model has potential to provide mechanistic interpretations of a "causal relationship" in combined treatment, offering a novel perspective.

1 Introduction

PULSAR (personalized ultra-fractionated stereotactic adaptive radiotherapy) aims to deliver tumoricidal doses in a pulsed mode with long intervals. It offers several advantages. Longer intervals allow considerably more normal tissue recovery after injury. At the same time, longer interval provides time for the tumor and tumor microenvironment (TME) to undergo changes, and thus allows for more meaningful adaptation based upon the evolving characteristics of the tumor (e.g., the change of tumor size, shape, and biomarker expression). Achieving this in conventional radiation therapy is challenging because not only is the tumor unlikely to undergo significant changes between two closely spaced radiation doses, but repeated dosing over a short interval may impede potential synergies in combined radiation therapy and immunotherapy. Despite promising progress, combination of the two treatments has yielded lackluster results in both clinical and preclinical studies. Whether synergy exists or if the efficacy mainly resides within each modality, remains an intriguing research question.

One aspect of particular interest is determining the optimal timing and sequence to harness the immune-mediated antitumor effects in combined radiation therapy (RT) and immune checkpoint blockade (ICI) therapy (e.g., PD-L1, CTLA4).[1-10] For instance, one clinical trial studied the outcome of multisite Stereotactic Body Radiation Therapy (SBRT) followed by pembrolizumab, reporting an overall response rate of 13.2% in advanced solid tumors.[6] The PEMBRO-RT phase 2 clinical trial in non-small cell lung cancer demonstrated a doubling in overall response when patients were treated with SBRT (3×8 Gy) followed by pembrolizumab.[7] The synergistic impact of combining RT with ICI therapy has been extensively explored in preclinical models as well. One study delivered 10 to 24 Gy in 1 to 3 daily, or every other day fractions and began PD1/PD-L1 checkpoint blockade therapy within a day of radiation.[5] Another preclinical study included 3 daily SBRT fractions of 8 Gy, followed by anti-CTLA4 treatment, beginning on the day of the last fraction.[8] Both studies demonstrated additive benefits when ICI blockade given concomitantly with radiation. Moreover, there are conflicting reports regarding the optimal timing of PD-L1 therapy in relation to radiation. While some studies showed no additive benefit when PD-L1 therapy was administered six days after a single dose of radiation,[9] others suggested clear benefits of PD-L1 therapy given even 21 days after radiation.[10]

Our team has performed an exploration of PULSAR in conjunction with anti-PD-L1 therapy using both immune activated and resistant syngeneic mouse models.[11] Tumor growth was studied as a function of radiation dose and dose scheduling. The results revealed that the effectiveness of anti-PD-L1 therapy depended on how the radiation was sequenced relative to the timing of ICI therapy. For instance, less tumor control was seen with the anti-PD-L1 drug if radiotherapy was given daily or every four days, compared to an interval of ten days. This motivates us to dive deeper into to answer the following question: How do the responses resulting from radiation therapy and ICI therapy interact temporally with each other? The Lewis Lung Carcinoma (LLC) cell line was chosen as the tumor model due to its aggressive nature and immunogenic properties in our study. Acknowledged as a "cold" tumor characterized by low T cell infiltration and high myeloid suppressive cell infiltration, LLC offers valuable insights into the temporal behavior of the adaptive immune response.[12] This is crucial for studying the synergy in vivo, given the longer time required for the adaptive immune response to develop and reach its full effectiveness. In addition, differing from most preclinical models using radiation and with either daily or every other day fractions, we tested radiations pulses with a longer spacing.

Our goal is to develop a modeling tool to better understand the interaction between radiation therapy and anti-PD-L1 therapy, discerning whether synergy exists or if efficacy is confined to each treatment modality independently. Given the complexities of both physical and biological processes involved, it is a very challenging task to model the response of individual components such as tumor growth, radiation damage, immune response, TME and T-cell dynamics.[13-19] Instead of going with a conventional analytical approach (e.g., solving differential equations), we deployed an AI using an LSTM-RNN (**Fig. 1**). RNN is a generalization of feed-forward neural networks that have an internal memory, offering unique strengths in identifying correlation in sequential signals.[20-23] Radiation and anti-PD-L1 were viewed as two external stimulation pulses occurring in a temporal sequence. To isolate the role of radiation, we calculate the difference map, which disentangles the interplay between tumor growth and radiation, highlighting the sole contribution of radiation. AI model may offer unique mechanistic interpretations regarding a "causal relationship", from a completely novel perspective.

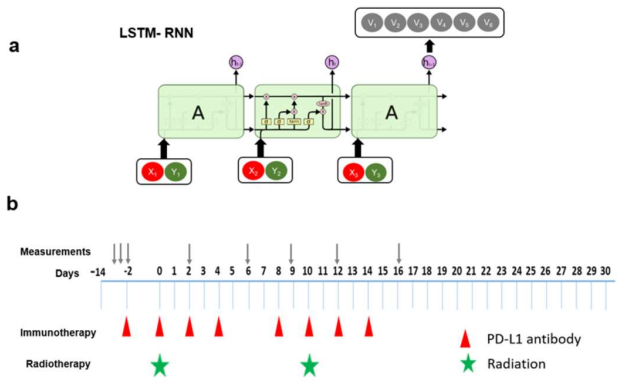


Fig. 1. a. The LSTM-RNN (recurrent neural network) model. b. The timing diagram of the combined therapy in our study. For immunotherapy, either anti-PD-L1 or isotype control was administered. For radiation, different doses were delivered (10, 15, 20, 40 Gy). The first pulse of radiation was always delivered 14 days after the implantation. Tumor volume measurements were carried out sequentially on certain days.

2 Materials and Methods

Data pre-processing. For each treatment group, seven or eight animals were studied, and the measurement of tumor volume exhibited large variance **(Fig. 2**). From a machine learning standpoint, this diminishes the discernible correlation, if any, between an input sequence and an output sequence. For training the LSTM-RNN model, data augmentation was necessary. The details of data processing are summarized below. The total time course was set to be 21 days (three weeks) for all samples to ensure data consistency.

We had 26 treatment groups in total. Each group included the volume change trajectories of seven or eight animals. We initially computed the mean and standard deviation using the tumor volumes of seven or eight animals on a specific measurement day. Assuming a Gaussian distribution for the average response, we then generated 50 samples randomly from this distribution. To address the issue of large error bars (**Fig. 2**), we intentionally reduced the standard deviation by a factor of five, which we believe might be achieved by employing more accurate volume measuring techniques. We acknowledge that the described data augmentation lacks statistical rigor but is the only feasible approach given the available data. As a side note, we consider the inclusion of 50 samples through data augmentation as a strategy to mitigate overfitting and enhance generalization.

A total of 1300 samples were used for machine learning (26 groups and 50 samples per group). Each sample consisted of a sequence of 28 steps with two inputs and one output. The two inputs corresponded to radiation dose and anti-PD-L1, respectively. For radiation, a single value encoded the dose, representing the killing effect using an LQ model which is dose dependent.[13] Otherwise, it was set to zero. In the case of anti-PD-L1, the input was coded as 0 or 1, indicating the absence or presence, respectively. These two inputs were concatenated at each step. The output represented the change in tumor volume between two consecutive measurements. For example, if we conducted six measurements for each animal, there would be five non-zero values. Positive change signifies volume increase, while negative change signifies volume decrease. This choice is primarily attributed to the design of our animal experiments preceding the consideration of AI modeling.

RNN modeling and training details. We developed a LSTM-RNN model using the Pytorch package, consisting of 100 hidden units at each time step and a fully connected layer in the last time step to predict volume change. The model used a "many to one" rather than a "many to many" structure. The LSTM-RNN used recurrent connections to keep track of the context hidden in the sequence (i.e. internal memory). The ADAM algorithm with default settings was used for optimization (learning rate = 0.001, beta1 = 0.9, beta2 = 0.999, epsilon = 1×10^{-8}).[24] Training was implemented on a PC with a NVIDIA GeForce RTX 3090 and an Intel(R) Core(TM) i7-10700K CPU. The batch size was 100 and the total epoch iteration was 5000. The loss function defined in Eqs. 1 and 2 was calculated based on the non-zero points (e.g. with measurements performed) only. It represents the discrepancy between the predicted volume change ($\Delta VC'_i$) and experimental counterpart (ΔVC_i), where i is the index of measurement. ΔVC could be either positive or negative.

$$\Delta VC_i = Volume_{i+1} - Volume_i \qquad (i=1, 2, 3, 4, 5) \qquad (1)$$

$$Loss = \sum_i (\Delta VC_i - \Delta VC'_i)^2 \qquad (2)$$

In total, 1200 samples (24 groups, 50 samples in each group) were used for training (80%) and 100 samples for testing (20%). We performed cross-validation using the following approach: two groups were randomly selected and excluded from the training dataset, and this process was repeated five times to train five models. Subsequently, these five models were applied to a new sequence to assess their robustness. This is different not a holdout validation, as no measurement data available for the new sequence. In our trial, no significant differences were observed among the five models, and we selected only one for subsequent analyses. It is important to note that our primary focus is not on quantitative aspects, given the inherent challenges in achieving both precision and accuracy.

Interpretation of hidden units and predictive modeling. To gain deeper insights into the internal signal flow within the LSTM-RNN model, specifically regarding the "causal relationship", we extract the values of 100 hidden units for 18 time steps. By performing arithmetic operations on the maps of hidden units between two groups (e.g., with or without anti-PD-L1, with or without radiation), we attempted to separate the respective effects of each treatment and provide plausible explanations. Additionally, we proceeded to assess the model's predictive power, aiming to estimate the outcomes of new delivery sequences not included in the experiments. To determine overall tumor control, analogous to the endpoint in clinical studies, we utilized the accumulated volume change (AVC) as the criterion (Eq. 3). We tested three cases: case 1 (two pulses of 10 Gy with a 10-day interval), case 2 (two pulses of 20 Gy with a 10-day interval), and case 3 (a single pulse of 40 Gy), with a cycle of anti-PD-L1 on four or two consecutive days. We analyzed AVCs at various offsets ranging between 1 and 14 days. For example, an offset of five days implies that five days after the delivery of the first radiation pulse, anti-PD-L1 is administered on four consecutive days. While there are numerous permutations to explore, we provided only two cases above to exemplify the potential role of the LST-RNN model.

$$AVC = \sum_{i=1}^{5} \Delta VC'_i \qquad (3)$$

3 Results

Experimental stud. To exemplify the influence of radiation schedules on tumor control in combined treatment, we summarize the experimental findings of a subset of groups in **Fig. 2**. In the absence of radiation, there is no significant difference observed between the group 1 (without anti-PD-L1) and group 2 (anti-PD-L1) (**Fig. 2a**). Similarly, when comparing group 6 (a single pulse of 40 Gy, anti-PD-L1) and group 8 (two pulses of 20 Gy with a 10-day interval, anti-PD-L1) (**Fig. 2b**), both effectively suppress tumor growth. The tumor control between 40 Gy and 20 Gy, appears comparable in the presence of relatively large error bars, as indicated by the data points following the first pulse on day 14. In **Fig. 2c** where a single pulse of 20 Gy was delivered on day 14, there is a noticeable benefit in tumor control for group 4 (anti-PD-L1) compared to group 3 (without anti-PD-L1), which becomes apparent two days after the first pulse. In **Fig. 2d**, when two radiation pulses (20 Gy with a 10-day interval) are delivered, we observe a general trend similar to **Fig. 2c**. The difference becomes more pronounced after the second pulse, between group 7 (without anti-PD-L1) and group 8 (with anti-PD-L1).

AI-based modeling and prediction. **Fig. 3a** illustrates the convergence of both the training and validation loss of the LSTM-RNN model after approximately 2000 iterations. The validation loss exhibits a minor

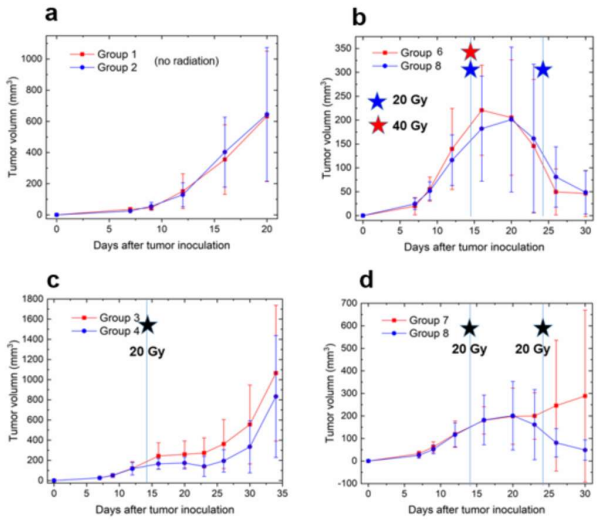


Fig. 2. Illustration of tumor volume change under different delivery schemes. The vertical line on day 14 represents the first radiation pulse.

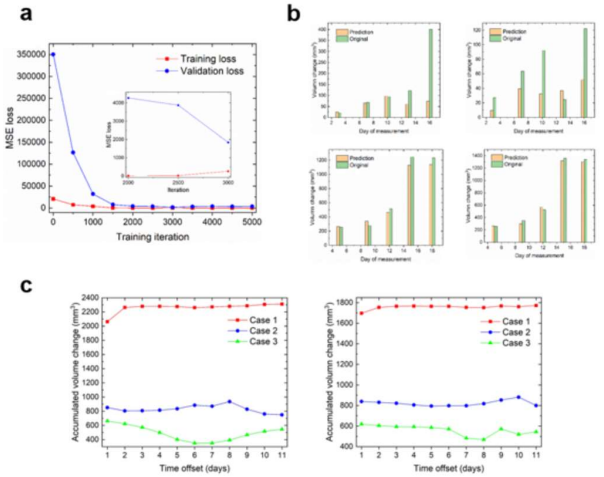


Fig. 3. a. Training-loss and validation-loss curves in terms of mean-squared error as a function of total iteration. b. The results of validation for two cases, each case having two groups not included for training (top: groups 12 and 13, bottom: groups 11 and 14). c. The accumulated volume change (AVC) for new schemes not included in the experimental study.

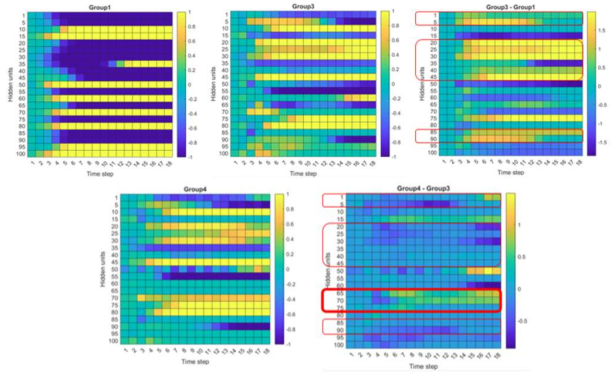


Fig. 4. Analyses of hidden units inside the LSTM-RNN model, which has 100 hidden units at each step and 18 steps in total. The first time step on the horizontal axis corresponds to the day when the first anti-PD-L1 was administered.

deviation from the training loss, suggesting the absence of significant overfitting. The results of the testing are displayed in **Fig. 3b**, showcasing an overall favorable prediction accuracy concerning volume change. The majority of differences at each time point reside within the range of -100 to 100 mm^3, except for a single point in group 11, which surpasses 300 mm^3 on day 16. Considering that the tumor volume in the control group exceeds 600 mm^3 around day 20 (**Fig. 2a**), the predictive accuracy offered by the LSTM-RNN model is deemed satisfactory.

To illustrate the potential use of the LSTM-RNN model for in-silico analysis, the response of AVC for different treatment schemes is depicted in **Fig. 3c** These schemes involve either two or four consecutive daily administrations of anti-PD-L1, slightly different from the every other day administration in the experiments. In the case of a single pulse of 40 Gy (case 3), optimal tumor control is achieved with an offset of seven to eight days. Four doses of anti-PD-L1 result in slightly improved tumor control compared to two doses, with a minimum volume of 350 mm^3 (left) versus 468 mm^3 (right). By contrast, for both case 1 and case 2 (two pulses of either 10 Gy or 20 Gy with a 10-day interval), AVC remains relatively stable, except for an abrupt increase observed in case 1 on the left end. This suggests that the synergistic therapeutic effect is anticipated to be similar, irrespective of when anti-PD-L1 is administered after the first radiation pulse. Another surprising observation in case 1 is the noticeable difference in AVC between two and four anti-PD-L1 administrations, with volumes of approximately 1750 mm^3 (right panel) and 2250 mm^3 (left panel), respectively. In other words, for a radiation pulse of relatively low dose, such as 10 Gy, doubling the dose of anti-PD-L1 adversely affects the treatment outcome, possibly due to treatment toxicities or other mechanisms. The reason for the two aforementioned patterns needs additional exploration, which could be attributed to the real interaction between radiation and immune response, or to the limited generalization capability of the LSTM-RNN model.

Analysis of hidden units and interpretation. To gain further insights into the "causal relationship", we extracted the values of 100 hidden units for each of the 18 time steps and performed arithmetic operations. For instance, we examined the difference between group 1 and group 3 in **Fig. 4** (top), where both groups received the isotype control, but only group 3 received a single pulse of 20 Gy on the third time step. Therefore, in group 1, the temporal evolution of the map of the hidden units primarily reflects the process of tumor growth, while the map of group 3 capture the processes of both tumor growth and the influence of radiation. Therefore, the difference map between the two groups, denoted as Group 3 - Group 1, disentangles the interplay between tumor growth and radiation, highlighting the sole contribution of radiation. We speculate that the presence of more positive values within the highlighted regions yields more effective tumor control (e.g., vertical index: 1-5, 20-45, 85-90, three highlighted boxes).

The impact of anti-PD-L1 is demonstrated in **Fig. 4** (bottom), by comparing group 3 (isotype control every other day) and group 4 (anti-PD-L1 every other day). Both groups received a single pulse of 20 Gy on the 3rd time step. The difference between the two reveals the effect solely attributable to anti-PD-L1, as manifested in a newly highlighted region (vertical index: 65-75). This region contains predominantly positive values and does not overlap with the three previously highlighted regions linked to radiotherapy. While the selection of a region lacks solid evidence and requires additional validation, the results are somewhat expected, considering the distinct mechanisms through which radiation and anti-PD-L1 operate. Moreover, inside this region, we see the effect begins to emerge approximately seven steps after the first pulse and continues throughout the treatment.

4 Discussion

The LSTM-RNN model has demonstrated its unique strength in modeling the process of tumor control in combined therapy up to three weeks. It capitalizes on its inherent ability to process sequential information, which makes it well suited for achieving our objective (**Fig. 1a**). Its recurrent computation property is advantageous as it prevents the growth of the number of model parameters as the number of time steps increases, particularly beneficial when dealing with small datasets.

Each group in our study corresponds to a specific input configuration, consisting of radiation dose and anti-PD-L1 drug dose. In essence, the LSTM-RNN model attempts to find the mapping between inputs and outputs for all 26 groups, which resembles a challenging curve fitting task considering the intricate and unknown mechanisms of interaction involved. Two additional challenges we encountered are: 1) the presence of many zeros in the input sequence (e.g., representing days without either radiation or anti-PD-L1 treatment), and 2) the absence of tumor volume measurement on multiple days. As a matter of fact, we conducted comparative tests and found that a feed forward network (FNN) performed much less effectively than the LSTM-RNN model. We also tested a "many-to-many" structure aiming to predict responses for all time points, including those with missing measurements. However, the results of this approach are inferior compared to the "many-to-one" structure employed in this manuscript. Such disparity is attributed to the increased difficulty for an AI model in establishing temporal correlations when a sequence contains numerous missing data points.

Another novel of our approach is the utilization of hidden units to identify temporal correlations between radiation and anti-PD-L1 (**Fig. 4**). This allows us to gain insights into the "causal relationship" from a completely new perspective. However, the biological interpretations drawn from observing the dynamics of hidden units in our work should be approached with caution. The latent representations in our model, as well as selecting different regions, are very abstract and require additional evidence to support our speculations. In this regard, we would like to reemphasize that our current goal is to identify the general trend in tumor control and potential synergy in a qualitative manner, rather than focusing on quantitative accuracy. Further quantitative analysis and interpretation may need to be postponed until additional biological correlates become available.

Additionally, the LSTM-RNN model may support in silico studies (**Fig. 3c**), generating essential data to address other relevant research questions in the field of combined PULSAR and ICI therapy. Various factors can be combined as inputs, such as dose per pulse, pulse structure, anti- PD-L1 dose, and relative timing. Moving forward, we will conduct a more in-depth exploration of various permutations to identify the optimum treatment strategies for maximizing the synergy between PULSAR and ICI therapy, in parallel with the design of the next phase of experiments. In our view, daily fractions are very unlikely to be the optimal choice.

Ordinary Differential Equation (ODE)-based mathematical models are extensively used for analyzing complex biological systems. These models include components such as tumor cells, immune effector cells, antigens, regulatory cells, antibodies, and natural killer cells. For instance, Eftimie et al. developed models focusing on tumor cells and immune effector cells, resembling a predator-prey model.[25] Serre et al. utilized discrete-time equations to model the synergy between immunotherapy and radiotherapy.[26] In the future, we plan to adopt a similar approach and compare the performance between the ODE-based and AI models. The two may complement each other and enhance the interpretation. Additionally, in our current study, the AI model focuses on the mean output of each group, without accounting for inter-animal variation. A broader question arises regarding the level of personalization in combined PULSAR and immunotherapy. In the long run, it might also be possible for us to take into account different temporal responses and identify optimum treatment strategy for each animal through reinforcement learning, similar to two previous studies.[27, 28]

5 Conclusion

We have successfully demonstrated that the proposed LSTM-RNN model exhibits unique strengths in predicting tumor growth and identifying potential "causal relationships", based on the results of preclinical experiments. Our study introduces a novel perspective on studying the temporal interaction between radiation and ICI therapy. When utilized as an in-silico modeling tool, the LSTM-RNN model has the potential to maximize therapeutic effects through the optimization of variables such as pulse interval, radiation dose, drug dose, and timing. By implementing more rigorous experiments and dedicating future efforts to biomarker identification, we can deepen our comprehension of biological mechanisms and improve quantitative capabilities.

References

1. Sharabi AB, Tran PT, Lim M, Drake CG, Deweese TL. Stereotactic radiation therapy combined with immunotherapy: Augmenting the role of radiation in local and systemic treatment. Oncology (Williston Park) 2015;29:331-340.
2. Gutiontov SI, Pitroda SP, Chmura SJ, Arina A, Weichselbaum RR. Cytoreduction and the optimization of immune checkpoint inhibition with radiation therapy. Int J Radiat Oncol Biol Phys 2020;108:17-26.
3. Jagodinsky JC, Harari PM, Morris ZS. The promise of combining radiation therapy with immunotherapy. Int J Radiat Oncol Biol Phys 2020;108:6-16.
4. Zhu, S., Zhang, T., Zheng, L., et al. Combination strategies to maximize the benefits of cancer immunotherapy. Journal of Hematology & Oncology 2021;14, 156.
5. Gong J, Le TQ, Massarelli E, Hendifar AE, Tuli R. Radiation therapy and PD-1/PD-L1 blockade: The clinical development of an evolving anticancer combination. J Immunother Cancer 2018;6:46.
6. Luke JJ, Lemons JM, Karrison TG, et al. Safety and clinical activity of pembrolizumab and multisite stereotactic body radiotherapy in patients with advanced solid tumors. J Clin Oncol 2018;36:1611-1618.
7. Theelen W, Peulen HMU, Lalezari F, et al. Effect of pembrolizumab after stereotactic body radiotherapy vs pembrolizumab alone on tumor response in patients with advanced non-small cell lung cancer: Results of the PEMBRO-RT phase 2 randomized clinical trial. JAMA Oncol 2019;5:1276-1282.
8. Vanpouille-Box C, Alard A, Aryankalayil MJ, et al. DNA exonuclease Trex1 regulates radiotherapy-induced tumour immunogenicity. Nat Commun 2017;8:15618.
9. Azad A, Yin Lim S, D'Costa Z, et al. PD-L1 blockade enhances response of pancreatic ductal adenocarcinoma to radiotherapy. EMBO Mol Med 2017;9:167-180.
10. Liang H, Deng L, Chmura S, et al. Radiation-induced equilibrium is a balance between tumor cell proliferation and T cell-mediated killing. J Immunol 2013;190:5874-5881.
11. Moore C, Hsu CC, Chen WM, et al. Personalized Ultrafractionated Stereotactic Adaptive Radiotherapy (PULSAR) in Preclinical Models Enhances Single-Agent Immune Checkpoint Blockade. *Int J Radiat Oncol Biol Phys*. 2021;110:1306-1316.
12. Lin H, Wei S, Hurt EM, et al. Host expression of PD-L1 determines efficacy of PD-L1 pathway blockade-mediated tumor regression. J Clin Invest 2018;128:1708.
13. Fowler JF. The linear-quadratic formula and progress in fractionated radiotherapy. Br J Radiol. 1989;62:679-694.
14. Deng L, Liang H, Xu M, et al. STING-dependent cytosolic DNA sensing promotes radiation-induced type I interferon-dependent antitumor immunity in immunogenic tumors. Immunity 2014;41:843-852.
15. Qu Y, Zhang B, Liu S, Zhang A, Wu T, Zhao Y. 2-Gy whole-body irradiation significantly alters the balance of CD4+ CD25- T effector cells and CD4+ CD25+ Foxp3+ T regulatory cells in mice. Cell Mol Immunol. 2010;7:419-427.
16. Kachikwu EL, Iwamoto KS, Liao YP, DeMarco JJ, Agazaryan N, Economou JS, McBride WH, Schaue D. Radiation enhances regulatory T cell representation. Int J Radiat Oncol Biol Phys. 2011;81:1128-1135.
17. Liu R, Xiong S, Zhang L, Chu Y. Enhancement of antitumor immunity by low-dose total body irradiationis associated with selectively decreasing the proportion and number of T regulatory cells. Cell Mol Immunol. 2010;7:157-162.
18. Gough, M. J., & Crittenden, M. R. (2022). The paradox of radiation and T cells in tumors. Neoplasia (New York, N.Y.), 31, 100808. https://doi.org/10.1016/j.neo.2022.100808
19. Arina A, Beckett M, Fernandez C, et al. Tumor-reprogrammed resident T cells resist radiation to control tumors. Nat Commun 2019;10:3959.
20. Bengio Y, Patrice S and Paolo F 1994 Learning long-term dependencies with gradient descent is difficult IEEE Trans. Neural Netw. 5 157-66
21. Chuang Liu, Zhongxing Li, Wenbin Hu, Lei Xing, Hao Peng. Range and dose verification in proton therapy using proton-induced positron emitters and recurrent neural networks (RNNs) Phys Med Biol. doi: 10.1088/1361-6560/ab3564, 2019.
22. Yao S, Hu Z, Zhang X, Lou E, Liang Z, Wang Y, Peng H. Feasibility study of range verification based on proton-induced acoustic signals and recurrent neural network. Phys Med Biol. 2020; 65(21):215017.
23. Hu Z, Li G, Zhang X, Ye K, Lu J, Peng H. A machine learning framework with anatomical prior for online dose verification using positron emitters and PET in proton therapy. Phys Med Biol. 2020; 14;65(18):185003.
24. D. Kingma and J. Ba, Adam: A method for stochastic optimization, 2014.
25. Eftimie, R., Bramson, J. L., Earn, D. J. Interactions between the immune system and cancer: a brief review of non-spatial mathematical models. Bulletin of Mathematical Biology, 2011; 73(1), 2-32.
26. Serre R, Benzekry S, Padovani L, et al. Mathematical Modeling of Cancer Immunotherapy and Its Synergy with Radiotherapy. Cancer Res. 2016; 76(17):4931-40.
27. Tortora, Matteo, et al. Deep reinforcement learning for fractionated radiotherapy in non-small cell lung carcinoma. Artificial Intelligence in Medicine 2021;119: 102137.
28. Tseng, Huan-Hsin, et al. Deep reinforcement learning for automated radiation adaptation in lung cancer. Medical Physics 2017; 44: 6690-6705.

Integration of white matter tracts in stereotactic brain radiotherapy treatment planning with a fully automated atlas-based pipeline

Sophie Bockel[1, 2], Mohammed El Aichi[1, 2], Roger Sun[1, 2], Nathan Benzazon[1, 2], Alexandre Carré[1, 2], François de Kermenguy[1, 2], Eric Deutsch[1, 2] and Charlotte Robert[1, 2]

[1]U1030, ImmunoRadAI, Université Paris-Saclay, Institut Gustave Roussy, Inserm, Villejuif, France
[2]Department of Radiation Oncology, Gustave Roussy, Villejuif, France

Abstract

This study introduces a Python-scripted pipeline for integrating white matter tractography atlases into stereotactic brain radiotherapy (SBRT) treatment planning. The method utilizes a probabilistic atlas of 52 white matter fibers, offering an accessible and automated approach to visualize and analyze white matter tracts in individual patients undergoing SBRT. The pipeline involves image preprocessing, affine then non linear registration to a common space, and dose map transformation with different interpolation methods. The results demonstrate accurate registration and suggest that bSpline interpolation provides optimal consistency in dose distribution. The study aims to derive dose constraints for white matter tracts, offering a potential tool for guiding therapeutic decisions in SBRT. The pipeline enhances the accessibility of white matter structural connectivity in radiotherapy planning.

1 Introduction

Understanding the white matter (WM) anatomy has been facilitated by recent advances in neuro-imaging, such as Magnetic Resonance Diffusion Tensor Imaging (DTI). Complementarily, anatomo-functional atlases of the WM tracts (WMT) and their associated functions have been elaborated based on awake surgery thanks to brain mapping through direct electrical stimulation [1,2]. All these major achievements underscore the crucial importance of preserving WMT in order to maintain patients' neurocognition and quality of life.

At the same time, stereotactic brain radiotherapy (SBRT) has dramatically changed the outcome of brain metastasis (BM) patients. While focal control and extended progression-free survival have been achieved, decreasing neurocognitive side effects of these patients is of major importance [3,4].

In clinical routine, WM structural connectivity is not considered in SBRT treatment planning of BM. This can be explained by the lack of accessible resources to facilitate such an approach.

Recently, Yeh et al. provided in open access a probabilistic atlas of 52 white matter fibers derived from tractographic data of 1065 healthy subjects. This atlas is normalized to the ICBM 2009a non-linear asymmetric space[5].

Here we introduce an automated method that is easily accessible, and straightforward, enabling the visualization of WM tracts provided by Yeh et al. probabilistic atlas on individual patient pre-SBRT imaging (MRI and CT scan) and planning dose maps. Ultimately, the methodology will enable WMT dose-volume histograms (DVH) to be calculated, making it possible to implement tract sparing dosimetry.

2 Materials and Methods

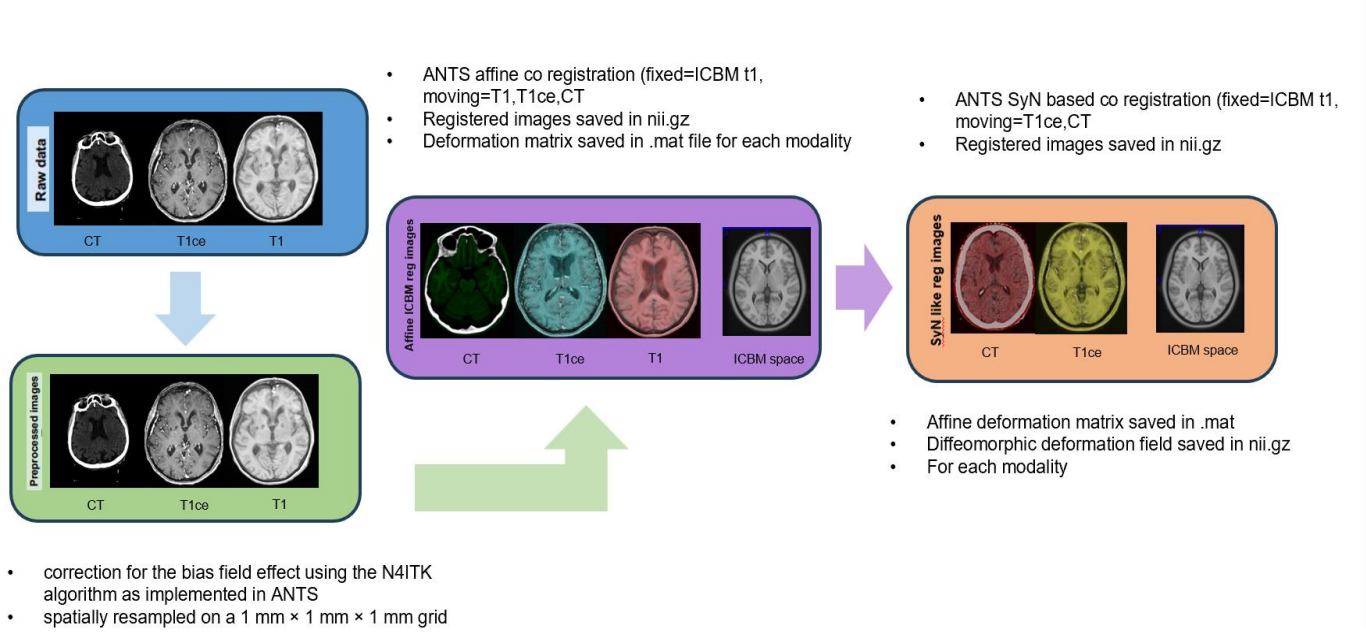


Figure 1. Co registration to ICBM2009a nonlinear asymmetric space: affine then non linear SyN registration.

We focused on BM patients treated for small (< 2cm) lesion(s) with one 20Gy fraction SBRT in our institution.
All pre-SBRT brain images (T1 gadolinium enhanced MRI and CT scan) were preprocessed as follow: DICOM to NIFTI conversion, correction for the bias field effect using the N4ITK algorithm as implemented in Advanced Normalizations Tools for Python (ANTSPy) ecosystem [6] for MRIs, resampling on a 1 mm × 1 mm × 1 mm grid and brain extraction (HD-BET tool for MRIs and CTseg tool by SPM for CTs). Using ANTSPy, an affine registration to the ICBM 2009a nonlinear asymmetric space (T1 ponderation) was performed, follow by a nonlinear registration (Symmetric diffeomorphic image registration (SyN)) (Figure 1).

For both MR and CT, Mutual Information was used as similarity metric for the affine and SyN co-registrations . As opposed to the affine stage, skull stripped images were used for the SyN step for moving and fixed images. To improve the accuracy of the SyN deformable registration, especially on cortical sulci and ventricles, brain cortex masks were auto segmented from the affine output images (CT or MR), corresponding to WM + grey matter (ANTS Atropos [6] for MRI and SPM CTseg for CT scan [7]), and were included in the parameters of the ANTS registration function. Cortical masks were then produced by auto-segmentation on the CT and MRI images obtained after affine + SyN transformations.
The anatomic accuracies of the final registrations (MR to ICBM space and CT ICBM space) were evaluated by computing DICE metric from cortical auto segmentations. We also computed DICEs after affine registration step only to verify the anatomical advantage of the combined final registration (affine + SyN).

With the goal to compute doses received by the WMT, transformation matrix from CT affine registration and CT deformation field from nonlinear registration were applied to the dose maps considering four different interpolators (bSpline, linear, gaussian and lanczosWindowedSinc) for both affine and SyN part, resulting in 16 different combinations of dose warping.
As well, the transformation matrix were applied to RT structures segments with GenericLabel interpolator.
Finally, to evaluate the consistency of warped dose maps and select the best interpolator combination, clinically relevant DVH metrics, such as organs at risk (OAR)(e.g., brainstem, brain, brain-PTV, cochlea, optic nerves, chiasm) mean dose and dose delivered to at least 0.035cc of the organ (D0.035cc) and PTV mean dose and dose delivered to 95% of the PTV (D95) were compared between initial dose map and the 16 final transformed dose maps depending on the interpolator type. DVH extraction was performed using PlatiPy python library.
Comparison of clinically relevant DVH metrics from initial dose maps and warped dose maps were made using boxplot representation of dose distribution.

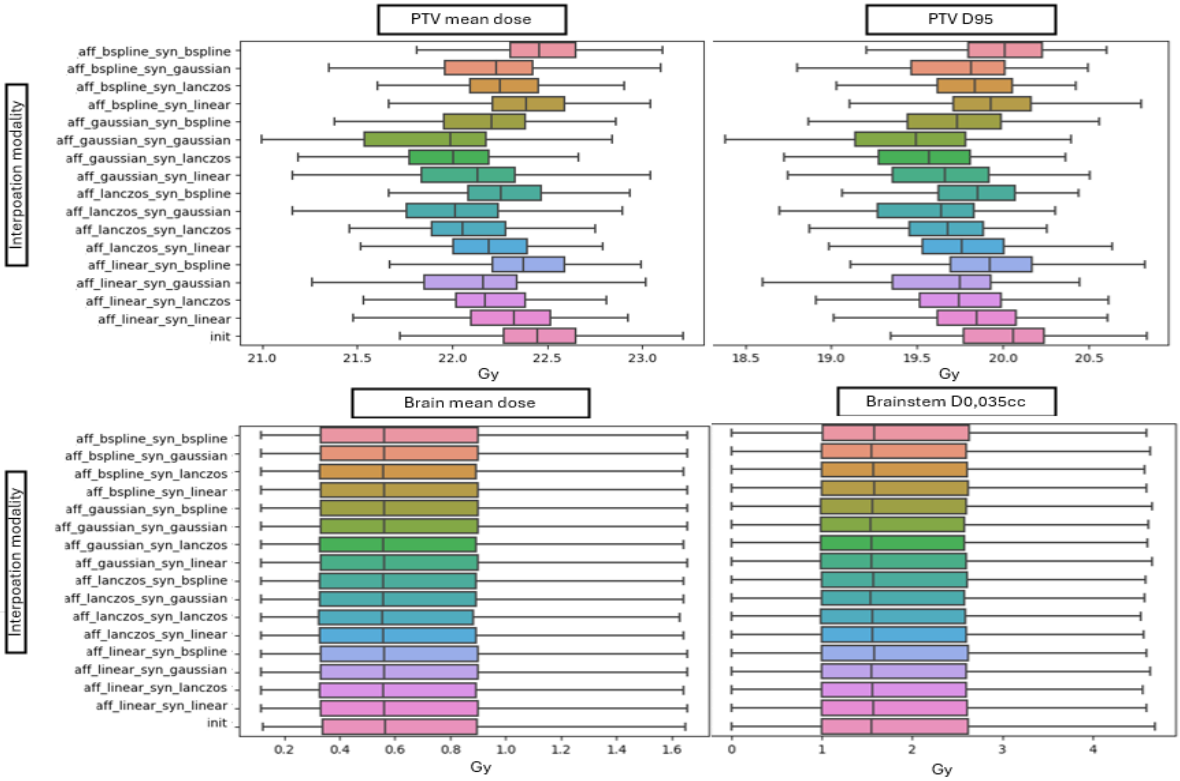


Figure 2. Boxplots of several DVH metric value distribution depending on the dose warping modality.

Init = initial dose map ; 16 dose warping modalities : bSpline, gaussian, linear and lanczos interpolator for affine and SyN registration.

3 Results

We applied our fully Python scripted pipeline to 112 patients treated by one 20Gy fraction SBRT for small BMs (<2cm).
For the T1 gadolinium enhanced MR to ICBM registration, mean DICE index of brain cortex was 0.82 (range: 0.66-0.89) for the affine step and 0.91 (range: 0.81-0.94) when the affine + SyN steps were applied. For the CT to ICBM registration, the corresponding values were equal to 0.79 (range: 0.54-0.89) and 0.89 (range: 0.62-0.92), respectively. For the dose transformation, 4 patients with cortex DICE index <0.80 were excluded to avoid erratic DVH values.
Figure 2 illustrates PTV D95% and mean dose, brainstem D0.035cc and brain mean dose distribution in boxplots, depending on the warping dose modality (linear, bSpline, gaussian and Lanczos interpolator for affine and SyN part). Dose distributions were quite similar between initial dosimetry and the 16 warped dose maps for OARs D0.035cc and Dmean. The transformation using bSpline interpolator for both affine and SyN parts seems to be the better option for PTV mean dose and D95, as the distribution of the dose values were closer to the initial dose map for this warping dose modality.

4 Discussion

Here, we implemented a fully Python scripted pipeline allowing the visualization of brain pre-SBRT MRI and CT scan, dose plan and WM tracts on an open-source tractography atlas (as illustrated in Figure 3).

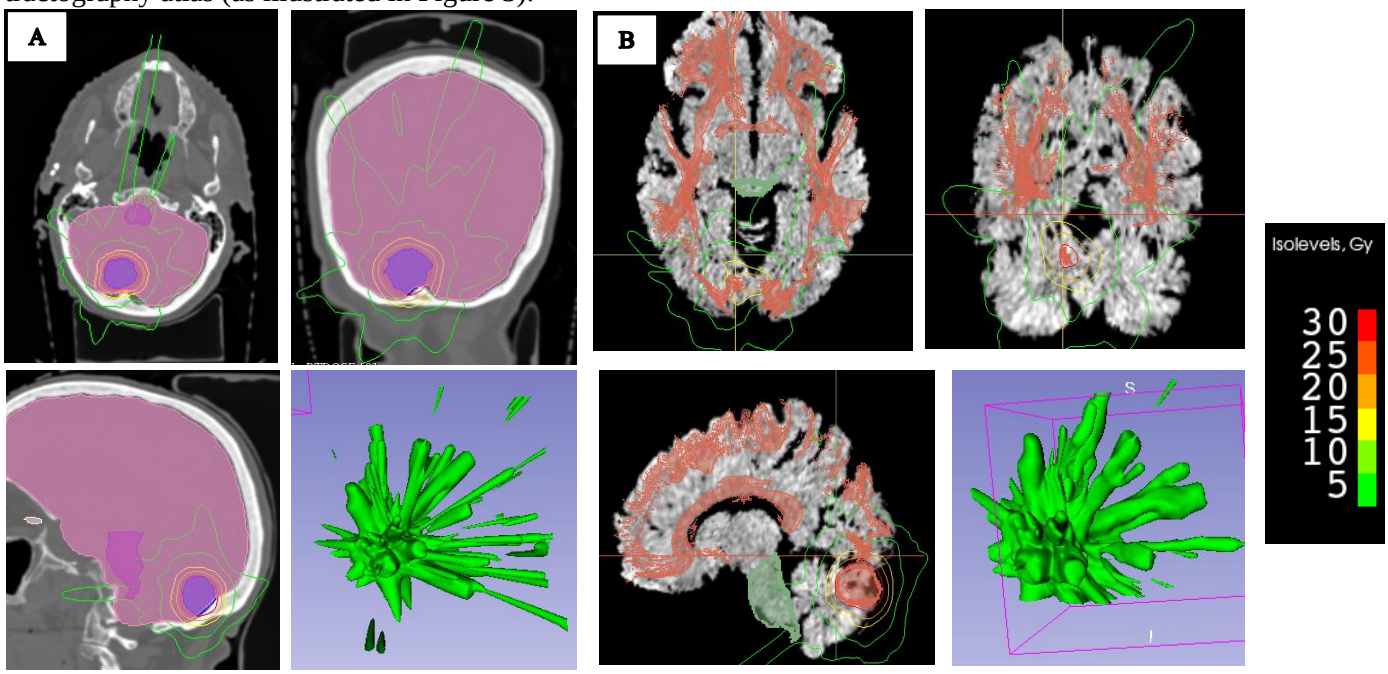


Figure 3. Workflow application on a patient treated with 1x20Gy SBRT for a right cerebellar BM.

A. Initial treatment planning on brain CT scan. B. Transformed brain CT scan and its associated dose map.

We choose to propose both MR and CT registration pipeline for different purpose. Because of its best anatomical accuracy, the T1 enhanced MR registered images could be used in a clinical setting for tract visualization and its relation to tumor volume before SBRT planification and could be a tool to guide the radiation oncologist therapeutic decision. On the other hand, because planning dose maps and RT structures segments are derived from the planning CT, the CT registered images will allow DVH calculation on WM tracts. We seek to derive dose constraints to WM tracts by applying our pipeline to a large amount of BMs patients already treated by SBRT in our institution. As it stands, the computational time is about 30 minutes per patient on a NVIDIA RTX A1000 Laptop 4 GB GPU, but we seek to improve the performance of the pipeline with more powerful GPU.
Regarding our choice of Mutual Information as similarity metric, several intensity-based image metrics appear in the literature with varying performance depending on their application, but mutual information has advantages in dealing with scanner variations and pathomorphological changes [8] and was more suitable for BM population and for multimodality (T1 gado weighted MR to T1 weighted MR and CT to T1 weighted MR) image registration.
Furthermore, adding an individual auto segmented cortical mask in the parameters of the SyN registration provided an anatomical constraint to the registration, favoring clinical plausibility of the transformation.
Despite recent guidelines on dose mapping/accumulation in radiotherapy [9], there is still some discussion on how dose distributions should be resampled when deforming them with a transformation and no consensus has been made on the choice of the interpolation method. To address this, we choose to evaluate 4 different type of interpolation for affine and SyN dose warping, resulting in 16 warped dose maps. It appears that for mean and maximum doses to OAR, there

is a very slight difference between the 16 interpolation methods. On the other hand, the bSpline interpolation method showed better similarity to the initial dose regarding the mean dose and the D95% delivered to the PTV. BSpline interpolation seem to provide better consistency in dose distribution when dose warping, especially for high doses.

5.Conclusion

Here, we present the first fully automated pipeline allowing registration of pre-SBRT planning imaging, dose map and atlas derived WM tracts. This easily applicable pipeline will make more accessible WM structural connectivity to the radiation oncologist community. From this method, applicable without performing DTI in a large number of patients, we seek to derive dose constraints to WM tracts. In the future we believe that this method will enable radio oncologists and physicists to propose WM sparing dosimetry.

References

1. Sarubbo S, De Benedictis A, Merler S, Mandonnet E, Balbi S, Granieri E, Duffau H. Towards a functional atlas of human white matter. Hum Brain Mapp. 2015 Aug;36(8):3117-36. doi: 10.1002/hbm.22832. Epub 2015 May 9. PMID: 25959791; PMCID: PMC6869563.

2. Sarubbo S, Tate M, De Benedictis A, Merler S, Moritz-Gasser S, Herbet G, Duffau H. Mapping critical cortical hubs and white matter pathways by direct electrical stimulation: an original functional atlas of the human brain. Neuroimage. 2020 Jan 15;205:116237. doi: 10.1016/j.neuroimage.2019.116237. Epub 2019 Oct 15. PMID: 31626897; PMCID: PMC7217287.

3. Aizer AA, Lamba N, Ahluwalia MS, Aldape K, Boire A, Brastianos PK, Brown PD, Camidge DR, Chiang VL, Davies MA, Hu LS, Huang RY, Kaufmann T, Kumthekar P, Lam K, Lee EQ, Lin NU, Mehta M, Parsons M, Reardon DA, Sheehan J, Soffietti R, Tawbi H, Weller M, Wen PY. Brain metastases: A Society for Neuro-Oncology (SNO) consensus review on current management and future directions. Neuro Oncol. 2022 Oct 3;24(10):1613-1646. doi: 10.1093/neuonc/noac118. PMID: 35762249; PMCID: PMC9527527.

4. Le Rhun E, Guckenberger M, Smits M, Dummer R, Bachelot T, Sahm F, Galldiks N, de Azambuja E, Berghoff AS, Metellus P, Peters S, Hong YK, Winkler F, Schadendorf D, van den Bent M, Seoane J, Stahel R, Minniti G, Wesseling P, Weller M, Preusser M; EANO Executive Board and ESMO Guidelines Committee. Electronic address: clinicalguidelines@esmo.org. EANO-ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up of patients with brain metastasis from solid tumours. Ann Oncol. 2021 Nov;32(11):1332-1347. doi: 10.1016/j.annonc.2021.07.016. Epub 2021 Aug 6. PMID: 34364998.

5. Yeh FC. Population-based tract-to-region connectome of the human brain and its hierarchical topology. Nat Commun. 2022 Aug 22;13(1):4933. doi: 10.1038/s41467-022-32595-4. PMID: 35995773; PMCID: PMC9395399.

6. Tustison NJ, Cook PA, Holbrook AJ, Johnson HJ, Muschelli J, Devenyi GA, Duda JT, Das SR, Cullen NC, Gillen DL, Yassa MA, Stone JR, Gee JC, Avants BB. The ANTsX ecosystem for quantitative biological and medical imaging. Sci Rep. 2021 Apr 27;11(1):9068. doi: 10.1038/s41598-021-87564-6. PMID: 33907199; PMCID: PMC8079440.

7. Brudfors M. Generative Models for Preprocessing of Hospital Brain Scans. University College London (UCL) PhD, 2020. https://discovery.ucl.ac.uk/id/eprint/10090849.

8. Avants BB, Tustison NJ, Song G, Cook PA, Klein A, Gee JC. A reproducible evaluation of ANTs similarity metric performance in brain image registration. Neuroimage. 2011 Feb 1;54(3):2033-44. doi: 10.1016/j.neuroimage.2010.09.025. Epub 2010 Sep 17. PMID: 20851191; PMCID: PMC3065962.

9. Hub M, Kessler ML, Karger CP. A stochastic approach to estimate the uncertainty involved in B-spline image registration. IEEE Trans Med Imaging. 2009 Nov;28(11):1708-16. doi: 10.1109/TMI.2009.2021063. Epub 2009 May 12. PMID: 19447703.

10. Murr M, Brock KK, Fusella M, Hardcastle N, Hussein M, Jameson MG, Wahlstedt I, Yuen J, McClelland JR, Vasquez Osorio E. Applicability and usage of dose mapping/accumulation in radiotherapy. Radiother Oncol. 2023 May;182:109527. doi: 10.1016/j.radonc.2023.109527. Epub 2023 Feb 10. PMID: 36773825.

Comparison of Geant4-DNA and RITRACKS/RITCARD: microdosimetry, nanodosimetry and DNA break predictions

Victor LEVRAGUE[1], Rachel DELORME[1], Mathieu ROCCIA[1], Ngoc Hoang TRAN[2], Sebastien INCERTI[2], Michael BEUVE[3], Etienne TESTA[3], Ianik PLANTE[4], Shirin RAHMANIAN[5] and Floriane POIGNANT[5]

[1]Univ. Grenoble Alpes, CNRS, Grenoble INP, LPSC-IN2P3, 38000 Grenoble, France
[2]Université de Bordeaux, CNRS/IN2P3, LP2i, UMR 5797, 33170 Gradignan, France
[3]IP2I, Université Claude Bernard Lyon 1, CNRS/IN2P3, 4 rue Enrico Fermi, 69622 Villeurbanne Cedex, France
[4]KBR, Houston, TX 77058, USA
[5]Analytical Mechanics Associates Inc., Hampton, VA 23666, USA

Abstract This work aims at investigating the impact of DNA geometry, compaction and calculation chain on DNA break and chromosome aberration predictions for high charge and energy (HZE) ions, using the Monte Carlo codes Geant4-DNA, RITRACKS and RITCARD. To ensure consistency of ion transport of both codes, we first compared microdosimetry and nanodosimetry spectra for different ions of interest in hadrontherapy and space research. The Rudd model was used for the transport of ions in both models. Developments were made in Geant4 (v11.2) to include periodic boundary conditions (PBC) to account for electron equilibrium in small targets. Excellent agreements were found for both microdosimetric and nanodosimetric spectra for all ion types, with and without PBC. Some discrepancies remain for low-energy deposition events, likely due to differences in electron interaction models. The latest results obtained using the newly available Geant4 example "dsbandrepair" will be presented and compared to DNA break predictions obtained with RITCARD.

1 Introduction

At the cellular scale, energy deposition events induced by low linear energy transfer (LET) ionizing radiation create biological damage including DNA double strand breaks (DSB) that are relatively homogeneously distributed. On the contrary, high LET radiation, such as high charge and energy (HZE) ions, are well known to induce heterogeneously distributed energy deposition events that can create complex biological damage, including clustered DSB. Clustered DSB are critical damage events difficult to repair, leading to increased cell death and chromosome aberrations (CAs) compared to low LET radiation exposure. This increased biological effectiveness makes high LET ions of great interest in medicine for increasing the therapeutic efficacy of cancer treatment, an advantage utilized in hadrontherapy. However, exposure to HZE ions also comes with carcinogenesis risks, which is an issue for both hadrontherapy (secondary cancer) and long-term space travel. Cellular biological mechanisms driving cancer initiation, promotion and progression for high-LET exposure remain to be fully elucidated[1] and constitute an active field of research. In that context, Monte Carlo track structure simulations are valuable tools to identify key cellular events and quantify the mechanisms.

The track structure code RITRACKS (Relativistic Ion Tracks) [Plante 2008], together with the code RITCARD (Relativistic Ion Tracks, Chromosome Aberrations, Repair, and Damage) [Plante 2019a], allow to simulate the transport of HZE ion tracks in micro and nanostructures. The yield of DNA damage is determined by calculating energy deposited in nanovoxels containing DNA, without details of the DNA geometry. Estimation of CA yields are obtained by modeling DSB (mis)repair and classifying the aberrations. However, DNA geometry influence the yield of DNA breaks. Previous numerical studies have shown that chromatin compaction decreases the yield of DNA damage generated by ionizing radiation, with dependence on ion charge and energy [Tang 2019]. Such changes in DNA damage yield are likely to impact CA yields and are thus important to consider in DNA damage and repair simulations.

We intend to use highly detailed DNA geometries available in the Geant4-DNA toolkit to compute DNA breaks for different level of DNA compaction and combine the results with RITCARD to determine the impact of DNA compaction on CA predictions. A first necessary step to fulfill this objective is to ensure the consistency of radiation transport models used in RITRACKS and Geant4-DNA.

The work that will be presented corresponds to this first phase of the project. We compared energy deposition at micrometric and nanometric scales for different ions with LET ranging from ~0.4 keV/µm up to 235 keV/µm. The calculation of DNA breaks and their complexity with the new calculation chain available in Geant4 11.2 is an ongoing work, and comparison with RITCARD will be presented based on availability at the time of the conference.

2 Materials and Methods

2.1 Geant4-DNA

Geant4-DNA is an open-source code developed through an international collaboration, that was originally initiated in the context of space radiation and extended to medical physics applications. It includes physics processes that allow event-by-event particle transport down to very low energy (typically down to 7 eV for electrons). It includes

[1] https://humanresearchroadmap.nasa.gov/

different geometric models of DNA for a full human cell nucleus with a detailed atomic-scale structure of the molecule[2] [Incerti 2018, Kyriakou 2022]. This precision allows to account for damage coming from both direct ionization of the DNA by ion tracks and secondary electrons, and the attack of chemical species on DNA, produced by the radiolysis of the surrounding water.

The presented results were obtained with the latest version of Geant4 (v11.2) using DNA-physics option 2, for which the model of heavy ion transport (*G4DNARuddIonisationExtendedModel)* was significantly updated. Calculations of the energy distribution of secondary electrons induced by water-ion ionization showed that Geant4 v11.2 results in a much better agreement with RITRACKS compared to previous Geant4 versions. In addition, developments were made in our Geant4 example to include periodic boundary conditions (PBC), based on previous work[3], as a solution to account for energy deposition due to delta electrons generated by distant tracks, simulating electronic equilibrium in small targets. The PBC is applied to electrons. We have also implemented an efficient nano-voxel mesh scorer compatible with the PBC option in order to compute the nanodosimetry spectra. Finally, we modified the physics list to extend the maximum energy limit for the *Rudd Ionisation Extended Model* from 300 MeV/n to 1 GeV/n for ions. We also modified the physics list to use the *Rudd Ionisation Extended Model* for proton and alpha particles. For alpha particles, cross sections were calculated to extend them from 100 MeV/n up to 1 GeV/n. This allows a consistent use of the Rudd ionization model for all ions.

The Geant4-DNA version 11.2 also contains two new DNA break calculation chains, called "*moleculardna*" and "*dsbandrepair*" that allows to calculate early radiation-induced DNA damage, such as DSBs and to model repair processes. The *dsbandrepair* application also contains different atomic scale models of DNA, including both the radioresistant heterochromatin (HC) compact form and the radiosensitive euchromatin (EC) less compact form. In a second step of this benchmark process, we intend to use this example to compare results of simple and complex strand break yields in our irradiation configurations, using both EC and HC DNA forms and compare it with the RITRACK/RITCARD DSB predictions. This is an ongoing work that will not be described further in this abstract.

2.2 RITRACKS/RITCARD

RITRACKS/RITCARD has been developed by NASA, focusing on space radiation applications. RITRACKS is a Monte Carlo track-structure code that simulates the full transport of protons and high charge and energy particles using the Rudd model, as well as the secondary electrons down to very low energy, as detailed elsewhere [Plante 2008, Plante 2011a]. Recent developments include the possibility to activate PBC in micrometric volumes to account for energy deposition in targets by energetic delta electrons generated by distant tracks. RITRACKS can calculate energy deposition in nanovoxels as well as in spherical and cylindrical micrometric targets [Plante 2011b, Plante 2021]. Contrary to Geant4-DNA, RITCARD calculates DNA DSBs based on energy deposited within nanovoxels containing 2,000 kbp of DNA, encompassing both the direct and indirect effects without going down to details at the atomic scale. RITCARD also includes repair processes which can account for dose-rate effect. Finally, RITCARD allows to classify the damage events into CA, an option not yet available in Geant4-DNA.

2.3 Micro- and nanodosimetry spectra

We calculated single track energy deposition distributions in micrometric and nanometric targets for several monoenergetic ion beams. The energy imparted to the target, ε (in eV or J) is,

$$\varepsilon = \sum_i \varepsilon_i,$$

where the summation over ε_i represents all the energy deposition events (ionizations/excitations). $f_1(\varepsilon)$ is defined as the probability distribution of ε for a single ion track generating energy deposition events in the target. Comparisons were made for the following ions: $^{1}H^{+}$ (150 and 250 MeV), $^{4}He^{2+}$ (250 MeV/n), $^{12}C^{6+}$ (290 MeV/n), $^{16}O^{8+}$ (55 and 350 MeV/n), $^{28}Si^{14+}$ (170 MeV/n), $^{48}Ti^{22+}$ (300 MeV/n) and $^{56}Fe^{26+}$(300, 450 and 600 MeV/n). These ions were chosen because of their interest in hadrontherapy and space research, and experimental chromosome aberration data are available [Hada 2014, Plante 2019a, Plante 2019b]. The simulation set up for microdosimetry calculation is described in Figure 1. For both microdosimetry and nanodosimetry calculations, the geometry was set as a water box of 12 µm in the longitudinal direction (direction of ion transport), and 16 µm x 16 µm in transverse directions. It is worth mentioning that with the PBC option activated, when a delta-electron is re-injected in the world volume, it is considered as a new track, as it mimics contribution from distant tracks. To mimic particle equilibrium, PBC were applied as followed: when an electron passed through one side of the irradiated volume, it appeared on the opposite side with the same velocity vector. The irradiated box contained a micrometric target which was a centered ellipsoid volume with a transverse radius of 7.1 µm and a longitudinal radius of 2.5 µm, representing the cell nucleus of a human fibroblast in an adherence configuration[4]. The nanometric targets were cubical voxels of 75 nm side that

[2] http://geant4-dna.org/

[3] https://github.com/amentumspace/g4pbc

[4] https://geant4-dna.github.io/molecular-docs/

were distributed within the cell nucleus, which aims at representing a fragment of DNA. Single ion energy distribution $f_1(\varepsilon)$ were calculated using a logarithmic binning (1,000 bins) between 6 eV to 2×10^6 eV for microdosimetry and 6 eV to 2×10^4 eV for nanodosimetry. For microdosimetry, results are normalized so that the first moment of $f_1(\varepsilon)$ gives the average energy deposited by one track to a target. For nanodosimetry, results were normalized so that $\int_0^\infty f_1(\varepsilon)\,\mathrm{d}\varepsilon = 1$.

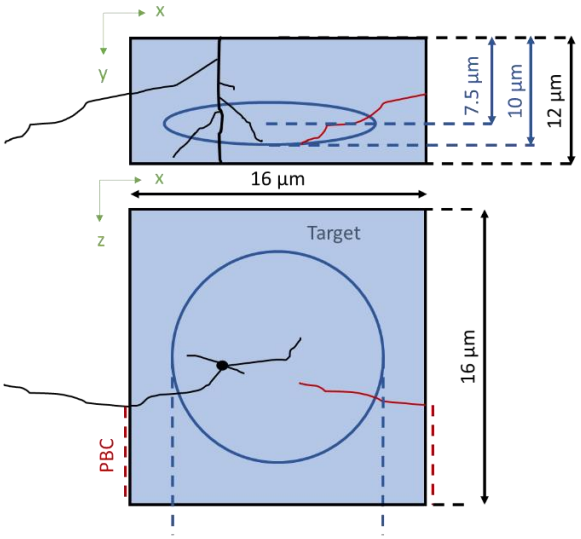


Figure 1. Simulation set-up for microdosimetry calculation. The light blue volume represents the simulated water volume, the dark blue circle the microdosimetry target, the black lines an ion track, the red line an electron that exited the volume and for which PBC were applied.

3 Results

Figure 2 shows the microdosimetry results obtained with Geant4 11.2 and RITRACKS with activated PBC for 9 different ions. Figure 3 shows the nanodosimetry results obtained with Geant4 11.2 and RITRACKS with activated PBC for 9 different ions. Excellent agreements were found for both microdosimetric and nanodosimetric spectra for all ion types and energies, with and without PBC *(without PBC are not shown)*. Microdosimetry spectra differ mainly for low-energy deposition events below 100 eV, especially for He, C and O ions, likely due to differences in electron interaction models. In the nanodosimetry spectra, slight discrepancies are observed in the 100 eV-10 keV ε range for heavy ions. These results will be presented at the conference with a deeper analysis on the comparisons.

4 Discussion and Conclusion

The present work aims at investigating the impact of DNA geometry on DNA break and CA predictions for HZE ions using the codes Geant4-DNA, RITRACKS and RITCARD. The first phase of the comparison process showed excellent agreements in microdosimetry and nanodosimetry spectra. This allows us to validate the Geant4-DNA newly released *Rudd Extended Ionisation Model* for heavy ion transport by comparing it with RITRACKS, that has been previously validated on several experimental data for HZE ions. In addition, we included changes in the Geant4 *Rudd Extended model* to also allow its use for alpha particles beyond 100 MeV/n. Finally, practical improvements were made to allow the use of PBC and registering physical data in both microstructures and nanovoxels, improvements that future Geant4 users may benefit from. The second step of code comparison between Geant4 and RITRACK/RITCARD is ongoing and will be dedicated to DNA break comparisons. It will focus on the impact of DNA geometry, compaction, and detailed direct/indirect effect consideration for DSB, and how such changes impact CA yields.

References

[Plante 2008] Plante, I., & Cucinotta, F. A. (**2008**). Ionization and excitation cross sections for the interaction of HZE particles in liquid water and application to Monte Carlo simulation of radiation tracks. *New Journal of Physics*, 10(12), 125020. https://doi.org/10.1088/1367-2630/10/12/125020

[Plante 2019a] Plante, I., Ponomarev, A., Patel, Z., Slaba, T., & Hada, M. (**2019**). RITCARD: radiation-induced tracks, chromosome aberrations, repair and damage. *Radiation Research*, 192(3), 282-298. https://doi.org/10.1667/RR15250.1

[Tang 2019] Tang, N., Bueno, M., Meylan, S., Incerti, S., Tran, H. N., Vaurijoux, A. *et al.* (**2019**). Influence of chromatin compaction on simulated early radiation-induced DNA damage using Geant4-DNA. *Medical Physics*, 46(3), 1501-1511. https://doi.org/10.1002/mp.13405

[Incerti 2018] Incerti, S., Kyriakou, I., Bernal, M.A., Bordage, M.C., Francis, Z., Guatelli, S. *et al.*, J.M.C. (2018), Geant4-DNA example applications for track structure simulations in liquid water: A report from the Geant4-DNA Project. Med. Phys., 45: e722-e739. https://doi.org/10.1002/mp.13048

[Kyriakou 2022] Kyriakou, I.; Sakata, D.; Tran, H.N.; Perrot, Y.; Shin, W.-G.; Lampe, N. *et al.* Review of the Geant4-DNA Simulation Toolkit for Radiobiological Applications at the Cellular and DNA Level. *Cancers* **2022**, *14*, 35. https://doi.org/10.3390/cancers14010035

[Plante 2011a] Plante, I., & Cucinotta, F. A. (**2011**). Monte-Carlo simulation of ionizing radiation tracks. *Application of Monte Carlo methods in biology, medicine and other fields of science*. InTech, Rijeka, Croatia, 315-356. https://doi.org/10.5772/15674

[Plante 2011b] Plante, I., Ponomarev, A., & Cucinotta, F. A. (**2011**). 3D visualisation of the stochastic patterns of the radial dose in nano-volumes by a Monte Carlo simulation of HZE ion track structure. *Radiation Protection Dosimetry*, 143(2-4), 156-161. https://doi.org/10.1093/rpd/ncq526

[Plante 2021] Plante, I., Poignant, F., & Slaba, T. (**2021**). Track structure components: Characterizing energy deposited in spherical cells from direct and peripheral HZE ion hits. Life, 11(11), 1112. https://doi.org/10.3390/life11111112

[Hada 2014] Hada, M., Chappell, L. J., Wang, M., George, K. A., & Cucinotta, F. A. (**2014**). Induction of chromosomal aberrations at fluences of less than one HZE particle per cell nucleus. Radiation Research, 182(4), 368-379. https://doi.org/10.1667/RR13721.1

[Plante 2019b] Plante, I., Slaba, T., Shavers, Z., & Hada, M. (**2019**). A bi-exponential repair algorithm for radiation-induced double-strand breaks: Application to simulation of chromosome aberrations. *Genes*, 10(11), 936. https://doi.org/10.3390/genes10110936

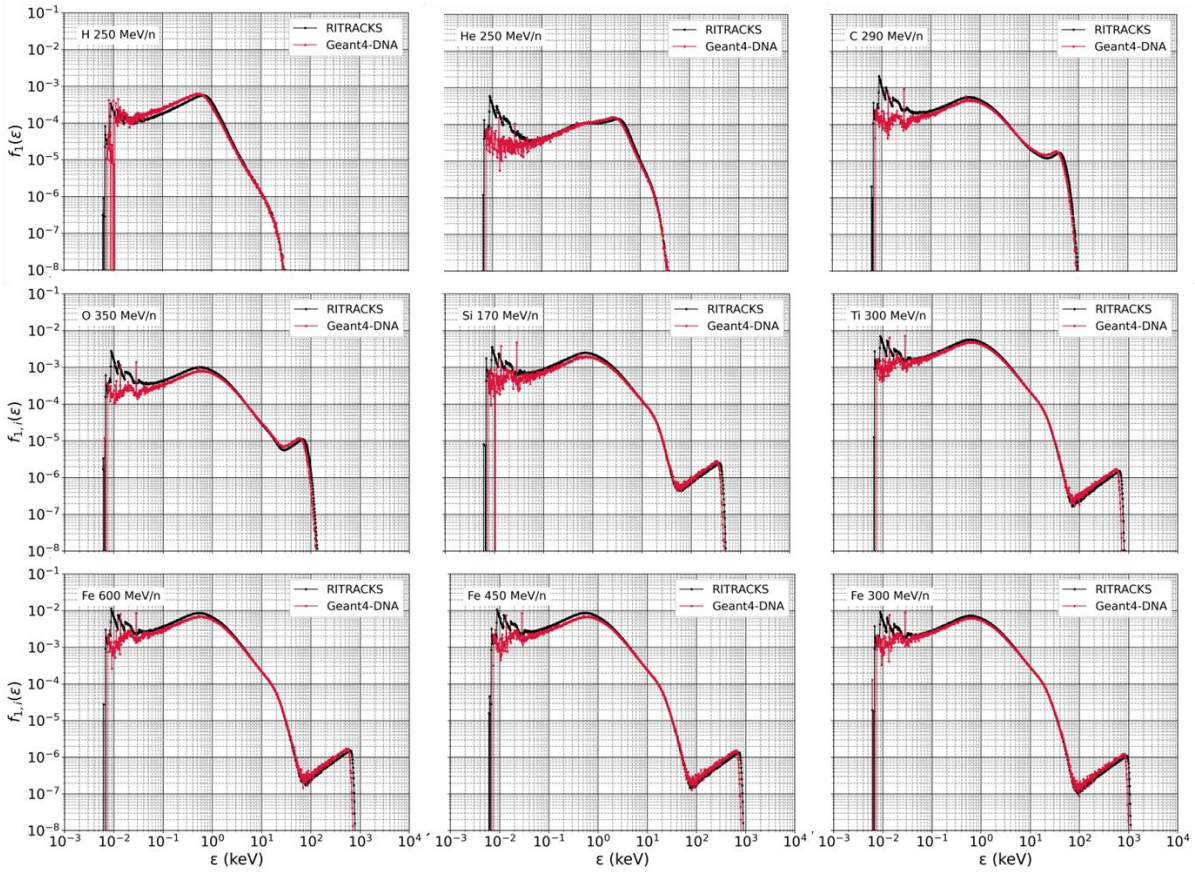


Figure 2. Comparison of single ion energy distribution in micrometric targets between RITRACKS (black) and Geant4-DNA (red).

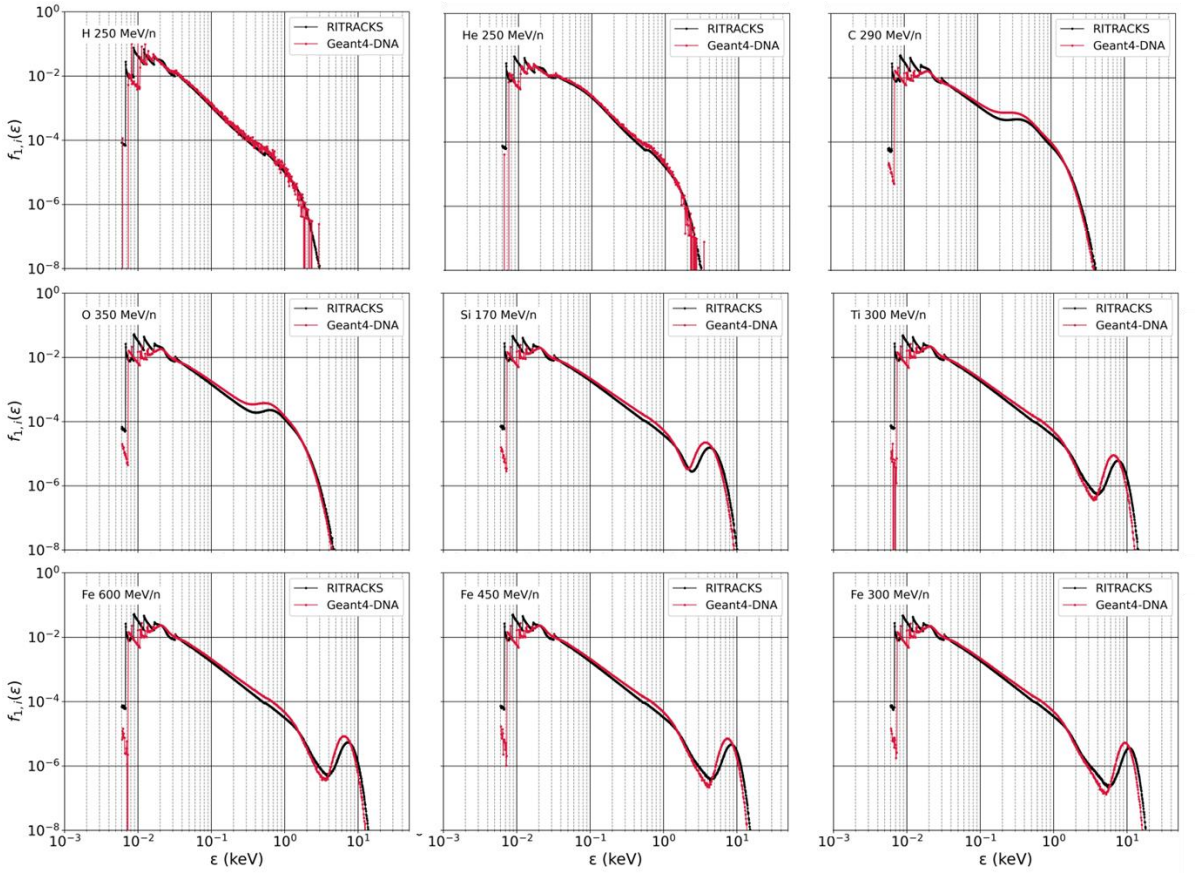


Figure 3. Comparison of single ion energy distribution in nanometric targets between RITRACKS (black) and Geant4-DNA (red).

Whole-body blood circulation modeling and dose to circulating blood in intensity modulated radiation therapy

Bingqi Guo, PhD[1], Ping Xia, PhD[1]

[1]Department of Radiation Oncology, Cleveland Clinic, Cleveland, OH, USA

Abstract Dose to circulating blood is gaining interest in radiation oncology due to the potential impact of blood dose on the response of the immune system to radiation therapy. However, modeling the interplay of blood circulation with the dynamic delivery of intensity-modulated radiation therapy is challenging. In this study, a novel technique to model whole-body blood circulation was introduced. The model uses patient-specific anatomy from medical images and preserves realistic blood dynamics, distribution, and perfusion. To demonstrate the use of dynamic blood modeling in radiation oncology, the model was used to evaluate the dose to circulation blood in total body irradiation treated with multi-isocentric volumetric modulated arc therapy, one of the most complex techniques in radiation oncology.

1 Introduction

Dose to circulating blood is gaining interest in radiation oncology due to the potential impact of blood dose on the response of the immune system to radiation therapy. However, modeling the interplay of blood circulation with the dynamic delivery of intensity modulated radiation therapy is challenging. This study aims to develop a technique for realistic, patient-specific blood circulation modeling and explore its application in radiation oncology.

2 Materials and Methods

A. Creation of the whole-body blood circulation model

i. Organ, tissue and blood vessel segmentation

A whole-body CT scan with IV contrast was used for creation of the whole-body blood circulation model. Main organs of the body, heart chambers, aorta and vena cava, and pulmonary artery and veins were segmented using a commercial deep-learning based auto-segmentation software. Main arteries and veins were first segmented using an in-house algorithm which searches for tube-shaped contrast enhanced structures, then sorted into large arteries, hepatic veins, portal veins and other veins manually. Smaller vessels were assumed to distribute uniformly throughout the organs and tissues. Figure 1(a) shows the contours of the heart chambers, lungs, liver and main blood vessels.

ii. Blood dynamics

Figure 1(b) shows the blood dynamics diagram including the systole, diastole and portal circulations and the average blood concentration and perfusion values for main organs and tissues[1].

iii. Blood density map

The blood density map was created by dividing the blood concentration of each organ and tissue by its volume. Figure 1(c) shows the blood density map normalized to maximal value.

iv. Blood velocity maps

Blood velocity maps in the whole body in systole and diastole were calculated based on Table 1. In aorta, vena cava and pulmonary artery/veins, blood speed was assumed to decrease linearly as its travel distance from the heart chambers increases. For small vessels, organs and tissues, blood speed decreases linearly as the distance to the nearest large artery/vein increases. Figure 1 (d) and (e) shows the blood velocity map in systole and diastole, respectively.

Table 1. Blood speed

speed of blood [cm/s]	systole	diastole
Lt and Rt ventricals	60	28
Lt and Rt atriums		28
aorta and pul artery	20 - 60	
vena cava and pul veins		10 - 28
large arterie	20	
large veins		10
lungs, liver and smaller vessles	5 - 20	1.5 - 10
tissue	2 - 5	1.3 - 1.5

v. Time to heart/porta maps

Based on the blood velocity distributions and blood dynamics, the time to heart maps in systole and diastole were derived. For example, the systole time to heart in aorta is 0 at where aorta connects with left ventricle and increases as blood moves along the aorta. In every step of moving along the aorta, time to heart increases by delta distance divided by velocity. For portal circulation, time to porta hepatis, where the portal vein enters liver was calculated. Liver has negative time to porta hepatis values. Figure 1 (f)-(h) shows the time to heart/porta maps.

vi. Blood perfusion maps

Blood perfusion to each voxel in the whole body was calculated by dividing the blood perfusion of each organ and tissue by its volume. To create the (cumulative) perfusion maps, the path of blood flow from the heart to each voxel in systole and from each voxel to the heart in diastole was determined and the blood perfusion of the voxel was projected to the entire path. Time to heart/porta maps were used to determine the path of each voxel. Blood was assumed to move in the direction which minimizes the time to heart/porta while observing physiological limitations. The interface of aorta/vena cava/pulmonary artery/pulmonary veins with heart chambers should have the highest cumulative perfusion value of 100%. Figure 1 (i) and (j) shows the systole and diastole/portal blood perfusion maps, respectively.

vii. Simulation of blood circulation

In this study, heartbeat rate of 70 times per minute was used. One third of the heartbeat is systole and 2/3 is diastole. The blood density map, velocity maps, time to heart/porta maps, and perfusion maps were all used to simulate the blood circulation in the whole body. A blood "particle" was sampled from the whole-body blood density distribution at the beginning and its trajectory was calculated every 1/3 of heartbeat (dt). During systole, blood particle moves to a new location which satisfies the following conditions (1) it is a physiologically permitted move, e.g. from aorta to large artery; (2) the new location has a time to heart equals the current time to heart + dt; (3) the distance of the new location is no more than blood speed of current location times dt. All the voxels meeting conditions (1)-(3) were identified and the next step was sampled from the voxels using the systole perfusion map. Similar method was used to simulate trajectories in diastole and portal circulations.

Within each heartbeat, not all blood enters the circulation, some remain in the tissue/organ. For example, approximately 60% of blood exits the ventricles in each systole. The probability of a blood particle remaining in the current organ/tissue rather than continuing in the circulation is blood density / (blood perfusion + blood density).

Liver is an important organ in both diastole and portal circulations. In each diastole, the probability of a blood particle in liver continuing in portal circulation (moving away from portal veins) rather than diastole circulation (moving towards hepatic veins) equals portal perfusion / (portal perfusion + diastole perfusion).

1000 blood particles were simulated for 5 minutes. Two types of simulations were performed: assuming all blood particles exist left ventricle at the initial step, or randomly sample particles from the whole-body at the initial step.

B. Dose to circulating blood in VMAT-TBI

i. Multi-isocentric VMAT-TBI planning

Following our institutional protocol for total body irradiation (TBI), a multi-isocentric, volumetric modulated arc therapy (VMAT) plan was created to prescribe 12 Gy in 8 fractions to the whole-body while reducing doses to lungs and kidneys[2]. The VMAT-TBI plan consisted of 7 isocenters and 16 beams. 600MU/min was used for all isocenters except lung for which 200MU/min was used to reduce the mean lung dose rate.

ii. Dose to circulating blood

To calculate dose to circulating blood, segment (control point) doses of the VMAT-TBI plan were exported. Trajectories of the 1000 particles were synchronized with the VMAT-TBI delivery at each isocenter using an in-house delivery emulator. For each blood particle, doses received at every step in the trajectory during delivery were calculated and summed up. For comparison, dose to "static" blood pool was calculated by convolving the dose distribution of the VMAT-TBI plan with the blood density map.

3 Results

Figure 2 shows the trajectories of 100-1000 blood particles (a) existing the left ventricle and (b) randomly sampled from the body in 10 seconds-5 minutes. Each particle is assigned a random color. For small number of particles and/or short duration of simulation, the blood particles concentrate on large vessels and major organs (heart, lungs, and liver). With a larger number of particles and longer simulation time, trajectories of blood particles occupy the whole-body.

Figure 2(c) shows the dose distribution of the VMAT-TBI plan and the isocenters, beams, and delivery times at each isocenter. Delivery time ranges from 50s to 332s with lung isocenter having the longest delivery time. For this VMAT-TBI plan, figure 2(d) compares mean doses and cumulative dose volume histogram (DVH) to the "static" and dynamic blood pool in one fraction. The mean dose is comparable between the circulating blood and the "static" blood pool (1.50 +/- 0.35 Gy vs 1.43 +/- 0.36 Gy for the whole plan, and 0.66 +/- 0.34 Gy vs 0.75 +/- 0.60 Gy for the lung isocenter, respectively). However, dose volume histogram is very different. The mean dose to the circulating blood in VMAT-TBI equals the prescription dose (1.5 Gy in one fraction).

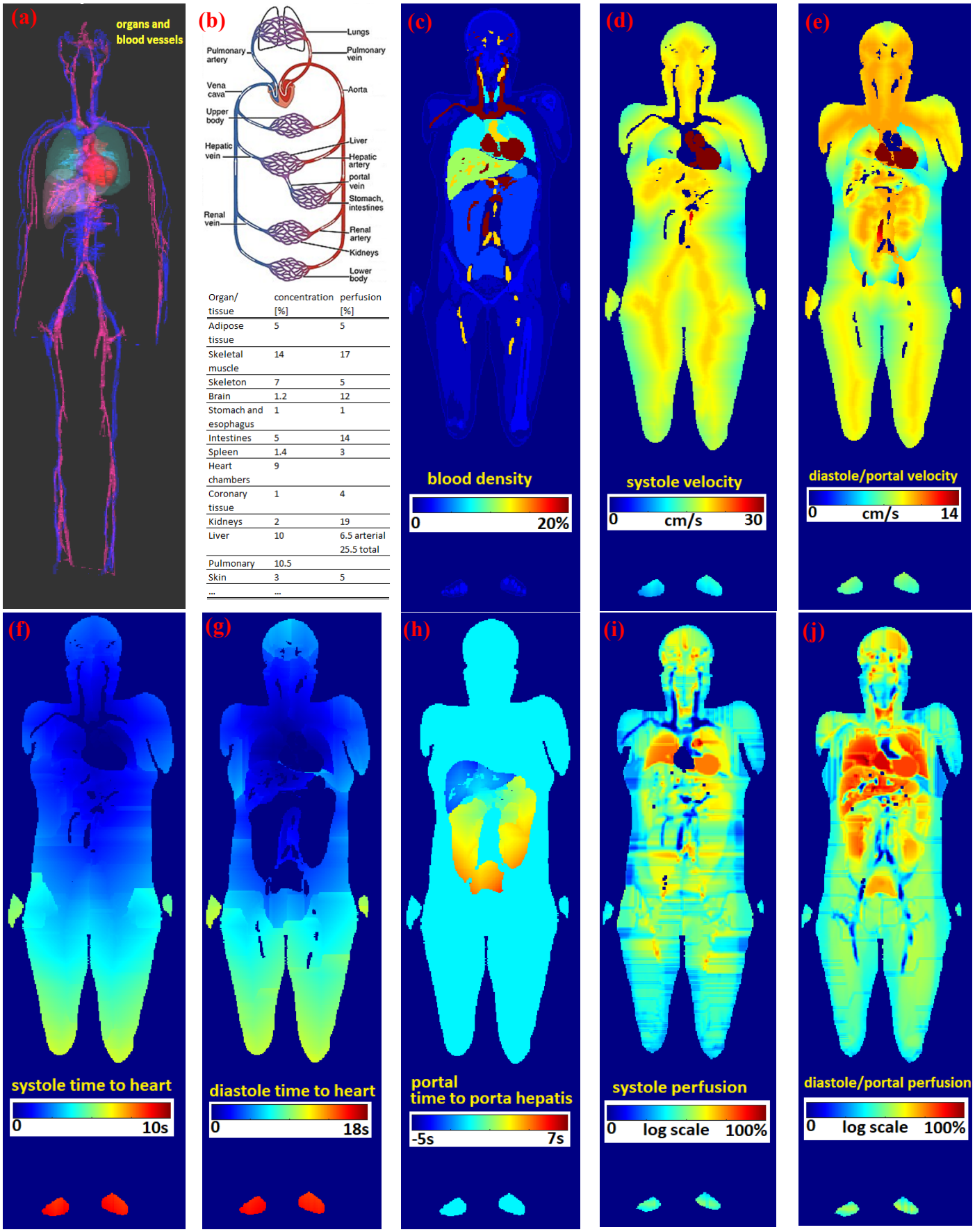
4 Conclusion

In this study, a novel technique to model the whole-body blood circulation was introduced. The model used patient-specific anatomy from CT images and realistic blood dynamics, distribution and perfusion. To demonstrate the application of dynamic blood modeling in radiation oncology, the model was used to calculate the dose to circulation blood in total body irradiation treated with multi-isocentric volumetric modulated arc therapy, one of the most complex techniques in radiation oncology.

References

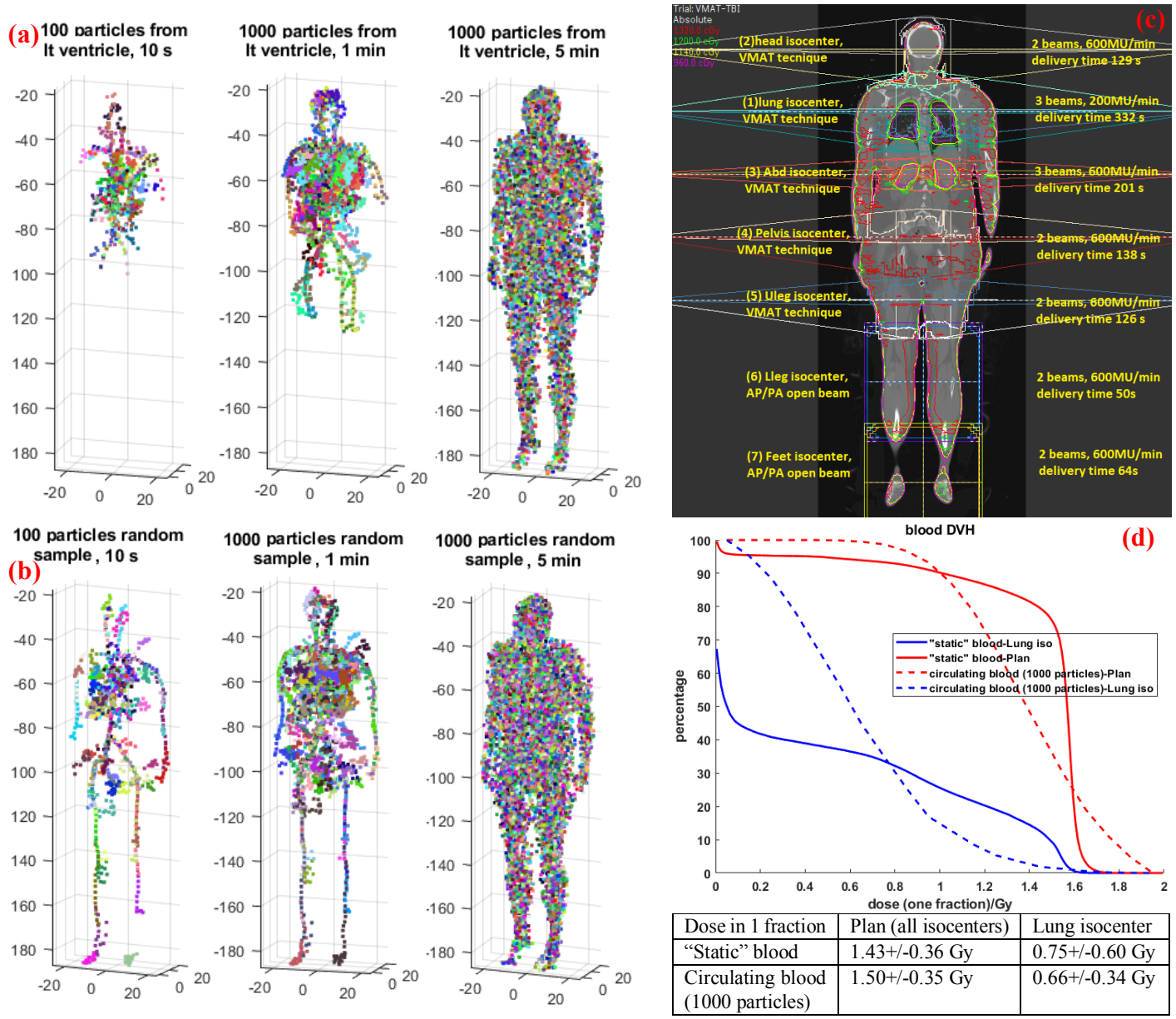
[1] Leggett RW, Williams LR. A proposed blood circulation model for Reference Man. Health Phys. 1995 Aug;69(2):187-201. doi: 10.1097/00004032-199508000-00003. PMID: 7622365.

[2] Guo B, Sheen C et al. Image-guided volumetric-modulated arc therapy of total body irradiation: An efficient workflow from simulation to delivery. 2021 Oct;22(10):169-177. doi: 10.1002/acm2.13412. PMID: 34480829.



Organ/ tissue	concentration [%]	perfusion [%]
Adipose tissue	5	5
Skeletal muscle	14	17
Skeleton	7	5
Brain	1.2	12
Stomach and esophagus	1	1
Intestines	5	14
Spleen	1.4	3
Heart chambers	9	
Coronary tissue	1	4
Kidneys	2	19
Liver	10	6.5 arterial 25.5 total
Pulmonary	10.5	
Skin	3	5
...	...	

Figure 1. Creation of the whole-body blood circulation model. (a) Segmentation of organs and main blood vessels. (b) Blood circulation diagram and blood concentration/perfusion to main organs/tissues[1]; (c) Normalized blood density distribution; (d) Blood velocity in systole; (e) Blood velocity in diastole; (f) Time to heart in systole; (g) Time to heart in diastole; (h) Time to porta hepatis in portal circulation; (i) Systole blood perfusion; (j) Diastole/portal blood perfusion.



Dose in 1 fraction	Plan (all isocenters)	Lung isocenter
"Static" blood	1.43+/-0.36 Gy	0.75+/-0.60 Gy
Circulating blood (1000 particles)	1.50+/-0.35 Gy	0.66+/-0.34 Gy

Figure 2. Results and application of the whole-body blood circulation model. (a) Trajectories of blood particles emitted from left ventricle; each particle is assigned a random color; (b) Trajectories of blood particles randomly sampled from the whole-body; (c) Total body irradiation (TBI) using multi-isocentric volumetric arc therapy (VMAT). The VMAT-TBI plan consists of 7 isocenters and 16 beams to prescribe 12 Gy in 8 fractions to the whole-body while reducing mean lung dose to 80% of the prescription and mean kidney dose to 50% of the prescription[2]; Delivery time at each isocenter ranges from 50s to 332s; (d) Dose to the circulating blood is compared with dose to "static" blood pool for VMAT-TBI. Mean doses and cumulative dose volume histogram (DVH) of the plan (all isocenters) and lung isocenter are compared between static and dynamic blood pool.

Using fluid flow simulations to model microscopic tumor invasion in muscle tissue for radiotherapy treatment

Gregory Buti[1], Ali Ajdari[1], Yen-Lin Chen[1], Christopher P. Bridge[2], Gregory C. Sharp[1], and Thomas Bortfeld[1]

[1]Massachusetts General Hospital and Harvard Medical School, Department of Radiation Oncology, Boston, USA
[2]Massachusetts General Hospital and Harvard Medical School, Athinoula A. Martinos Center for Biomedical Imaging, Charlestown, USA

Abstract
A framework is developed to include the preferred path of microscopic tumor invasion along muscle fibers in radiation target volumes. Following the novel idea that the directionality of muscle fibers inside the muscle can be represented by a flow velocity field, fluid flow simulations are performed with a computational fluid dynamics solver using the finite-elements method. The flow velocity fields are then embedded in a shortest path algorithm to model the microscopic infiltration around the gross tumor. The model assumes that the preferred path of microscopic tumor invasion is along the dominant muscle fiber direction estimated from the velocity field. The method was tested on a sacral chordoma tumor, a malignant tumor originating in the pelvis with suspected microscopic invasion of the gluteal muscles.

1 Introduction

Chondrosarcoma is a type of cancer that originates in cartilage cells and typically develops in bone structures of the pelvis, femur, upper arm, or shoulder blade [1]. One of the major uncertainties in the treatment of chondrosarcomas is how to account for microscopic tumor invasion into the surrounding musculature. It is known that tumor cells encounter greater resistance in the transverse direction of the muscle fiber than in the longitudinal direction [2]. Therefore, to maximize the therapeutic effect, one should use radiation target volumes with more generous margins in the longitudinal direction of the muscle fiber. However, defining such target volumes is a real challenge because CT images have poor soft tissue contrast and therefore cannot provide much intramuscular information. In addition, the typical clinical expansion algorithms that are used to generate radiation target volumes are unable to account for muscle-specific orientation and muscle fiber architecture. As a result, physicians are currently limited to manually delineating target volumes or using expansions along the major anatomical axes, which is an oversimplification of the complex muscle anatomy.
A promising computational approach in the field of biomechanics is the analysis of muscle force generation by representing muscle fibers as an idealized flow of a fluid within the boundaries of the muscle surface. In [3, 4], the equivalence was made between muscle fibers connecting bones through the origin and insertion with an irrotational and incompressible flow between a source and drain. It was shown that the orientation of the vector field from such fluid flow simulations can reproduce the microscopic muscle fiber orientations obtained from the patient's diffusion tensor image, which is a direct representation of the muscle microstructure.
The goal of this study is to combine recent advances in automatic expansion algorithms and fluid flow simulations to create radiation target volumes with local muscle fiber orientation into a single framework. The framework is tested for a sacral chordoma, a type of chondrosarcoma that develops in the sacrum with suspected infiltration in the surrounding gluteal muscles.

2 Materials and Methods

This section presents the framework for obtaining radiotherapy target volumes with muscle fiber information. The main goal is to generate the clinical target volume (CTV), which includes the volume of suspected microscopic tumor invasion. The CTV is obtained as a geometric expansion of the gross tumor volume (GTV), where the GTV-to-CTV expansion is spatially modulated according to the preferred mode of tumor invasion along the muscle fiber direction. The calculation is based on volume delineations of muscles and their corresponding attachment sites (muscle origins and insertions). Here we use a patient, called the *template* patient, with the complete set of delineations available, which serves as the basis for the flow calculations. The results can then be mapped to other different patients, called *target* patients, that do not have attachment delineations available. In short, the workflow consists of (a) performing fluid flow simulations in the template patient for the different gluteal muscles, (b) mapping the flow fields onto the target patient, and (c) generating radiation target volumes by including the mapped flow fields in a shortest path calculation [5]. The first step requires the use of the Finite Element Method (FEM) in a mesh-like geometry as explained below.

2.1 Flow simulation & mesh generation

The flow in a simply-connected domain can be described as a potential flow, through the use of a velocity potential ϕ. Assuming that the flow is both irrotational and incompressible, the Laplacian, ∇^2, of the velocity potential must be zero and the problem reduces to solving a Neumann boundary value

problem for the Laplace equation:

$$\begin{aligned} \nabla^2\phi &= 0, \quad \text{in muscle} \\ \nabla\phi \cdot \mathbf{n} &= 0 \quad \text{on muscle} \end{aligned} \tag{1}$$

$\mathbf{n}$ denotes the vector normal to the muscle surface, so the second equation imposes a zero flux condition at the muscle periphery. Fixed Dirichlet boundary conditions are imposed at the origin and insertion sites in order to simulate a flow between a source (origin) and drain (insertion):

$$\begin{aligned} \phi &= a \quad \text{in origin} \\ \phi &= b \quad \text{in insertion} \end{aligned} \tag{2}$$

with scalar values a and b. Since only the orientation of the potential flow is relevant to our problem, the values for a and b can be arbitrarily chosen as long as $a \neq b$. Finally, the flow velocity field $\mathbf{v}$ is obtained as $\mathbf{v} = \nabla\phi$. Fig. 1 A) gives a schematic overview of the problem statement.
All binary volumetric structures extracted from the patient dicom files were converted into volume meshes in Python. First, the *skimage* marching cubes algorithm is used to find a surface mesh from binary volumetric data. Then, the *pygalmesh* library is used to generate a volumetric tetrahedral mesh from the surface mesh [6], which is used as the domain for the FEM computation (see Fig. 1 B). Eq. (1) is solved using the open-source FEM computing platform FEniCS [7].

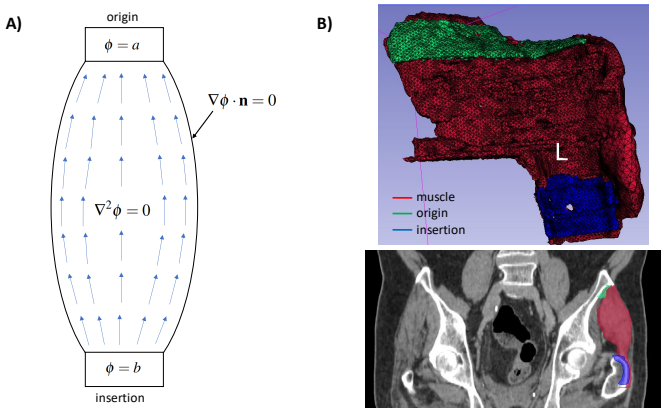


Figure 1: A) Schematic drawing of the Laplacian flow simulation problem. B) Example of the template gluteus medius mesh structures showing the muscle, origin and insertions (top), and associated CT image with structure segmentations (bottom).

2.2 Shortest path calculation & volume generation

A GTV-to-CTV margin can be obtained by using a shortest path calculation with GTV at the origin. Within this framework, the CTV is defined as an iso-distance contour in the 3D shortest distance map. The shortest distance S can be found as a solution to the Eikonal equation [5, 8]:

$$\nabla^{\mathsf{T}} S \cdot \mathbf{G}^{-1} \cdot \nabla S = 1, \tag{3}$$

where $\mathbf{G}$ is a positive semidefinite tensor of second order, called the Riemann metric tensor, or simply the metric. The metric plays an important role because it models the suspected tumor infiltration distance in a tissue. The relationship between the metric and the velocity field can be established by identifying the orientation of the velocity field as the orientation of greatest distance increase in the space defined by the metric. One way to formalize this relationship between $\mathbf{G}^{-1}$ and $\mathbf{v}$ is by first, creating a 3×3 tensor as the outer product, $\otimes$, of $\mathbf{v}$ with itself and performing a spectral decomposition:

$$\mathbf{v} \otimes \mathbf{v} = Q \cdot \Lambda \cdot Q^{-1}, \tag{4}$$

with Q and Λ the matrices of eigenvectors and eigenvalues, respectively. Second, replacing Λ by a parameterized diagonal matrix $diag(\alpha, 1, 1)$ and defining:

$$\mathbf{G}^{-1} \equiv Q \cdot diag(\alpha, 1, 1) \cdot Q^{-1}. \tag{5}$$

α is a dimensionless free parameter that controls the relative distance increase along the principal eigenvector which, in this context, corresponds to the preference for tumor invasion along the dominant muscle fiber direction.
The shortest path equation was solved using the Python-based fast marching solver HamiltonFastMarching [9].

2.3 Image registration & velocity field transformation

Image registration is used to transform the flow velocity fields from the template patient to the target patient for insertion in (3). Let $\mathbf{v}_1, ..., \mathbf{v}_n$ be the flow velocity fields for a set of n muscle compartments in the template patient. n separate displacement vector fields between the muscle compartment pairs. The transformed flow velocity fields in the target patient are then obtained by multiplying each velocity field by the *inverse* Jacobian matrix J of its associated displacement vector field, $\tilde{\mathbf{v}} = J^{-1} \cdot \mathbf{v}$ [10]. Similarly, the $\mathbf{G}^{-1}$ can be transformed directly according to the tensor transformation rule:

$$\tilde{\mathbf{G}}^{-1} = J^{-1} \cdot \mathbf{G}^{-1} \cdot J^{-\mathsf{T}}. \tag{6}$$

Eq. 6 allows the mapping of velocity fields between different patients by accounting for geometric differences such as differences in size and shape of muscles. Therefore, replacing $\mathbf{G}^{-1}$ by $\tilde{\mathbf{G}}^{-1}$ in (3) enables the calculation of a 3D shortest distance map in the target patient image.
Deformable image registrations between muscles was performed using DIPY (version 1.6.0) [11].

2.4 Patient data

A sacral chordoma patient, with manual delineations of the gluteus maximus, gluteus minimus and gluteus medius is used as a template image for obtaining muscle fiber information. The origin and insertion sites for all muscles of the template patient were delineated by an experienced radiation oncologist in our institution. Another sacral chordoma patient without origins/insertions available was used as a target image to generate the CTV. The hip bones, which act as

an impermeable barrier to tumor infiltration, were autosegmented on the CT image using TotalSegmentator (version 2.0.5) [12].

3 Results

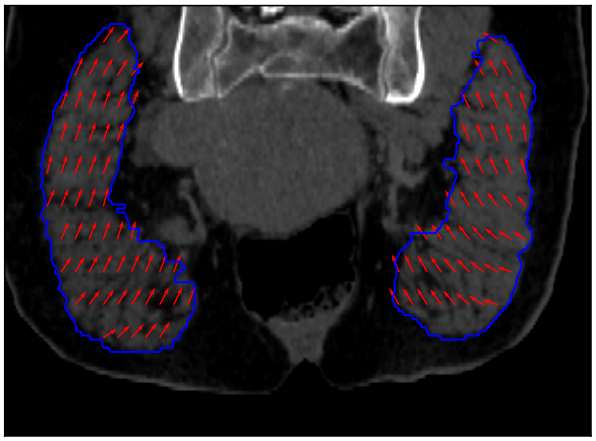
This section presents the results of the flow simulation in the template patient, followed by a comparison of the CTVs in the target patient with and without the inclusion of preferential models of tumor invasion.

3.1 Flow simulations

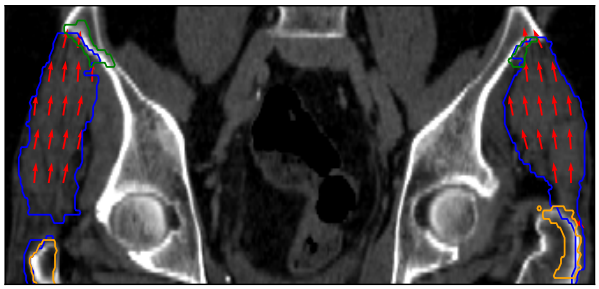
Fig. 2 shows the flow velocity fields in the template patient, overlayed on the CT image for the various gluteal muscles. As displayed, the FEM obtained smooth vector fields that can be mapped to the target patient.

3.2 CTVs

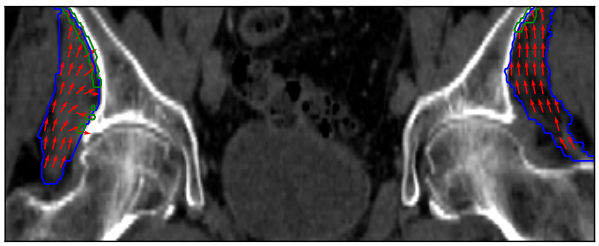
Fig. 3 compares the CTVs in the patients with and without muscle fiber orientation. The reference CTVs are defined as volumes with a 10 mm margin around the GTV in non-muscle tissue and larger margins of 20 mm and 30 mm in the gluteal muscles. The fiber-specific CTVs were generated by specifying a value of $\alpha > 1$ in the gluteal muscles and $\alpha = 1$ in the other tissues. As α increases, the model increases its preference for muscle invasion along the muscle fiber direction. We choose values for α equal to 4.0, and 9.1 because these specific values for α correspond to the standard CTV ranges mentioned earlier.
Significant differences in the shape of the CTVs are observed, especially as the CTV margins become more anisotropic. The difference between the reference CTV and the muscle fiber CTV was quantified using the 95% Hausdorff distance (HD95). The HD95 was 7.6 mm, and 13.9 mm for the (10 / 20) mm and (10 / 30) mm CTVs, respectively. This means that our proposed method has the potential to significantly impact current CTV delineation standards.



(a) Gluteus maximus



(b) Gluteus medius

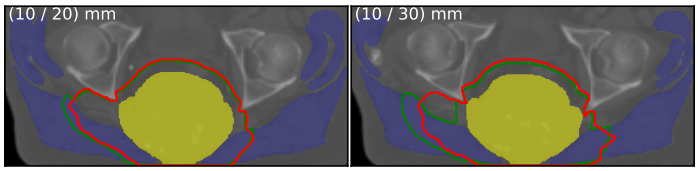


(c) Gluteus minimus

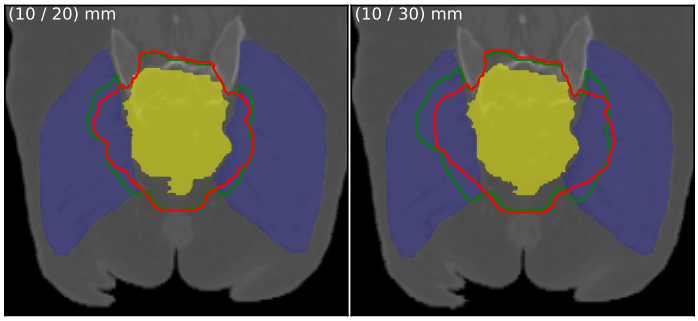
Figure 2: The flow velocity field obtained through Laplacian flow simulations are depicted by the red arrows within the (a) gluteus maximus, (b) gluteus medius and (c) gluteus minimus muscles. The flow vectors are assumed to be representative of the muscle fiber direction.

4 Discussion

In radiation therapy, the goal is to deliver a therapeutic dose to the CTV so that all tissues suspected of harboring tumor cells are treated. Currently, for chondrosarcoma treatments, the CTV is defined as a geometric expansion of the GTV by specifying a uniform margin in the muscle compartment. However, such an oversimplified approach does not take into account the mode of tumor invasion along the muscle fibers. As a result, tumorous tissue may be undertreated or radiation may be delivered to areas that are potentially tumor-free.
This study develops a computational method to improve the target volume definition for tumors with spread patterns along certain predefined structures such as skeletal muscles. Spread patterns were derived from Laplacian flow simulations to obtain patient-specific muscle fiber information. The Laplace equation is a fundamental equation in physics as it arises in a number of phenomena such as gravitational theory, electrostatics, fluid mechanics, and the heat equation. In this context, Laplace's equation is used to simulate a laminar flow that is representative of the directionality of the muscle fibers inside the muscle. Formally, potential flows governed by Laplace's equation are solutions to the time-independent Navier–Stokes equations for viscous (low Reynolds numbers), incompressible and irrotational fluids [4]. Realistic muscle fiber trajectories can be obtained by imposing different types of boundary conditions. First, a constant inflow and outflow was assumed at the origin and insertion sites, and second, a zero flux condition at the muscle periphery prevented muscle fibers from leaking through the outer muscle



(a) Axial slice.



(b) Coronal slice.

Figure 3: Comparison of the standard CTV with margins of (10 / 20) mm and (10 / 30) in non-muscle / muscle tissue in green, against the model-based CTVs in red, overlayed on the patient CT image. The GTV and muscle masks are shown in yellow and blue, respectively.

surface .

Once the flow fields have been computed with the FEM, insertion into the shortest path algorithm enables the generation of radiation target volumes. The shortest path map models the preferential invasion for miscroscopic tumor invasion in the tissues surrounding the gross tumor. Here, the flow velocity is equivalent to the presumed tumor infiltration distance along its principal direction.

A limitation of the current implementation is that muscle insertions are not currently delineated in clinical routine. Therefore, the decision was made to manually delineate origins and insertions in a single patient CT image, which served as a template to provide muscle fiber information. Automatic segmentation of insertion sites may be an interesting avenue for future research to further personalize the model.

Future research will focus on validating the model, and introducing tumor-specific effects such as the displacement of origin and insertion sites by the gross disease and disruptions of muscle fibers. In prospective setting, model validation requires the use of DTI. Alternatively, retrospectively, the orientation and magnitude of the flow velocity fields can be informed by areas of recurrence after treatment.

5 Conclusion

This study presents a novel framework for incorporating the preferred path of tumor invasion into the clinical target volume for radiotherapy treatments. The directionality of tumor invasion was estimated using fluid flow simulations to obtain realistic muscle fiber trajectories. The method has the potential to improve the consistency and accuracy of the clinical target volume, thereby reducing the physician's time burden.

References

[1] A. Gazendam, S. Popovic, N. Parasu, et al. "Chondrosarcoma: A Clinical Review". *Journal of Clinical Medicine* 12.7 (Mar. 2023), p. 2506. DOI: 10.3390/jcm12072506.

[2] L. Beunk, K. Brown, I. Nagtegaal, et al. "Cancer invasion into musculature: Mechanics, molecules and implications". *Seminars in Cell & Developmental Biology* 93 (Sept. 2019), pp. 36–45. DOI: 10.1016/j.semcdb.2018.07.014.

[3] G. G. Handsfield, B. Bolsterlee, J. M. Inouye, et al. "Determining skeletal muscle architecture with Laplacian simulations: a comparison with diffusion tensor imaging". *Biomechanics and Modeling in Mechanobiology* 16.6 (June 2017), pp. 1845–1855. DOI: 10.1007/s10237-017-0923-5.

[4] J. Varvik, T. F. Besier, and G. G. Handsfield. "Computational fluid dynamics simulations for 3D muscle fiber architecture in finite element analysis: Comparisons between computational fluid dynamics and diffusion tensor imaging". *International Journal for Numerical Methods in Biomedical Engineering* 37.12 (Sept. 2021). DOI: 10.1002/cnm.3521.

[5] G. Buti, A. Ajdari, K. Hochreuter, et al. "The influence of anisotropy on the clinical target volume of brain tumor patients". *Physics in Medicine & Biology* 69.3 (Jan. 2024), p. 035006. DOI: 10.1088/1361-6560/ad1997.

[6] N. Schlömer. *pygalmesh: Python interface for CGAL's meshing tools*. 2021. DOI: 10.5281/ZENODO.5564818.

[7] I. A. Baratta, J. P. Dean, J. S. Dokken, et al. *DOLFINx: The next generation FEniCS problem solving environment*. en. 2023. DOI: 10.5281/ZENODO.10447666.

[8] T. Bortfeld and G. Buti. "Modeling the propagation of tumor fronts with shortest path and diffusion models—implications for the definition of the clinical target volume". *Physics in Medicine and Biology* 67.15 (July 2022), p. 155014. DOI: 10.1088/1361-6560/ac8043.

[9] J.-M. Mirebeau and J. Portegies. "Hamiltonian Fast Marching: A Numerical Solver for Anisotropic and Non-Holonomic Eikonal PDEs". *Image Processing On Line* 9 (Feb. 2019), pp. 47–93. DOI: 10.5201/ipol.2019.227.

[10] G. B. (B. Arfken, H.-J. Weber, and F. E. Harris. *Mathematical methods for physicists : a comprehensive guide*. 7th ed. Elsevier, 2013.

[11] E. Garyfallidis, M. Brett, B. Amirbekian, et al. "Dipy, a library for the analysis of diffusion MRI data". *Frontiers in neuroinformatics* 8 (2014), p. 8.

[12] J. Wasserthal, H.-C. Breit, M. T. Meyer, et al. "TotalSegmentator: Robust Segmentation of 104 Anatomic Structures in CT Images". *Radiology: Artificial Intelligence* 5.5 (Sept. 2023). DOI: 10.1148/ryai.230024.

Characterization of Acute Radiation-Induced Vascular Changes in Brain Tumors Using Wavelet-Based Analysis of Dynamic-Contrast-Enhanced MRI Information

Hassan Bagher-Ebadian[1,2,3,4], Stephen L. Brown[1,2], Tavarekere N. Nagaraja[5,2], Prabhu Acharya[4], James R. Ewing[6,2,4], Indrin J. Chetty[7,4], Benjamin Movsas[1], and Kundan Thind[1,8,9]

[1]Department of Radiation Oncology, Henry Ford Cancer Institute, Detroit, USA, [2]Department of Radiology, Michigan State University, East Lansing, USA, [3]Department of Osteopathic Medicine, Michigan State University, East Lansing, USA, [4]Department of Physics, Oakland University, Rochester, USA, [5]Department of Neurosurgery, Henry Ford Health, Detroit, USA, [6]Department of Neurology, Henry Ford Health, Detroit, USA, [7]Department of Radiation Oncology, Cedars-Sinai Medical Center, Los Angles, USA, [8]Department of Medicine, Michigan State University, East Lansing, USA,[9]Department of Oncology, School of Medicine, Wayne State University, Detroit, USA

Abstract: The ability to better discern tumor from normal tissue has the potential to improve treatment efficacy. This study uses dynamic contrast enhanced (DCE)-based pathophysiological nested model-selection (NMS) technique along with wavelet-based spatiotemporal modeling of DCE-MRI information to characterize the RT response of tumor and normal brain in an animal model. Twenty-three immune-compromised-RNU rats were implanted orthotopically with human U251N glioma cells. 28 days after brain implantation, two DCE-MRI studies were performed 24hr apart. 20 Gy stereotactic radiation was delivered before the second MRI. NMS based DCE-MRI analysis was performed to distinguish different tissue regions. Time-traces of DCE-MRI information for different models were analyzed using a wavelet-based coherence analysis method to characterize and rank their RT-induced effects. Compared to the peritumoral tissues ($r=0.622$, $p<0.0001$), normal and highly permeable tissues ($r=0.650$ with $p<0.0001$ and $r=0.809$ with $p<0.0001$, respectively) are less impacted by RT. This study demonstrates a novel approach to discern peritumoral tissue from normal brain using acute changes in physiology in response to radiation therapy that has the potential to improve treatment planning and RT efficacy.

1 Introduction

An important issue affecting treatment efficacy in oncology is the ability to distinguish tumor from normal tissue. Recent studies [1-6] have shown that tumor and normal tissue vascular networks are characterized by different permeability and other physiological parameters. Furthermore, differential acute responses to treatment between tumor and normal tissue have potential to be exploited using decoded spatial and temporal signatures to facilitate and improve treatment planning.

In animal models, dynamic contrast enhanced (DCE)-MRI pharmacokinetic (PK) analysis shows promise as a stable descriptor of tumor physiology [1, 2], and of physiological reaction to treatments [1-6]. Our group has taken an important step toward the improvement of the image-based estimation of vascular permeability parameters by introducing a nested model selection (NMS) technique [7, 8].

This study combines a physiological NMS technique with a novel spatiotemporal modeling of wavelet-based information of DCE-MRI data, enabling evaluation of the RT-affected regions in an animal model of cerebral U251N tumors as well as the time-varying information content of the DCE-MR raw data on a voxel-by-voxel basis.

2 Materials and Methods

2.1 Animal Population, DCE-MR Studies, and Brain Irradiation: Twenty-three immune-compromised-RNU/RNU rats were stereotactically and orthotopically implanted with 5×10^5 U251N cancer cells (The U251N cells [9]were contributed by Dr. Tom Mikkelsen from Department of Neurosurgery, Henry Ford Health) in 10 µl of phosphate buffered saline (PBS) that formed a 3-4 mm diameter orthotopic glioma in two weeks post-implantation[10]. For each rat, two DCE-MRI studies (multi-slice, multi-echo gradient-echo sequence, with three 2.0 mm slices, no gap, matrix:128×64, FOV:32×32mm^2, $T_R/T_{E1}/T_{E2}$ = 24ms/2ms/4ms, flip-angle=18º, SW=150 kHz) were performed 24 hours apart using a 7T Varian (Agilent, 20cm bore system with Bruker console) scanner. Four-hundred slice packages at ~ 1.55s intervals (total experiment duration: ~10min) were acquired. A bolus of MR contrast-agent (MR CA-Magnevist) was injected (tail-vein) by hand push at acquisition 15. For each DCE-MRI experiment, two T1 mapping (TOMROP, T One by Multiple Read Out Pulses, or Look-Locker, LL) sequences [11, 12] were run to produce voxel-by-voxel maps of tissue T_1 prior to, and immediately after the Dual Gradient Echo (DGE) DCE-MRI study. All experiments in this study were performed in accordance with institutional regulations (IACUC#1509). A single 20Gy stereotactic-radiation exposure was delivered to rat brains before the second MRI which was acquired 1-6.5 hours after radiation.

2.2 Pharmacokinetic Analysis: We used the Standard Model (SM) for evaluating cerebral physiology [7, 8, 13] following measurement of contrast tracers using DCE-MRI experiment. A Tofts model [14, 15] was used to model the leakage of contrast agent (CA) in the vasculature system. After intravenous administration of CA, the CA concentration in tissue is given by:

$$C_t(t) = K^{trans} \int_0^t e^{-K_{ep}(t-\lambda)} C_p(\lambda) d\lambda + v_p C_p(t) \quad (1)$$

where: C_p and C_t are the plasma and tissue concentrations of the CA, K^{trans} is the volumetric forward transfer rate of the contrast agent into the interstitial space. The k_{ep} is the transfer rate from interstitial compartment to the vascular compartment, t and λ denote the experiment time, and v_p is the fractional volume of the vascular distribution space or plasma distribution space. We assumed that CA concentration, [Gd], is proportional to the change in the longitudinal relaxation rate (R_1) after CA administration:

[Gd] = CA(t) ~ $\Delta R_1(t)$, where $R_1 = 1/T_1$ (T_1 is the longitudinal relaxation time).

2.3 Nested Model Selection Technique: As shown in Figure 1, we have shown[7, 8] that three physiologically nested models (models 1, 2 and 3) can be derived from the standard Toft's model to describe possible pathophysiological conditions of underlying tissue pathology. We have generated a series of stable processing pipelines accordingly, for producing vascular parametric maps based on the NMS[7, 8]. As shown in this figure, NMS technique generated [16] three maps of brain regions labeled with the number of parameters used to describe the data: ***a)*** Model 1 region: normal vasculature with no leakage, the only parameter estimated is plasma volume, v_p; ***b)*** Model 2 region: leaky tissues with CA leakage without measurable back-flux to the vasculature, in which case, v_p and, K^{trans} can be estimated; or ***c)*** Model 3 region: leaky vessels with CA leakage and measurable back-flux and, thus, v_p, K^{trans}, and k_{ep}, or extracellular extra-vascular volume, v_e (ratio of K^{trans} and K_{ep}) can be estimated.

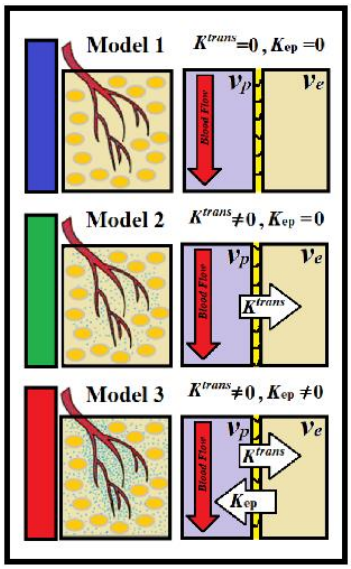


Figure-1: Three physiological Nested Models.

2.3 Nonstationary Time-Frequency Analysis: Normalized time-traces of DCE-MRI information (1st and 2nd echoes for pre, and post-RT cohorts) for three model regions extracted from all the rat brains were analyzed using a magnitude-squared wavelet-based coherence (MSWC) technique to characterize their time-frequency information before and after radiation. The MSWC technique is useful for analyzing nonstationary signals and recruits an analytic Morlet wavelet [17-19] to estimate the measures of the magnitude and phase correlations between two signals (reference and observation signals) in the time-frequency space. The wavelet cross-spectrum (WC) is a measure of the distribution of power of two signals (reference versus observation) [17-19] and for two time series time series, *R* and *O*, can be written as:

$$WC_{RO}(\Omega,\beta) = \vec{\varphi}(C_R^*(\Omega,\beta).C_O(\Omega,\beta)) \quad (2)$$

$C_R(\Omega, \beta)$ and $C_O(\Omega, \beta)$ denote the continuous wavelet transforms of signals R and O at scales Ω and positions β. The superscript * is the complex conjugate, and φ is a smoothing operator in time and scale domains. Therefore, the Wavelet coherence [17-19] is defined as the measure of the correlation between the two signals R, and O in the time domain as follows:

$$\frac{|\vec{\varphi}(C_R^*(\Omega,\beta).C_O(\Omega,\beta))|^2}{\vec{\varphi}(|C_R(\Omega,\beta)|^2).\vec{\varphi}(|C_O(\Omega,\beta)|^2)} \quad (3)$$

For each pathophysiological model, the pre- and post-RT signal coherence level was estimated using Equation 3. The coherence level of the time-frequency information (in the z space) was used to compare and rank the effect of radiation on different model regions corresponding to two different time-varying DCE-MRI raw signals (1st and 2nd echoes).

3 Results

An example of model selection map and its corresponding DGE (1st and 2nd echoes) images for a slice of a rat brain are shown in Figure-2. Figure 3A-F demonstrate amplitude and phase (black arrows) information changes within the cone of influence (certain area delineated by white dashed-line) of the two raw DCE-MRI signals (1st and 2nd echoes) due to the RT for different model regions (Subfigures 3A, 3B, and 3C for the 1st echo and 3D, 3E, and 3F, for the 2nd echoes) at different time points (10 sec/sample). Subfigures 4A-B illustrate the variation of the coherence level between the pre and post RT for the two raw-DCE-MRI signals (subfigures A and B for the 1st and 2nd echoes, respectively) for three different model regions (blue, green, and red for Models 1, 2, and 3 respectively).Figure 3A-F along with subfigures 4A-B demonstrate the magnitude-squared of coherence (RT affected coherence level) for normalized frequency at different timepoints for the two raw-DCE-MRI signals for three different model regions. These figures illustrate that the higher loss of coherence (RT-induced signal changes, r=0.622 with p<0.0001) belongs to the Model 2 regions compared to Models 1 and 3 (r=0.650 with p<0.0001 and r=0.809 with p<0.0001 for Models 1 and 3, respectively).

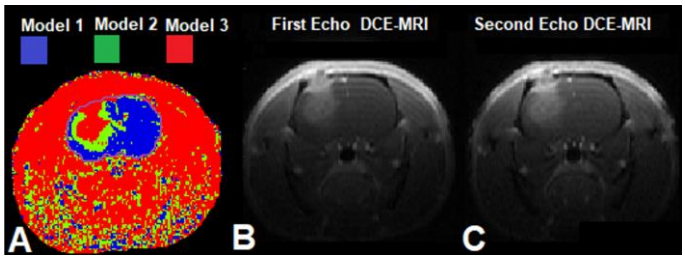


Figure-2: Subfigure 2A-C show model choice map, the 1st and 2nd echo images for a slice of a rat brain. The smooth muscles outside of the rat brain are classified as highly permeable tissues (Model 3, in red color).

4 Discussion

In this study, we assumed that CA concentration, [Gd], is proportional to the change in the longitudinal relaxation rate (R_1) after CA administration: [Gd] = CA(t) ~ $\Delta R_1(t)$, where $R_1 = 1/T_1$ (T_1 is the longitudinal relaxation time). This assumption is supported in most DCE-MRI studies because they are carried out under the conditions of rapid repetition time and Ernst tip-angle adjusted for a mean decrease in R_1. Under these conditions, studies in multicompartmental systems demonstrate little effect of exchange mechanisms [20, 21].

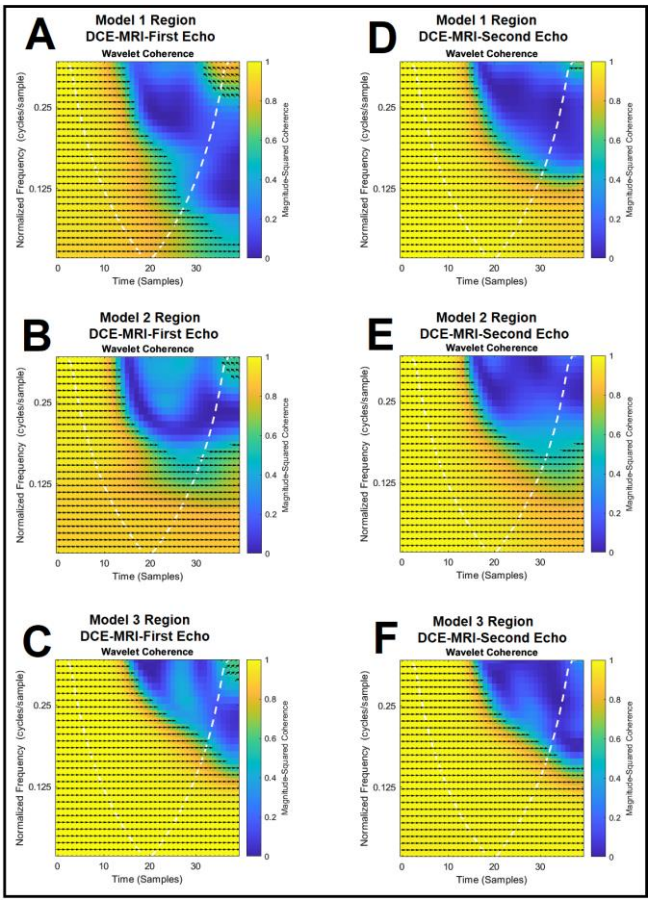


Figure-3: This figure shows the Magnitude-Squared of Coherence for normalized frequency at different timepoints (10 sec/sample) for the two raw-DCE-MRI signals for three different model regions: Subfigures 3A, 3B, and 3C for the 1st echo and 3D, 3E, and 3F, for the 2nd echoes.

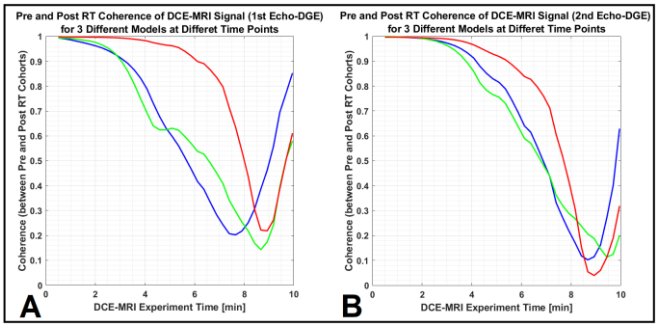


Figure-4: This figure shows the variation of the coherence level between the pre and post RT for the two raw-DCE-MRI signals (1st and 2nd echo signals) for three different model regions (blue, green, and red for Models 1, 2, and 3 respectively).

Figure 4A-B also illustrates that the highest RT-induced signal changes for all models occur during the 8 to 9 minutes of the DCE-MRI experiment (low-frequency components). As shown in Figure-4, compared to the 1st echo, the 2nd echo of the DCE-MRI information for all the three pathophysiological regions shows higher RT-induced changes at a later time (~ 9 min) of the DCE-MRI experiment. This is because that the 2nd echo signal is associated with the T_2* dephasing effect and directly affects the uncertainty level and heterogeneity content of the raw signal due to the local and static magnetic field inhomogeneity. Many studies have already shown that the DCE-MRI features of the U251N animal model tumors correlate with human Glioblastoma, including a necrotic center, poorly demarcated, infiltrative tumor borders, and an enhanced rim on T_2-weighted imaging (2nd echo in DGE exam). Our group is one of the few studying animal tumors using the DGE pulse sequences. In the paradigm, vascular permeability, and parametric maps, allow separation of T_1 and T_2* effects. Further, the raw information of these two components can be used to model the extravascular space of the tumor, which plays an important role in response to therapy. The cerebral physiology and mechanical characteristics exhibit a high degree of uniformity across various species, as established by previous research [22-24]. This study's analysis offers insightful revelations about the specific components and regions influenced by radiation within the context of DCE-MR imaging. Moreover, it underscores the significance of these findings in delineating the microenvironment of human tumors. The acquired data holds considerable promise for enhancing the evaluation and treatment strategies of lethal human tumors. It is imperative for the clinical research community to integrate this knowledge to deepen their understanding of the fundamental physiology of solid tumors. Consequently, replicating and refining this research using clinical pulse sequences (such as Spoiled Gradient-Recalled Echo, SPGRE with single echo) at varied temporal resolutions could greatly benefit clinical studies in radiation oncology, offering substantial translational potential.

5 Conclusion

This pilot study shows for the first time that wavelet-based radiomic analysis combined with an NMS technique can effectively decompose DCE-MRI information to rank the irradiated tumor regions and its surrounding normal tissues by their temporal MR signal changes. This is an important first step toward characterizing RT-affected brain regions and the RT-induced temporal changes in raw DCE-MR experiment which both are fundamental in treatment planning of cancer patients, design and optimization of DCE-MR imaging, and pharmacokinetic modeling.

References

[1] T. N. Nagaraja *et al.*, "Reproducibility and relative stability in magnetic resonance imaging indices of tumor vascular physiology over a period of 24h in a rat 9L gliosarcoma model," *Magn Reson Imaging*, vol. 44, pp. 131-139, Dec 2017, doi: 10.1016/j.mri.2017.09.003.

[2] M. P. Aryal *et al.*, "Intratumor distribution and test-retest comparisons of physiological parameters quantified by dynamic contrast-enhanced MRI in rat U251 glioma," *NMR Biomed*, vol. 27, no. 10, pp. 1230-8, Oct 2014, doi: 10.1002/nbm.3178.

[3] J. H. Kim, V. Y. Park, H. J. Shin, M. J. Kim, and J. H. Yoon, "Ultrafast dynamic contrast-enhanced breast MRI: association with pathologic complete response

in neoadjuvant treatment of breast cancer," *Eur Radiol,* Jan 22 2022, doi: 10.1007/s00330-021-08530-4.

[4] Y. Mori, N. Mori, T. Izumiyama, A. Inoue, K. Takase, and T. Aizawa, "Mathematical model for histogram analysis of dynamic contrast-enhanced MRI: A method to evaluate the drug treatment response in rheumatoid arthritis," *Eur J Radiol,* vol. 141, p. 109831, Aug 2021, doi: 10.1016/j.ejrad.2021.109831.

[5] T. Weikert *et al.*, "Early Prediction of Treatment Response of Neuroendocrine Hepatic Metastases after Peptide Receptor Radionuclide Therapy with (90)Y-DOTATOC Using Diffusion Weighted and Dynamic Contrast-Enhanced MRI," *Contrast Media Mol Imaging,* vol. 2019, p. 1517208, 2019, doi: 10.1155/2019/1517208.

[6] S. E. Heethuis *et al.*, "Dynamic contrast-enhanced MRI for treatment response assessment in patients with oesophageal cancer receiving neoadjuvant chemoradiotherapy," *Radiother Oncol,* vol. 120, no. 1, pp. 128-35, Jul 2016, doi: 10.1016/j.radonc.2016.05.009.

[7] J. R. Ewing and H. Bagher-Ebadian, "Model selection in measures of vascular parameters using dynamic contrast-enhanced MRI: experimental and clinical applications," *NMR Biomed,* vol. 26, no. 8, pp. 1028-41, Aug 2013, doi: 10.1002/nbm.2996.

[8] H. Bagher-Ebadian *et al.*, "Model selection for DCE-T1 studies in glioblastoma," *Magn Reson Med,* vol. 68, no. 1, pp. 241-51, Jul 2012, doi: 10.1002/mrm.23211.

[9] A. Torsvik *et al.*, "U-251 revisited: genetic drift and phenotypic consequences of long-term cultures of glioblastoma cells," *Cancer Med,* vol. 3, no. 4, pp. 812-24, Aug 2014, doi: 10.1002/cam4.219.

[10] C. Peng *et al.*, "Targeting orthotopic gliomas with renal-clearable luminescent gold nanoparticles," *Nano Res,* vol. 10, no. 4, pp. 1366-1376, Apr 2017, doi: 10.1007/s12274-017-1472-z.

[11] G. Brix, L. Schad, M. Deimling, and M. Lorenz, "Fast and precise T1 imaging using a TOMROP sequence," *Magn Reson Imaging,* vol. 8, no. 4, pp. 351-356, 1990, doi: DOI: 10.1016/0730-725x(90)90041-y.

[12] H. Bagher-Ebadian, R. Paudyal, T. Mikkelsen, Q. Jiang, and J. Ewing, "Rapid and Direct Quantification of Longitudinal Relaxation Time (T1) in Look-Locker Sequences Using an Adaptive Neural Network," *IEEE-IJCNN,* vol. 3, no. 4, pp. 236-241, 2009.

[13] R. Bai *et al.*, "Shutter-Speed DCE-MRI Analyses of Human Glioblastoma Multiforme (GBM) Data," *J Magn Reson Imaging,* vol. 52, no. 3, pp. 850-863, Sep 2020, doi: 10.1002/jmri.27118.

[14] P. Tofts and A. Kermode, "Measurement of the blood-brain barrier permeability and leakage space using dynamic MR Imaging. 1. Fundamental Concepts," *Magnetic Resonance in Medicine,* vol. 17, pp. 357-367, 1991.

[15] P. S. Tofts *et al.*, "Estimating kinetic parameters from dynamic contrast-enhanced T(1)-weighted MRI of a diffusable tracer: standardized quantities and symbols," (in eng), *J Magn Reson Imaging,* vol. 10, no. 3, pp. 223-32, Sep 1999, doi: 10.1002/(SICI)1522-2586(199909)10:3<223::AID-JMRI2>3.0.CO;2-S [pii].

[16] H. Bagher-Ebadian *et al.*, "Dynamic contrast enhanced (DCE) MRI estimation of vascular parameters using knowledge-based adaptive models," *Sci Rep,* vol. 13, no. 1, p. 9672, Jun 14 2023, doi: 10.1038/s41598-023-36483-9.

[17] A. Grinsted, J. C. Moore, and S. Jevrejeva, "Application of the cross wavelet transform and wavelet coherence to geophysical time series," *Nonlin. Processes Geophys.*, vol. 11, no. 5/6, pp. 561-566, 2004, doi: 10.5194/npg-11-561-2004.

[18] D. Maraun, J. Kurths, and M. Holschneider, "Nonstationary Gaussian processes in wavelet domain: Synthesis, estimation, and significance testing," *Physical Review E,* vol. 75, no. 1, p. 016707, 01/22/ 2007, doi: 10.1103/PhysRevE.75.016707.

[19] C. Torrence and P. J. Webster, "Interdecadal Changes in the ENSO-Monsoon System," *Journal of Climate,* vol. 12, pp. 2679-2690, August 01, 1999 1999, doi: 10.1175/1520-0442(1999)012<2679:Icitem>2.0.Co;2.

[20] H. A. Spencer RG, Ferretti JA, Weiss GH, "Spillover and Incomplete Saturation in Kinetic Measurements," *Journal of Magnetic Resonance,* vol. 101, no. 3, B, pp. 294-296, 1993.

[21] Bagher-Ebadian H. Paudyal R, Nagaraja TN, Panda S, Fenstermacher JD, Ewing JR,, "MRI Look-Locker Estimates of the Longitudinal Relaxation Rate Are Approximately Linear in Contrast Agent Tissue Concentration," *International Society for Magnetic Resonance in Medicine,* vol. 1, 16th Scientific Meeting & Exhibition, Toronto, CA., 2008.

[22] V. L. Jacobs, P. A. Valdes, W. F. Hickey, and J. A. De Leo, "Current review of in vivo GBM rodent models: emphasis on the CNS-1 tumour model," *ASN Neuro,* vol. 3, no. 3, p. e00063, Aug 3 2011, doi: 10.1042/AN20110014.

[23] E. Radaelli *et al.*, "Immunohistopathological and neuroimaging characterization of murine orthotopic xenograft models of glioblastoma multiforme recapitulating the most salient features of human disease," *Histol Histopathol,* vol. 24, no. 7, pp. 879-91, Jul 2009, doi: 10.14670/HH-24.879.

[24] M. Candolfi *et al.*, "Intracranial glioblastoma models in preclinical neuro-oncology: neuropathological characterization and tumor progression," *J Neurooncol,* vol. 85, no. 2, pp. 133-48, Nov 2007, doi: 10.1007/s11060-007-9400-9.

OpenTPS: an open-source treatment planning system to foster research and innovation.

A. Barragán-Montero[1], D. Dasnoy[2], S. Wuyckens[1], W. Zhao[2], E. Peeters[1], R. Schyns[1], M. Huet-Dastarac1, E. Löyen[2], G. Rotsart de Hertaing[2], G. Janssens[3], V. Hamaide[3], S. Deffet[3], K. Souris[3], E. Sterpin[1,4], B. Macq[2], J. A. Lee [1,2]

[1]Molecular Imaging Radiation Oncology (MIRO), UCLouvain, Brussels, Belgium
[2] ICTeam, UCLouvain, Louvain-La-Neuve, Belgium
[3] Ion Beam Applications S.A, Louvain-La-Neuve, Belgium
[4] Department of Radiation Oncology, KULeuven, Leuven, Belgium

Abstract Treatment planning systems (TPS) are an essential component for simulating and optimizing a radiation therapy treatment before administering it to the patient. However, TPS commercialized by private companies are often black-boxes when it comes to the implementation of most of their features (i.e. image registration, dose calculation and optimization algorithms). This hampers and delay researchers when working with these TPS for implementing and testing new ideas. We have developed OpenTPS, a new open-source software platform (opentps.org), to foster high quality research involving treatment planning for external beam radiotherapy, and, in particular, for proton therapy. It is designed to be a flexible and user-friendly platform for medical physicists, radiation oncologists, and other members of the radiation therapy and research community. It serves to create customized treatment plans for educational and research purposes, particularly for emerging delivery and optimization techniques such as FLASH, arc proton therapy, or beamlet free optimization, among others.

1 Introduction

Planning a radiotherapy treatment is a complex task, which requires the acquisition of advanced medical images and the use of sophisticated algorithms to analyze these images, simulate the effect of radiation on the patient (i.e. dose calculation) and optimize the radiation dose distribution. Today, the main treatment planning systems (TPS) are marketed by big companies, which charge up to several hundred thousand euros for a license of such software. They come with multiple features and a user-friendly GUI, but their big drawback is their low modularity, lack of transparency accessibility to the code. This difficulties their use in research environments, sometimes delaying or hampering research projects. This is the case, for example, of projects that attempt to study treatment planning for new dose delivery modalities, faster optimization algorithms, more precise treatments, etc. Additionally, small research groups and clinics that do not have access to these expensive TPS experience a greater difficulty in conducting relevant research in the field, further slowing scientific progress in the field.

To address this issue, multiple groups have contributed to the development of open-source tools for treatment planning over the past decades. Well-known platforms are matRad [1], FoCa [2], CERR [3], or some modules embedded in 3DSlicer [4]. Our group has also a long experience in open-source tools, with the development of REGGUI [5] for image-processing, MCSquare [6] for Monte Carlo dose calculation in proton therapy, and MIROpt [7] for robust optimization in proton therapy. While these tools were entirely developed in Matlab and/or C++, in 2020 we embarked on a project to merge these existing tools and develop our own open-source platform, OpenTPS, using the state-of-the-art and widely-available Python programming language. OpenTPS (opentps.org) is designed to serve as a flexible framework for both personal and collective research efforts. Its open-source code is publicly available under the Apache-2.0 license. The code is designed to be easily extended and customized. The software is organized into two packages: the Core package, which is a library that defines data classes, data processing methods, and IO methods, and the GUI package, which offers a graphical interface for viewing and interacting with the data. In the current release, the Core package includes a range of features that are essential for proton therapy treatment planning, but extension to conventional radiotherapy treatments is being addressed and will be included in the next release. Some of the key features available in the current Core package of OpenTPS include:

- Data management and processing: DICOM importing, exporting and processing patient data.
- Dose computation: computing the dose from proton therapy treatment plans using fast Monte Carlo simulations with MCsquare.
- Treatment planning: creating treatment plans for individual patients, including the ability to define beam angles, energy levels, spot spacing, etc.
- Treatment evaluation: performing quality assurance checks on treatment plans, including the ability to evaluate robustness, compare doses, and assess the conformity of the treatment plan to the patient's anatomy.
- Data augmentation: creating artificial variations of patient data for robustness evaluation and for machine learning applications.

In the next section, we detail these tools and give an example of their use for two use cases: 1) implementation and evaluation of advanced optimization techniques of arc proton therapy, and 2) generation of synthetic data to simulate and evaluate different scenarios of tumor motion.

2 Materials and Methods

2. 1. GUI

OpenTPS comes with an optional graphical user interface (GUI) for visualization of 3D images, dose maps, contours

and dynamic images. Several other interactive tools are available: profile plots, visualization of DVH, segmentation, etc. The implementation is based on the PyQt, VTK, and PyQtGraph libraries, which are widely recognized for their capabilities in data visualization and analysis in scientific and medical applications. Figure 1 shows a screenshot of the GUI loaded with a head-and-neck patient geometry and the optimized treatment dose associated to the case. In addition, graphical modules have been implemented to design and optimize a treatment plan. The entire GUI is extensible and modifiable by the user. Examples of extensions can be found in the OpenTPS documentation.

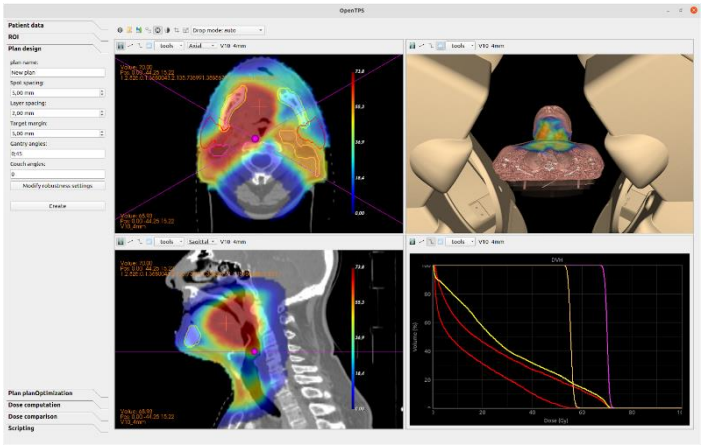


Figure 1. GUI with a head-and-neck cancer patient and an intensity-modulated-proton-therapy dose distribution.

2. 2. Core package

Data management and processing

The current version of OpenTPS supports the import patient data in DICOM format (CT, RT structures, RT dose, RT ion plan and vector field), MHD format (3D images), and OpenTPS-specific serialized object format. DICOM and MHD export is also supported.

The core library of OpenTPS includes a range of common image processing methods: resampling, coordinate sys- tem conversion, affine transformations, and filtering. Additionally, it has segmentation options such as automatic segmentation for body, lung and bones in CT images.

Rigid and deformable image registration methods are available, including the diffeomorphic Morphons algorithm [8]. The morphons registration is also used to create motion models. Deformation fields between the 4DCT phases are computed. Then with the deformations from all 4DCT phases, the midposition (MidP) image [9] can be computed. A motion model, called Dynamic3DModels in OpenTPS, is composed of the MidP image and the various deformation fields from the MidP image to each of the 4DCT phases.

Digitaly reconstructed radiographs (DRR) can also be computed using a 3DCT to simulate in room x-ray devices.

Dose calculation

A stable release of the fast Monte Carlo for proton therapy, MCsquare, is included in OpenTPS. However, OpenTPS is designed with the flexibility to support multiple dose engines and we are currently working on the implementation of photon dose calculation algorithms, both Monte Carlo and analytical.

MCSquare is available via the class MCsquareDoseCalculator and can be used to compute dose distributions, dose influence matrices, and LET distributions. Input paramaters such as the number of primaries, the beam model and the CT calibration are required to compute a dose based on CT and a treatment plan. From those inputs, MCsquareDoseCalculator generates the files required by MCsquare. Then, MCsquare simulates the interaction of each particle in batches until it reaches a user-defined minimal statistical uncertainty, or it simulates all the primaries.

Plan optimization

The treatment planning module of OpenTPS provides a set of tools and functions that are used to create and optimize a treatment plan based on several inputs. In general, the method followed depends on the treatment delivery modality chosen. As OpenTPS is initially built for research in proton therapy, the implementation of the treatment planning module conform to IMPT delivery, but implementation of IMRT and VMAT is being investigated. In proton pencil beam scanning (PBS), the dose is painted using scanning magnets that steer each pencil beam to a specific location in the target volume, namely a spot. Dividing the tumor into iso-energy layers, the treatment is therefore delivered by irradiating the volume spot-by-spot and layer-by-layer. This process is repeated for each irradiation direction. The modulation of the spot intensity is the aim of plan optimization. The optimal spot intensities result from solving a optimization problem where the user has built a cost function based on dose-volume objectives for the target volume and organs at risk (Figure 2).

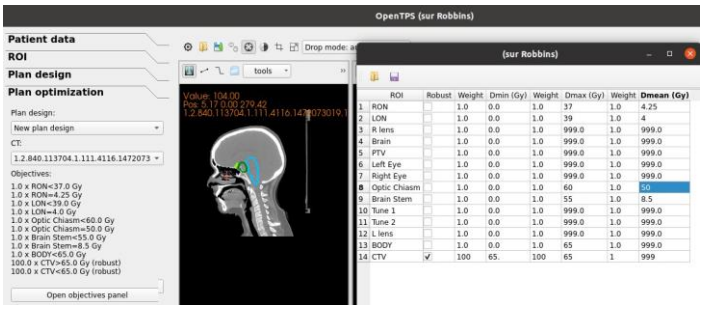


Figure 2. Objectives panel of the OpenTPS GUI.

OpenTPS comes with several built-in optimizers that solve the above problem. It is also designed so that other optimizers may be easily implemented. A variety of iterative solvers are provided: quasi-Newton's methods from Scipy, in-house implementation of LBFGS algorithms [10], gradient-based methods such as the classical gradient descent and the FISTA algorithm [11] from PyUNLockBox [12]. FISTA solves non-differentiable convex optimization problems by proximal splitting. For example, it can be of particular interest for ArcPT treatment plan optimization where one wants to maximize the spot sparsity by adding a l1-

norm penalty to the total objective function. Linear programming [13] methods are also available and can be used to solve the optimization problem. Based on the list of objectives in the plan design, a model is built with linear equations instead of the typical quadratic version. The model and the solver are built on the Gurobipy package [14], a Python interface to the powerful Gurobi solver, which requires a specific license. Gurobi includes a variety of algorithms to solve linear programming problems.

Treatment evaluation
In OpenTPS, a dose-volume histogram (DVH) evaluation is performed to assess the quality of the treatment plan. For proton therapy, it is important to assess the robustness of the plan to treatment uncertainties, thus robustness evaluation must also be performed on different uncertainty scenarios and not only on the nominal case, leading to a DVH-band instead. Two strategies are available depending on the scenarios definition. The first method, which is based on good practice rules and is commonly used in clinical routine, involves verifying the dose in the error space for a small number of extreme scenarios [15]. In OpenTPS, the dose is re-computed with MCsquare for 81 scenarios resulting from a combination of the setup and proton range erros. The second strategy generates a confidence interval in the so-called dosimetric space. It refers to all possible realizations of the treatment plan, or the different dose distributions that could be achieved. Described in [16], the scenarios are randomly sampled by MCsquare. The analysis in the dosimetric space reports the percentage of evaluation scenarios that meet a certain criterion such as minimum dose coverage for a target percentage (e.g., 95% coverage for 90% of scenarios).

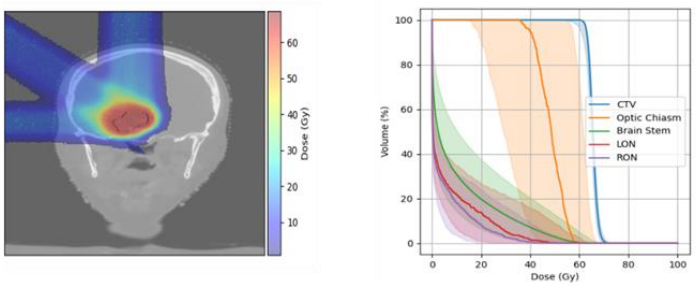


Figure 3. Example of a brain patient optimized in OpenTPS and robust evaluation of the plan (right) with the DVH-bands encompassing all uncertainty scenarios.

3 Results

Use case 1) Optimization in arc proton therapy.
Several solvers were implemented in OpenTPS to optimize arc proton therapy plans. Here, the arc proton problem is formulated as a MIP program and generates the best plan for a brain patient with only five upward energy switching (Figure 4). Details of the implementation and results can be found elsewhere [17].

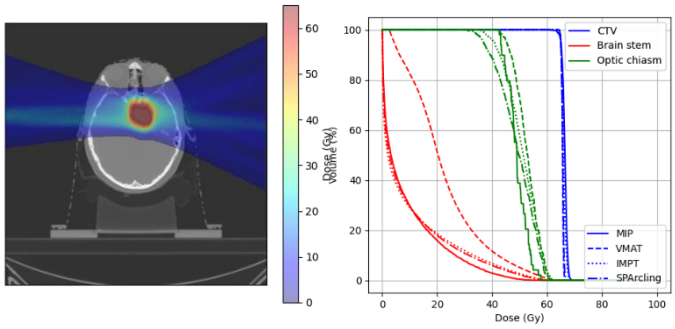


Figure 4. Left: Arc proton therapy dose distribution optimized in OpenTPS with Mixed-Integer Programming (MIP). Right: DVH comparison for Arc proton plans (in-house MIP optimizer, solid line and SPArc [18] technique re-implemented in OpenTPS, dashdot line), VMAT plan (dashed line) and conventional 2 lateral beam IMPT plan (dotted line).

Use case 2) Simulation of different tumor motion scenarios
Using a 4DCT based patient-specific breathing motion model, we can create sequences of variable length containing any desired variation of breathing pattern [19], as well as inter-fraction changes such as setup error simulations or target shrinks and baseline shifts. This way, the delivery of treatment plans can be simulated on these image sequences to assess the plan robustness against the chosen variation. This method of data augmentation is also used to train AI models [20].

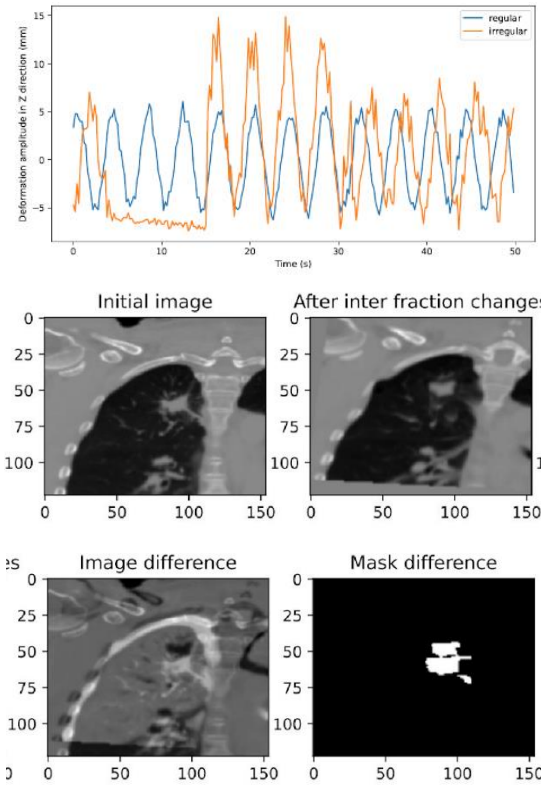


Figure 3. Example of synthetic breathing signals (regular and irregular), generated with OpenTPS tools (top), and example of synthetic inter-fraction deformations applied to and lung CT (bottom).

4 Discussion

OpenTPS is a unique platform for research in radiotherapy treatment planning, and particularly for proton therapy planning, in its current state. It is an open-source tool written in a user-friendly programming language. This allows for both internal contributions and international collaboration to enhance the platform quickly, making it increasingly sophisticated over time. Our research group is currently focusing on several exciting areas within OpenTPS, including arc proton therapy treatment planning and FLASH proton therapy. Depending on the needs of the users, new features might be added in the future such as (but not limited to) other treatment or image modalities, objects tracking or position prediction in dynamic sequences. Additionally, the platform has potential for real-time treatment delivery visualization and 2D to 3D image reconstruction.
The information to get started are available on our website (opentps.org) or on gitlab.

5 Conclusion

OpenTPS offers a flexible and dynamic platform for exploring new possibilities and pushing the boundaries of what is possible in radiation therapy and in proton therapy particularly.

6 Acknowledgements

OpenTPS has been funded by a Proof-of-Concept (POC) grant from the Walloon Region of Belgium.

References

1. Wieser H-P, Cisternas E, Wahl N, Ulrich S, Stadler A, Mescher H, et al. Development of the open-source dose calculation and optimization toolkit matRad. Med Phys. 2017;44:2556–68. http://dx.doi.org/10.1002/mp.12251

2. Sanchez-Parcerisa D, Udías J. Teaching treatment planning for protons with educational open-source software: experience with FoCa and matRad. J Appl Clin Med Phys. 2018;19:302–6. http://dx.doi.org/10.1002/acm2.12326

3. Deasy JO, Apte AP. A Computational Environment for Radiological Research. https://cerr.github.io/CERR/

4. SlicerRT. https://slicerrt.github.io/

5. https://openreggui.org/

6. Souris K, Lee JA, Sterpin E. Fast multipurpose Monte Carlo simulation for proton therapy using multi- and many-core CPU architectures. Med Phys. 2016;43:1700. http://dx.doi.org/10.1118/1.4943377

7. Barragán Montero AM, Souris K, Sanchez-Parcerisa D, Sterpin E, Lee JA. Performance of a hybrid Monte Carlo-Pencil Beam dose algorithm for proton therapy inverse planning. Med Phys . 2018;45:846–62. https://aapm.onlinelibrary.wiley.com/doi/abs/10.1002/mp.12688

8. Janssens G, Jacques L, Orban de Xivry J, Geets X, Macq B. Diffeomorphic registration of images with variable contrast enhancement. Int J Biomed Imaging. 2011;2011:891585. http://dx.doi.org/10.1155/2011/891585

9. Wolthaus JWH, Sonke JJ, van Herk M, Damen EMF. Reconstruction of a time-averaged midposition CT scan for radiotherapy planning of lung cancer patients using deformable registration. Med Phys . 2008;35:3998–4011. http://dx.doi.org/10.1118/1.2966347

10. Byrd RH, Lu P, Nocedal J, Zhu C. A Limited Memory Algorithm for Bound Constrained Optimization. SIAM J Sci Comput . 1995;16:1190–208. https://doi.org/10.1137/0916069

11. Beck A, Teboulle M. A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems. SIAM J Imaging Sci . 2009;2:183–202. https://doi.org/10.1137/080716542

12. Defferrard M, Pena R, Perraudin N. PyUNLocBoX: Optimization by Proximal Splitting. doi. 2017;

13. Dantzig GB, Thapa MN. Linear Programming 2: Theory and Extensions . Springer Science & Business Media; 2006. https://play.google.com/store/books/details?id=kWUKBwAAQBAJ

14. Gurobi Optimization LLC. Gurobi Optimizer Reference Manual (Gurobi Optimization, LLC). 2022;

15. Sterpin E, Rivas ST, Van den Heuvel F, George B, Lee JA, Souris K. Development of robustness evaluation strategies for enabling statistically consistent reporting. Phys Med Biol. 2021;66:045002. http://dx.doi.org/10.1088/1361-6560/abd22f

16. Souris K, Barragan Montero A, Janssens G, Di Perri D, Sterpin E, Lee JA. Technical Note: Monte Carlo methods to comprehensively evaluate the robustness of 4D treatments in proton therapy. Med Phys . 2019;46:4676–84. http://dx.doi.org/10.1002/mp.13749

17. Wuyckens S, Saint-Guillain M, Janssens G, Zhao L, Li X, Ding X, et al. Treatment planning in arc proton therapy: Comparison of several optimization problem statements and their corresponding solvers. Comput Biol Med. 2022;148:105609. http://dx.doi.org/10.1016/j.compbiomed.2022.105609

18. Ding X, Li X, Zhang JM, Kabolizadeh P, Stevens C, Yan D. Spot-Scanning Proton Arc (SPArc) Therapy: The First Robust and Delivery-Efficient Spot-Scanning Proton Arc Therapy. Int J Radiat Oncol Biol Phys. 2016;96:1107–16. http://dx.doi.org/10.1016/j.ijrobp.2016.08.049

19. Dasnoy-Sumell D, Aspeel A, Souris K, Macq B. Locally tuned deformation fields combination for 2D cine-MRI-based driving of 3D motion models. Phys Med. 2022;94:8–16. http://dx.doi.org/10.1016/j.ejmp.2021.12.010

20. Loÿen E, Dasnoy-Sumell D, Macq B. 3DCT Reconstruction from a Single X-Ray Projection Using Convolutional Neural Network. 2022 IEEE International Conference on Image Processing (ICIP) . IEEE; 2022. p. 1111–5. http://dx.doi.org/10.1109/ICIP46576.2022.9897902.

GATE 10: A new versatile Python-driven Geant4 application for medical physics

N. Krah[1], L. Maigne[2], M. Favaretto[3], A.F. Resch[3], C. Trigila[4], G. Mummaneni[4], E. Roncali[4], M. Jacquet[1], T. Baudier[1], A. Pereda[2], H.Fuchs[5], A.Coussat[6], and D. Sarrut[1]

[1]CREATIS, CNRS, Centre Léon Bérard, INSA Lyon, France
[2]LPC, CNRS, University of Clermont Auvergne, France
[3]MedAustron Ion Therapy Centre, Wiener Neustadt, Austria
[4]University of California at Davis, USA
[5]Medical, University of Vienna, Vienna, Austria
[6]Jagiellonian University in Kraków, Poland

Abstract In this contribution, we present GATE 10, the new version of the long-standing Geant4 application GATE, dedicated to medical physics, e.g. radiation imaging and radiation therapy systems. It features a Python-based user interface and integration of artificial intelligence methods. Moreover, being a Python package, GATE 10 can be easily integrated into external software, e.g. as a Monte Carlo dose engine. The user interface is easy to use, yet provides access to the rich functionality of Geant4 and preconfigured components of GATE (e.g. actors, sources).
GATE 10 is currently under rapid development and a first non-beta release is planned for May 2024. The software is open-source and available via the GitHub platform. Installation is performed via the Python package manager (`pip`). In this contribution, we present an overview of GATE 10 and a selection of new features and improvements.

1 Introduction

Since the beginning of GATE almost 20 years ago, users configure a simulation via macro commands grouped in text files [1, 2] similar to those in Geant4 [3]. Since then, the rise of high-level programming languages such as Python and new machine learning approaches have revolutionized scientific computing. The new version GATE 10 features a Python-based user interface and integration of artificial intelligence methods. Moreover, GATE 10 can be easily integrated as Python package into external software, e.g. as a Monte Carlo dose engine.

We have designed and implemented GATE 10 completely from scratch building onto the legacy of previous versions. On the one hand, it retains most of the features contributed to the code over time. This includes actors as well as preconfigured particle sources. Actors are collections of user actions to hook into Geant4's particle tracking to extract physics quantities, e.g. particle distributions, energy deposit, etc. GATE's particles sources are preconfigured primary particle generators including spatially distributed sources inside a patient for the use in nuclear medicine and imaging. On top of that, GATE 10 includes numerous new features and improvements, of which we will present a selection in this contribution.

2 Design principle of GATE 10

We have been developing GATE 10 with the following goals in mind:

1. Ease of use: The user should be able to concentrate on configuring the simulation rather than worrying about technical aspects of the software.
2. The Python script through which the simulation is set up should appear like an input file with the advantage of having access to all the computational tools available in Python and Python libraries (e.g. `numpy` [4]).
3. Give the user access to the rich Geant4-functionality, with a focus on applications in medical physics and medical imaging.
4. Provide an inner code structure with interfaces within which contributors can implement new features.
5. Keep a reasonable resemblance to the logic of older versions of GATE, i.e. version 9 and older.
6. Support a multitude of platforms including Linux, MacOS, and Windows.
7. Provide functionality to facilitate the logistics of running GATE simulations.
8. Do not compromise the simulation speed or even improve simulation speed compared to older versions of GATE.

3 Code structure

GATE 10 is composed of a part implemented in C++ and a part implemented in Python. The C++ library includes functionality that has a relevant impact on computational speed, i.e. mainly functions frequently called during particle tracking, e.g. after each step or at the beginning of each new track. This mainly concerns scoring (actors) and primary particle generation (sources). By implementing this part of GATE 10 in C++, Geant4 does not need to execute callbacks into Python code which would be associated with a computational overhead. Together with Geant4 classes and

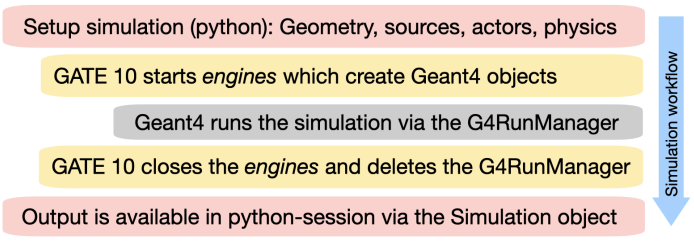


Figure 1: Workflow of a GATE 10 simulation from set-up to simulation run.

functions, the C++-portion of GATE 10 is exposed to Python via *pybind11* [5].

The pure Python part of GATE 10 provides a user interface to build simulations and an internal interface to Geant4. In particular, GATE 10 has two distinct kinds of classes which handle a simulation. *Managers* provide an interface to the user to set-up and configure a simulation and collect and organize user parameters in a structured way. *Engines* provide the interface to Geant4 and are responsible for creating all Geant4 objects. Managers and engines are divided into sub-managers and sub-engines responsible for certain logical parts of a simulation, such as geometry, physics, primary particle generation (sources), and scoring (actions, actors). Additionally, many components in GATE are implemented as classes which provide interfaces to the managers and engines. Figure 1 shows a scheme of the simulation workflow. The user sets up a simulation starting with the `Simulation` class, the main manager. It collects general parameters, e.g. about verbosity and visualization, and it manages the way the simulation is run (e.g., in a subprocess or not). Sub-managers are: `VolumeManager`, `PhysicsManager`, `ActorManager`, `SourceManager`. These managers can be thought of as bookkeepers. For example, the VolumeManager keeps a dictionary of all the volumes added to a simulation, a dictionary of all the parallel world volumes, etc. Furthermore, it also provides the user with methods to perform certain tasks, e.g. `VolumeManager.add_parallel_world()`.

The `SimulationEngine` is the main driver of the Geant4 simulation. Given a simulation object `sim`, it creates a new `SimulationEngine` every time the user calls `sim.run()`, which in turn creates all the sub-engines: `VolumeEngine`, `PhysicsEngine`, `SourceEngine`, `ActorEngine`, `ActionEngine`. The method `SimulationEngine.run_engine()` actually triggers the construction and run of the Geant4 simulation. It takes the role of the `main.cc` file in a pure Geant4 simulation. When the Geant4 simulation has terminated, `SimulationEngine.run_engine()` returns the simulation output which is then collected by the `Simulation` object.

GATE classes representing a Geant4 object (or multiple Geant4 objects combined as, e.g., in the Volume classes) are designed to do multiple things:

- Store user parameters. Example: the `Region` class holds the user parameters `user_limits`, `production_cuts`, and `em_switches`.
- Provide interface functions to the user and to manager classes to configure the object or inquire about it. Examples: `Region.associate_volume()`, `Region.need_step_limiter()`
- Provide interface functions to the engines to handle the Geant4 objects, such as `initialize()` and `close()`.
- Provide convenience functionality such as yielding the object's configuration as a dictionary (`to_dictionary()`), reloading it from a dictionary (`from_dictionary()`), dumping info about the object (e.g. `Region.dump_production_cuts()`), and copying the configuration from another object (`volume2.copy_user_info(volume1)`).
- Handle technical aspects such as serialization via pickling (e.g. for subprocesses dispatching) in a unified way.

4 Example: Dose calculation

In this section, we present a simple example to illustrate the use of GATE 10. The simulation consists of a patient parametrised via a CT image and a conversion table from CT numbers (in Hounsfield units) to material, a proton beam source, and a dose actor which scores dose on a voxel grid attached to the patient geometry. Secondary electrons are only produced inside the patient if their range is greater than 3 mm. The option `start_new_process=True` lets GATE 10 run the Geant4 simulation in a separate process. In this way, the simulation can be run multiple times in a loop, each time with a different beam energy. Dispatching to a subprocess also allows the user to run the same simulation multiple times in an interactive Python session, such as *ipython* or *jupyter*.

```
import opengate as gate
sim = gate.Simulation()
mm = gate.g4_units.mm
MeV = gate.g4_units.MeV
patient = sim.add_volume("Image", name="patient")
patient.image = "patient-4mm.mhd"
patient.mother = "world"
patient.voxel_materials = [
    [-2000, -900, "G4_AIR"],
    [-900, -100, "Lung"],
    [-100, 0, "G4_ADIPOSE_TISSUE_ICRP"],
    [0, 300, "G4_TISSUE_SOFT_ICRP"],
    [300, 800, "G4_B-100_BONE"],
    [800, 6000, "G4_BONE_COMPACT_ICRU"]]
patient.set_production_cut(
    particle_name="electron",
    value=3 * mm)
src = sim.add_source("GenericSource", name="p_src")
src.particle = "proton"
src.position.type = "sphere"
src.position.radius = 10 * mm
src.position.translation = [0, 0, -140 * mm]
src.n = 1e4 # number of primaries
src.direction.type = "momentum"
src.direction.momentum = [0, 0, 1]
dose = sim.add_actor("DoseActor", "dose")
```

```
dose.mother = "patient"
dose.size = [99, 99, 99]
dose.spacing = [2 * mm, 2 * mm, 2 * mm]
dose.img_coord_system = True
dose.translation = [2 * mm, 3 * mm, -2 * mm]
for en in [130, 140, 150, 160]:
    src.energy.mono = en * MeV
    dose.output = f"patient_dose_energy_{en}.mhd"
    sim.run(start_new_process=True)
```

5 Memory management between Python and Geant4

In this section, we describe a rather technical detail about the combination of Geant4 and Python which is, however, crucial to make simulations work correctly. GATE 10 creates certain kinds of Geant4 objects within code on the Python-side while other Geant4 objects are created on the C++-side, in particular by Geant4 itself. To avoid memory leakage, Geant4 objects need to be deleted at the end of a simulation, either by the C++-side, usually meaning Geant4, or by the garbage collector on the Python-side. This requires some understanding of how lifetime is managed in Geant4. GATE 10 uses the `G4RunManager` to initialize, run, and deconstruct a simulation. The `G4RunManager`'s destructor triggers a nested sequence of calls to the destructors of many objects handled by the run manager, e.g. geometry and physics. Volumes, for example, are created through code on the Python-side, but they should never be deleted on the Python-side because the `G4RunManager` is responsible for deletion. Segmentation faults occur if the objects no longer exist at the time of supposed destruction triggered by the `G4RunManager`. This is achieved by correct configuration of the wrapping in *pybind11*.

There is also an important aspect on the Python-side concerning Geant4 objects that are deleted by the `G4RunManager`: Python usually stores a reference to the Geant4 objects whose creation is implemented on the Python-side. This reference will not find the object anymore once the `G4RunManager` has deleted it upon its own destruction. Therefore, we have implemented a mechanism that makes sure that references to Geant4 objects are unset before the `G4RunManager` is garbage collected (and thus its destructor is called).

GATE 10 is now very robust in terms of memory handling and segmentation faults.

6 Novel features

GATE 10 includes several novel features compared to previous versions. We briefly describe a selection of them in this section.

6.1 Native Geant4 multithreading

As a new feature in GATE 10, native Geant4 multithreading capability is available (for most of the functionalty so far). The user simply sets the number of threads, e.g. `sim.number_of_threads=4` and the `SimulationEngine` automatically sets up the `G4MTRunManager` appropriately.

6.2 Export to structured file, e.g. JSON

A simulation can be exported as a structured plain-text file in JSON format. This JSON file can be reloaded at a later stage to re-create the simulation. The user can choose, via an option, whether such as JSON file is automatically created and stored alongside the simulation output:

```
import opengate as gate
sim = gate.Simulation()
sim.store_json_archive = True
sim.store_input_files = True
sim.json_archive_filename = "simu_iccr.json"
sim.output_dir = "output"
...
sim.run()
```

The above example creates a JSON file in a folder called *output*, and alongside with it also stores a copy of all potential input files (images, material descriptions, etc.). The JSON file is structured, human-readable, and compact in size. The simulation can be re-run later by:

```
import opengate as gate
sim = gate.Simulation()
sim.from_json_file("output/simu_iccr.json")
sim.run()
```

This feature helps a lot in data analysis and is particularly interesting in applications where simulation output needs to be archived, such as clinical dose calculation. It can also be helpful for dispatching simulations in a cluster environment and for debugging. The functionality could be easily extended to other file formats such as XML.

6.3 Generative Adversarial Networks for phase-space generation

GATE 10 gives the user the possibility to generate particles via a generative adversarial network (GAN) instead of using, e.g., a conventional phase space file. The idea behind this approach is that particles are not tracked from the sources on, but generated at a later stage, thereby skipping part of the simulation geometry [6–8]. The GAN is also more convenient to store than typically huge phase space files. GATE 10 relies on the `torch` library and provides functionalities to train and apply the neural network.

6.4 User-friendly repeated volumes

Many simulations in medical imaging require geometries which contain many repetitions of the same kind of volume in space. For example, a detector head could consist of a grid of crystal segments where all segments share the same shape, size, and physical properties. Geant4 allows volumes to be

present in multiple locations in space by associating multiple G4PhysicalVolumes to the same G4LogicalVolume. Previous versions of GATE already allowed the user to repeat volumes, but GATE 10 now exposes this feature in a simple manner to the user, as can be seen in the following example snippet:

```
crystal = sim.add_volume("Box", "crystal")
crystal.size = [1 * cm, 1 * cm, 1 * cm]
crystal.translation = [
    [1 * cm, 0 * cm, 0],
    [0.2 * cm, 2 * cm, 0],
    [-0.2 * cm, 4 * cm, 0],
    [0, 6 * cm, 0]]
crystal.material = "LYSO"
```

Instead of setting translation to a single vector to define the placement of the crystal volume, we provide a list of translation vectors. GATE 10 automatically creates the correct set of G4PhysicalVolumes, thereby effectively replicating the volume 4 times in space. Similarly, a list of rotation matrices can be set instead of a single matrix, to inform GATE 10 how each copy of the volume should be rotated.
Multiple utility functions are available to automatically generated sets of translation vectors and/or rotation matrices for common geometric dispositions, such as cartesian grids and circular arrangements, e.g. for PET simulation.

6.5 Voxelized representation of the geometry

The user can generate a voxelized representation of the simulation geometry or a portion thereof. The output is a label image and a lookup table where each voxel contains a label which corresponds to a volume name in the lookup table. Slices of such an image can be easily visualised with common Python libraries such as matplotlib [9]. Such an image helps checking and verifying a geometry or can be used in post-processing simulation data. The user can specify the spatial binning and the extent of the geometry to be voxelized. For the latter, it is possible to provide a list of volumes and GATE 10 will automatically determine the suitable extent which encompasses all volumes.

7 Discussion

GATE 10 is currently under fast development, with new features being added at least at a monthly frequency. The current version as of writing this is beta7, which is stable, fully usable, and which includes more than 80% of functionality offered by previous GATE versions. A first non-beta release is planned for May 2024 and will be presented at the ICCR conference.
Until then, we will refactor the actors and sources to make the code more efficient and the user experience more intuitive. We are also working on improving computation speed of the dose actor (and derived variants) in multithreading mode. Special physics such as optical surfaces are currently being implemented as well as tesselated volumes. Several institutions are already using GATE 10 for their software projects. For example, it is currently being integrated into the IDEAL 1.2 platform for independent dose calculation in light ion beam therapy [10].
Part of our development strategy is that every feature in GATE 10 must have a test case to ensure code integrity over time. Currently, more than 140 test simulations are run automatically each time the code is modified. These tests also serve as examples for users.

8 Conclusion

GATE 10 is open-source and available via the GitHub platform [11]. Installation is performed via the Python package manager (pip), which allows for a simple installation process with one single command. All major platforms, i.e. Linux, MacOS, and Windows are supported. The user is not required to manually install Geant4 as it ships pre-compiled with GATE 10. GATE 10 can be easily used by other software as an external library.

References

[1] S. Jan, D. Benoit, E. Becheva, et al. "GATE V6: A major enhancement of the GATE simulation platform enabling modelling of CT and radiotherapy". *Physics in Medicine and Biology* 56.4 (2011), pp. 881–901. DOI: 10.1088/0031-9155/56/4/001.

[2] D. Sarrut, M. Bardiès, N. Boussion, et al. "A review of the use and potential of the GATE Monte Carlo simulation code for radiation therapy and dosimetry applications". *Medical Physics* 41.6Part1 (May 2014), p. 064301. DOI: 10.1118/1.4871617.

[3] S. Agostinelli, J. Allison, K. Amako, et al. "GEANT4 - A simulation toolkit". *Nuclear Instruments and Methods in Physics Research, Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 506.3 (2003), pp. 250–303. DOI: 10.1016/S0168-9002(03)01368-8.

[4] *Numpy python package*. URL: https://numpy.org.

[5] *Pybind11*. URL: https://github.com/pybind/pybind11.

[6] D. Sarrut, N. Krah, and J. M. Létang. "Generative adversarial networks (GAN) for compact beam source modelling in Monte Carlo simulations". *Physics in Medicine & Biology* 64.21 (Oct. 2019), p. 215004. DOI: 10.1088/1361-6560/ab3fc1.

[7] D. Sarrut, A. Etxebeste, N. Krah, et al. "Modeling complex particles phase space with GAN for Monte Carlo SPECT simulations: a proof of concept". *Physics in Medicine & Biology* 66.5 (Mar. 2021), p. 055014. DOI: 10.1088/1361-6560/abde9a.

[8] A. Saporta, A. Etxebeste, T. Kaprelian, et al. "Modeling families of particle distributions with conditional GAN for Monte Carlo SPECT simulations". *Physics in Medicine & Biology* 67.23 (Dec. 2022), p. 234001. DOI: 10.1088/1361-6560/aca068.

[9] *Matplotlib python package*. URL: https://matplotlib.org.

[10] L. Grevillot, D. J. Boersma, H. Fuchs, et al. "The GATE-RTion/IDEAL Independent Dose Calculation System for Light Ion Beam Therapy". *Frontiers in Physics* 9 (Aug. 2021). DOI: 10.3389/fphy.2021.704760.

[11] *OpenGATE github repository*. URL: https://github.com/OpenGATE/opengate.

PortPy: An Open-Source Initiative for Radiotherapy Treatment Planning Optimization Including Benchmark Data and Algorithms and Integration with Commercial TPS

Gourav Jhanwar[1], Mojtaba Tefagh[2], Qijie Huang[1], Vicki Trier Taasti[3], Seppo Tuomaala[4], Saad Nadeem[1], and Masoud Zarepisheh[1]

[1]Department of Medical Physics, Memorial Sloan-Kettering Cancer Center, New York, NY, USA
[2]Sharif University of Technology, Tehran, Iran
[3]Department of Clinical Medicine - Danish Center for Particle Therapy, Aarhus University Hospital, Denmark
[4]Varian Medical Systems Inc, CA, USA

Abstract We have developed PortPy (Planning and Optimization for Radiation Therapy in Python), an innovative open-source software package designed to expedite the research and clinical translation of novel treatment planning techniques in radiation therapy. PortPy addresses the limitations of existing packages by offering several key features: 1) Interoperability with Commercial Systems: PortPy is designed to work seamlessly with a commercial treatment planning system, Eclipse. This interoperability allows for the objective evaluation of new techniques in comparison to the current best clinical practices. 2) Inclusion of Benchmark Dataset: The package includes a benchmark dataset, currently 50 lung patients, complete with clinical plans and all necessary data for treatment planning optimization, such as voxels, beamlets, and the dose influence matrix. This data is extracted from Eclipse using its API, facilitating comprehensive and realistic testing scenarios. 3) Advanced Benchmark Algorithms: PortPy encompasses benchmark algorithms capable of identifying "global" optimal solutions, using computationally intensive mixed integer programming, for various non-convex optimization problems encountered in treatment planning. This includes challenges such as Dose Volume Histogram (DVH) constraints, direct machine parameter optimization for Volumetric Modulated Arc Therapy (VMAT), and beam orientation optimization for Intensity Modulated Radiation Therapy (IMRT). PortPy enables the design, testing, and clinical validation of both classical optimization-based and cutting-edge AI-based treatment planning techniques. It promotes transparency, reproducibility, and community-driven development in the field of radiation therapy treatment planning.

1 Introduction

The importance of open-source packages in advancing research, promoting transparency, and encouraging community-driven development cannot be overstated. In the field of radiotherapy treatment planning, existing Matlab-based open-source packages like CERR [1] and MatRad [2] have made significant contributions. The rising popularity of Python as a free, open-source programming language has inspired the development of Python-based packages such as OpenTPS, focusing on proton treatment planning, and PyMedPhys, mainly aimed at machine and plan quality assurance (QA). However, a key limitation of these existing packages is their inability to objectively evaluate new treatment planning techniques against current clinical practices. This limitation stems from the difficulty in fully replicating clinical environments, such as detailed linear accelerator configurations and commercial dose calculation engines. Consequently, plans generated using novel techniques, such as AI-based automated treatment planning, cannot be objectively compared with clinical plans used in patient treatment. We seek to overcome this barrier with the development of PortPy (Planning and Optimization for Radiation Therapy in Python), a new Python-based open-source package. PortPy is currently compatible with the Varian Eclipse treatment planning system (TPS), and future expansions to other TPSs are planned. It is important to emphasize that PortPy can be utilized without access to any TPS. However, access to a TPS is necessary if a final evaluation within the TPS is desired. Furthermore, PortPy addresses the absence of a benchmark dataset and algorithms in current packages, providing a crucial tool for researchers to evaluate novel techniques.

Figure 1 illustrates the inspiration behind PortPy's development, drawing from successful open-source practices in the AI and computer science community. Open-source tools such as PyTorch and TensorFlow, alongside benchmark datasets like ImageNet, and benchmark algorithms like AlexNet, have revolutionized AI and data science. Our goal is to replicate this successful model, equipping researchers with comprehensive tools (PortPy integrated with commercial TPS and the popular 3DSlicer imaging package), a benchmark dataset (currently 50 curated lung patients with expert-selected beams, all necessary optimization data, and benchmark IMRT plans generated using our in-house automated planning system, ECHO), and benchmark algorithms (including Mixed Integer Programming [MIP] algorithms for global optimal solutions).

2 Materials and Methods

Figure 2 illustrates the PortPy design and its three main modules: "Data Management", "Plan Generation", and "Plan Evaluation", which are discussed in the next sections.

2.1 Data Management

This module provides access to the curated benchmark PortPy dataset, which currently comprises data from 50 lung patients.

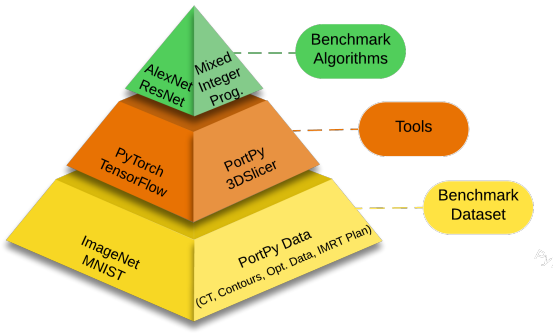


Figure 1: The development of PortPy has been inspired by the recent advancements in AI, facilitated by new tools, benchmark datasets, and algorithms. This analogy is depicted in this figure.

It allows researchers to test and develop their algorithms using the same dataset. The available data includes: 1) CT images and contours, 2) all necessary data for optimization for 72 uniformly selected beams (e.g., beamlets, voxels, dose influence matrix) extracted from Eclipse (version 16.1), 3) expert-selected beams for each patient, 4) an IMRT plan for each patient, generated using our in-house automated planning system, ECHO [3] (Expedited Constrained Hierarchical Optimization). ECHO is integrated with the Eclipse system and is currently in routine use in our clinic, having treated over 10,000 patients to date. It is important to note that the plans included in our dataset are not the clinical plans used for actual patient treatment, and therefore have not undergone the routine clinical scrutiny. However, they serve as valuable benchmark plans for research purposes.

2.2 Plan Generation

This module facilitates the generation of treatment plans using either classical optimization methods or emerging AI-based techniques. For optimization tasks, PortPy has been integrated with CVXPy [4], a widely-used open-source package. CVXPy enables the high-level formulation of optimization problems and offers out-of-the-box access to a range of of free (e.g., SCIP, SCIPY) and commercial (e.g., MOSEK, CPLEX) optimization engines (available for free for research purposes). Additionally, the PortPy.AI module is equipped with essential functionalities for AI-based planning. These include data access, data pre-processing, model training and testing, and patient-specific 3D dose prediction. The availability of both optimization and AI-based planning modules within PortPy allows researchers to not only compare these techniques but also explore their complementary aspects (Use Case 3, Section 3.3).

2.3 Plan Visualization and Evaluation

PortPy is equipped with basic built-in visualization tools, such as Dose-Volume Histograms (DVH) and dose distribution maps, as depicted in Figures 3, 4, 5. For more advanced visualizations, it integrates seamlessly with the popular open-source 3DSlicer package. When it comes to plan evaluation, PortPy provides a suite of methods to quantify optimized plans using clinically relevant metrics (for example, mean, maximum, DVH) and to compare these plans against established clinical criteria, as shown in Figure 4. Additionally, PortPy enables the export of plans in the DICOM RT plan format, allowing users to import these plans into any commercial TPS for final dose calculation and assessment. Figure5 illustrates how the results from PortPy align closely with Eclipse after importing the plan in Eclipse and running final dose calculation. It should be highlighted that although DICOM RT plan can be imported into various TPS, discrepancies are anticipated with TPSs other than Eclipse, due to the fact that our current data are derived from Eclipse, and there may exist variations in dose calculation methods between Eclipse and other TPSs. For IMRT, PortPy also supports exporting the optimal fluence map in a format compatible with Eclipse, facilitating the import into Eclipse for final leaf sequencing and dose calculation.
Furthermore, PortPy includes a collection of benchmark algorithms, useful for evaluating novel treatment planning optimization algorithms. These algorithms employ mixed integer programming (MIP) and are able to identify the best global optimal solutions for non-convex problems, including DVH constraints, VMAT, and beam angle orientations. While these algorithms are highly computationally demanding (typically taking days or weeks to solve) and not suited for real clinical application, they can be run on powerful computers and in the absence of time constraints to find global optimal solutions. These solutions then serve as benchmarks to validate new, computationally efficient algorithms (see Use Case 2, Section 3).

3 Results

In this section, we present three use cases to demonstrate the capabilities of PortPy. Additional examples are available on our GitHub page, provided in user-friendly Jupyter Notebook formats for ease of access and experimentation.

3.1 Use case 1: Developing and validating new treatment planning algorithms

The details of this simple and educational use case are provided in a Jupyter Notebook available on our GitHub page. To illustrate how researchers can develop new treatment planning techniques and validate them using PortPy, we performed the following tasks:

1. Utilizing the "Data Management" module, we loaded a lung patient case from the PortPy dataset, using the expert selected beams, which are also included in the dataset. We then employed the "Plan Generation" module and "Optimization" class to create a new IMRT plan using constraint optimization. We applied max/mean

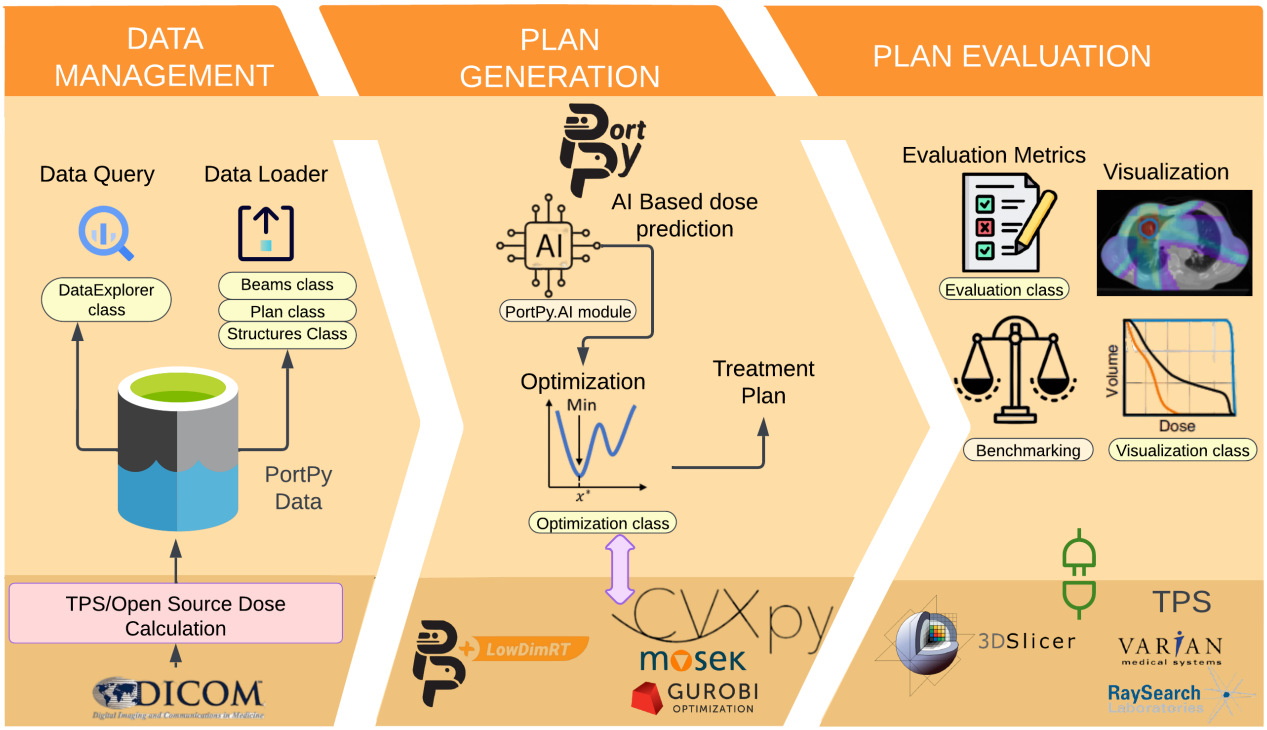


Figure 2: (PortPy workflow) PortPy comprises three primary modules: 1) "Data Management" for accessing the benchmark dataset, 2) "Plan Generation" for creating treatment plans using either classical optimization techniques or AI-based planning, and 3) "Plan Evaluation" for visualizing and assessing the generated plans.

hard constraints in line with our clinical criteria and developed a composite quadratic objective function, adjusting the weights manually. The integration with CVXPy facilitated the formulation of this optimization problem. Subsequently, we asked CVXPy to use the commercial optimization engine MOSEK (for which we obtained a free academic license) for solving the problem. Following this, we utilized PortPy's leaf sequencing algorithm to segment the resulting optimal fluence. Alternatively, the optimal fluence can be used without leaf sequencing.

2. The "Plan Evaluation" module and "Visualization" class were used to display the dose distribution and fluence profile of a beam in PortPy (see Figure 3). We then utilized the "Evaluation" class for a quantifiable assessment of the plan against our clinical criteria (refer to Figure 4).

3. We imported the DICOM RT plan file generated by PortPy, including the optimal control points, into Eclipse for the final dose calculation. Figure 5 shows that the results in Eclipse and PortPy are aligned, with minor discrepancies attributable to sampling inaccuracies and specific details regarding scattering and leaf transmissions.

4. We conducted a comparison of the plan with the PortPy IMRT benchmark plan (shown in Figure 6), which was generated using our in-house automated planning system, ECHO.

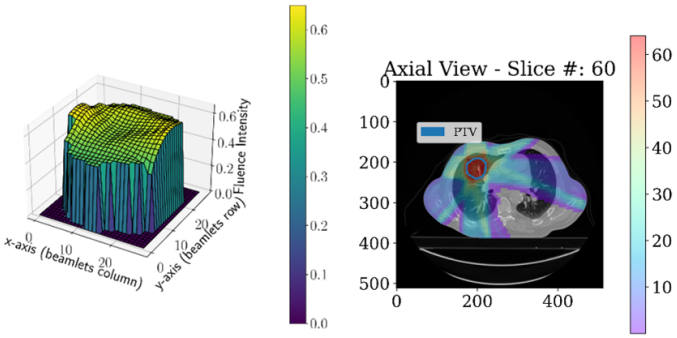


Figure 3: PortPy enables visualization of the generated plans, and for advanced visualizations, it offers integration with 3DSlicer and commercial treatment planning systems.

	constraint	structure_name	Limit	Goal	Plan Value
0	max_dose	GTV	69Gy	66Gy	65.88
1	max_dose	PTV	69Gy	66Gy	65.88
5	max_dose	HEART	66Gy		57.47
6	mean_dose	HEART	27Gy	20Gy	4.64
7	V(30Gy)	HEART	50%		1.00
8	V(30Gy)	HEART		48%	1.00
9	max_dose	LUNG_L	66Gy		65.88
10	max_dose	LUNG_R	66Gy		20.92
11	max_dose	CORD	50Gy	48Gy	10.50

Figure 4: PortPy allows for the evaluation of generated plans through quantifiable metrics, enabling comparisons with established clinical criteria. The color green is used to indicate areas where limits and goals are met.

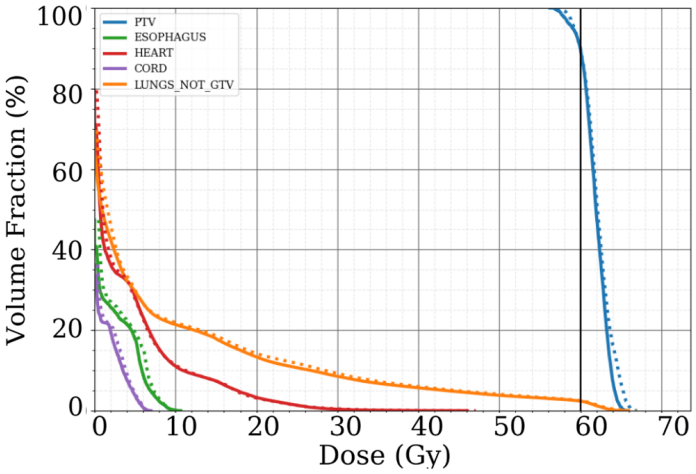


Figure 5: The DVH curves generated in PortPy (solid lines) for the optimized plan closely align with the DVH curves generated in Eclipse (dotted lines) following the final dose calculation

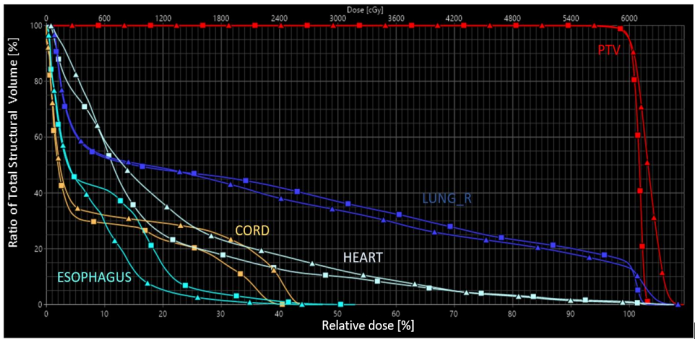


Figure 6: The plan generated in PortPy (lines with square marks) is imported into Eclipse and compared against the benchmark IMRT plan that is available in the PortPy dataset (lines with triangle marks). The benchmark plan in the dataset was generated using our in-house automated planning system, ECHO.

3.2 Use case 2: Bench-marking novel algorithms against the global optimal solutions

Certain non-convex optimization problems common in this field can be solved to global optimality through proper formulation using Mixed Integer Programming (MIP), which is notably computationally intensive. Figure 7 displays plans generated using both expert-selected beams, as provided in the dataset, and globally optimal beams. The global optimal beams were determined by solving the MIP-based problem, a process that required over four days of computation on our system. These plans can act as benchmarks for validating new, more computationally efficient algorithms for beam orientation optimization. Additionally, PortPy includes global optimal solutions for both VMAT and DVH constraints.

3.3 Use case 3: AI-based automated planning

The PortPy.AI module facilitates the end-to-end development and validation of innovative AI-based planning techniques. For instance, it enables the training of deep learning models using the PortPy benchmark dataset and benchmark IMRT

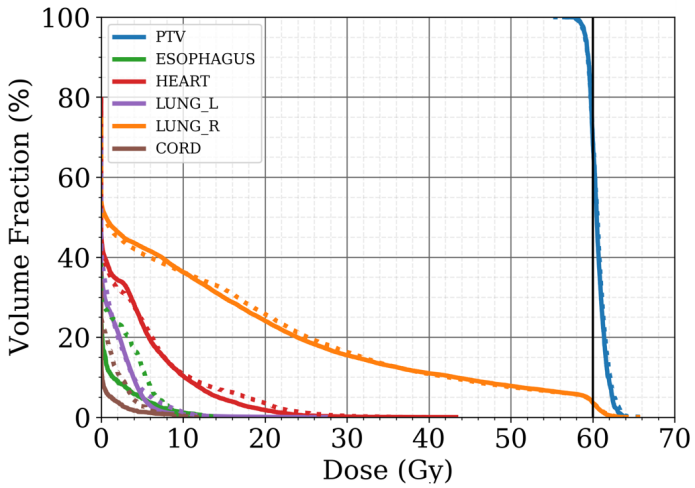


Figure 7: DVH curves generated with expert-selected beams (solid lines) are compared to those from globally optimal beams (dashed lines), the latter obtained by solving a mixed integer programming problem (> 4 days to solve).

plans. Subsequently, these models can predict the 3D dose distribution for a new patient. Following this prediction, the "Optimization" module can be employed for dose mimicking optimization, and the optimal control points (or fluence map) can be imported into a TPS for final evaluation. Illustrative workflows and examples are provided in our GitHub repository's Jupyter Notebooks.

4 Discussion and Conclusion

PortPy has been developed to address the limitations of the current packages, with a primary aim to advocate transparency, reproducibility, and foster community-driven development within the realm of radiotherapy treatment planning. PortPy stands as an open-source community project, under continuous enhancement, and we invite contributions and involvement from the broader scientific community. PortPy has been downloaded over 7,000 times in the past four months, a testament to its adoption within the scientific community.

References

[1] J. O. Deasy, A. I. Blanco, and V. H. Clark. "CERR: a computational environment for radiotherapy research". *Medical physics* 30.5 (2003), pp. 979–985.

[2] H.-P. Wieser, E. Cisternas, N. Wahl, et al. "Development of the open-source dose calculation and optimization toolkit matRad". *Medical physics* 44.6 (2017), pp. 2556–2568.

[3] M. Zarepisheh, L. Hong, Y. Zhou, et al. "Automated and clinically optimal treatment planning for cancer radiotherapy". *INFORMS journal on applied analytics* 52.1 (2022), pp. 69–89.

[4] S. Diamond and S. Boyd. "CVXPY: A Python-embedded modeling language for convex optimization". *The Journal of Machine Learning Research* 17.1 (2016), pp. 2909–2913.

Pylinac: An open-source quality assurance image analysis engine

James Kerns[1]

[1]Radformation, New York, NY, United States of America

Abstract Pylinac is an open-source image analysis library written in Python that can analyze routine quality assurance images and datasets for medical physicists in radiotherapy. The library is concise and usually implemented in a few lines of code. The engine is used in a commercial application and receives regular updates. Several modules are available for complete analysis of starshots, CatPhan phantom, and planar (2D) phantoms from numerous commercial manufacturers.

1 Introduction

The use of software to automate quality assurance (QA) analysis has been well established in medical physics [1–3]. Its use can assist or automate procedures or analysis of images or data. Although publishing studies of in-house software is common, these studies rarely publish the code itself for use by other researchers. As well, the American Association of Physicists in Medicine (AAPM) TG-142 [4] and TG-198 [5] has added more tests for a physicist to be responsible for; thus quick, accurate, and automatic test analysis can reduce the physicist's time spent on routine QA items.

The goal of this work is to showcase one part of a QA software analysis programming library that is robust, automatic, extensible, and open-source. The library can be integrated into a QA workflow to automate image analysis and data exploration. Currently, the library has numerous modules that each address a common QA test. Examples include volumetric-modulated arc therapy (VMAT) QA, starshots, cone-beam CT (CBCT)/CT, machine logs, Winston-Lutz isocenter size [6], and picket fences.

Several commercial options are available to automate or assist in completing TG-142 items, however they are proprietary. This library is open-source and thus all code and algorithms can be inspected by users for robustness. The library is also vendor-agnostic; several manufacturers are supported and is largely driven by user requests. The library can be integrated with other libraries to provide more features, such as a graphical user interface or web server interface. Given its open-source nature and liberal licensing, it can be extended and used by the community to conform to the needs of individual clinics.

The aim of the library is to provide an easy-to-use programming interface for doing complex but routine QA tasks or data exploration that is tedious or error-prone to do manually. The library automates as much of the analysis as possible, including data filtering and image localization; only input that is necessary should be given by the physicist. Output includes raw values, figures, and PDF reports. It is object-oriented in nature, allowing for code reuse and extension, and most algorithms can be utilized with fewer than 10 lines of code. The library is available on Github [7].

2 Materials and Methods

For the sake of space, only the CatPhan analysis is addressed here.

2.1 Data Collection

All data was collected on a TrueBeam accelerator, equipped with an on-board kV imager. Results were measured manually to benchmark the library.

2.2 Data Analysis

The library is written in the general programming language Python [8]. Although MATLAB is a very common language for numerical data analysis, including uses in medical physics, the extensibility beyond numerical analysis is limited compared to general programming languages. The library uses several open-source Python packages for reading, processing, and plotting data.

Cone-beam CT and CT QA commonly use the CatPhan phantom to measure several image quality metrics. The analysis software can automatically register CatPhan 503, 504, 600, and 604 phantoms, correct for position offsets in all three Cartesian directions, and correct for modest phantom roll, pitch, and yaw. The library can analyze the CTP404 (HU linearity & geometric scaling), CTP528 (high-contrast line pair resolution), CTP486 (HU uniformity), and CTP515 (low-contrast detectability) modules. Where applicable, the analysis procedure follows the CatPhan manual procedures for each phantom. The analysis procedure is as follows:

Register the phantom z-location. Upon loading, all the images are scanned to determine where the houndsfield units (HU) linearity module (CTP404) is located. This is accomplished by examining each image slice and looking for 1) a disk the size of the CatPhan and 2) a circular profile at the locations of the HU linearity regions of interest (ROIs) that has characteristics like the CTP404 module. If the profile contains low, high, and mostly medium values then it is thought to be the HU linearity module. All such slices are found and the median slice is set as the HU linearity module location. All other CatPhan modules are at known offsets from this position.

Determine phantom roll. Although there is often little axial roll in the CatPhan, even small rotations can misalign the ROIs and lead to inaccurate results. The roll of the phantom is determined by converting the HU linearity slice to a binary image based on a threshold. The air-filled ROIs, which are below the threshold, are then located and the angle made by the two holes determines the phantom roll.

Determine HU linearity. The HU module (CTP404) contains several materials with different HU values. Using angle positions corrected for roll and the radius from the center of the phantom, circular ROIs are sampled which correspond to the HU material regions. The mean pixel value of the ROI is the determined HU value.

Determine HU uniformity. HU uniformity (CTP486) is calculated in a similar manner to HU linearity, but within the CTP486 module/slice looking at a top, bottom, left, right, and central ROI.

Calculate Geometry/Scaling. The HU module (CTP404) also contains several "nodes" which have a known spacing. Using corrected angles, the area around the 4 nodes are sampled and then a threshold is applied which identifies the node within the ROI sample. The center of mass of the node is determined and then the distance between nodes is calculated and averaged.

Calculate Spatial Resolution/MTF. The Spatial Resolution module (CTP528) contains 21 pairs of aluminum bars having varying thicknesses. One unique advantage of these bars is that they are all focused on and equidistant to the phantom center. Thus, a circular profile is taken at the radius of the line pairs. The peaks and valleys of the profile are located; peaks and valleys of each line pair are used to calculated the modulation transfer function (MTF). The relative MTF (rMTF), i.e. normalized to the first line pair, is then calculated from these values. Up to the first eight line pairs are analyzed; however, each line pair region is analyzed sequentially and if the MTF falls below a detectable level the analysis stops at the last reasonable line pair region.

Calculate Low Contrast Resolution. Low contrast is difficult to determine quantitatively since visual acuity has several factors. The library uses both the contrast value of the ROI as well as the ROI size to compute a "detectability" score (equation 1), which is based on the contrast-to-noise equation. ROIs above the score are said to be "seen", while those below are not seen. Only the 1.0% supra-slice ROIs are examined. Two background ROIs are sampled on either side of each ROI and averaged. The score for a given ROI is calculated using equation 1:

$$Visibility(I) = \frac{I_{max} - I_{min}}{I_{max} + I_{min}} \times \frac{\sqrt{\pi r^2}}{\sigma_I} \quad (1)$$

where I is the region of interest and r is the radius of the ROI, and σ_I is the standard deviation of the ROI.

Calculate Slice Thickness. Slice thickness is measured by determining the full-width half-max (FWHM) of the wire ramps in the CTP404 module. A profile of the area around each wire ramp is taken, and the FWHM is determined from the profile. The profiles are averaged and the value is converted from pixels to mm and multiplied by 0.42 as per the CatPhan manual. Measurements of the image quality metrics were also determined by hand according to the CatPhan 504 manual at the treatment console at the time of measurement and compared to algorithm analysis results where applicable.

3 Results

CBCT results for a small field of view (FOV) protocol scan can be visualized in Figure 1, which shows a summary of results. The figure shows the slices where the analyses were done, markers to show the ROIs of each test, and plots of the HU, rMTF, and uniformity slice. The acquired values were compared to the manual ones taken at the time of measurement and are presented in Table 1.

The difference between the manual measurements and those of the algorithm are quite small for all measurements taken. On average, the HU analysis had a difference of 1.6 and 2.8 HU for the head and thorax acquisition modes, respectively. HU differences between manual and algorithm values was never more than 10, indicating good agreement. Geometric scaling showed agreement to within 0.2mm. The slice thickness measurements showed the largest relative difference in measurement. Given that the nominal slice thickness was 2 mm however, the automatic analysis showed better agreement to the nominal value on average, being within 0.2 mm for both scans. It should be noted that slice thickness accuracy degrades with lower image quality since the wire ramps become more difficult to delineate from the noise; this may explain the higher thorax scan manual value.

4 Discussion

The results of the benchmark tests all agreed well with the manual measurements taken. The library has been shown to have algorithms that can analyze routine QA test files with a high level of accuracy. The library is also well documented for easy setup and use, creating a low barrier to entry, especially for users without a deep knowledge of programming.

Robustness of the algorithms are ensured by continuous code testing. Each improvement or new feature is backed by corresponding code tests to ensure that changes have not introduced unintentional errors to existing features and that the new code works properly. The library has a large and growing test data suite of hundreds of images and datasets to ensure data from different machines, different acquisition conditions, and different data qualities give consistent results. These images have been contributed from a variety of clinics, ensuring a range of procedures and results and equipment performance. Users are encouraged to donate images to the test suite. Also, because of the transparency by being open-source, users can examine the algorithms for accuracy,

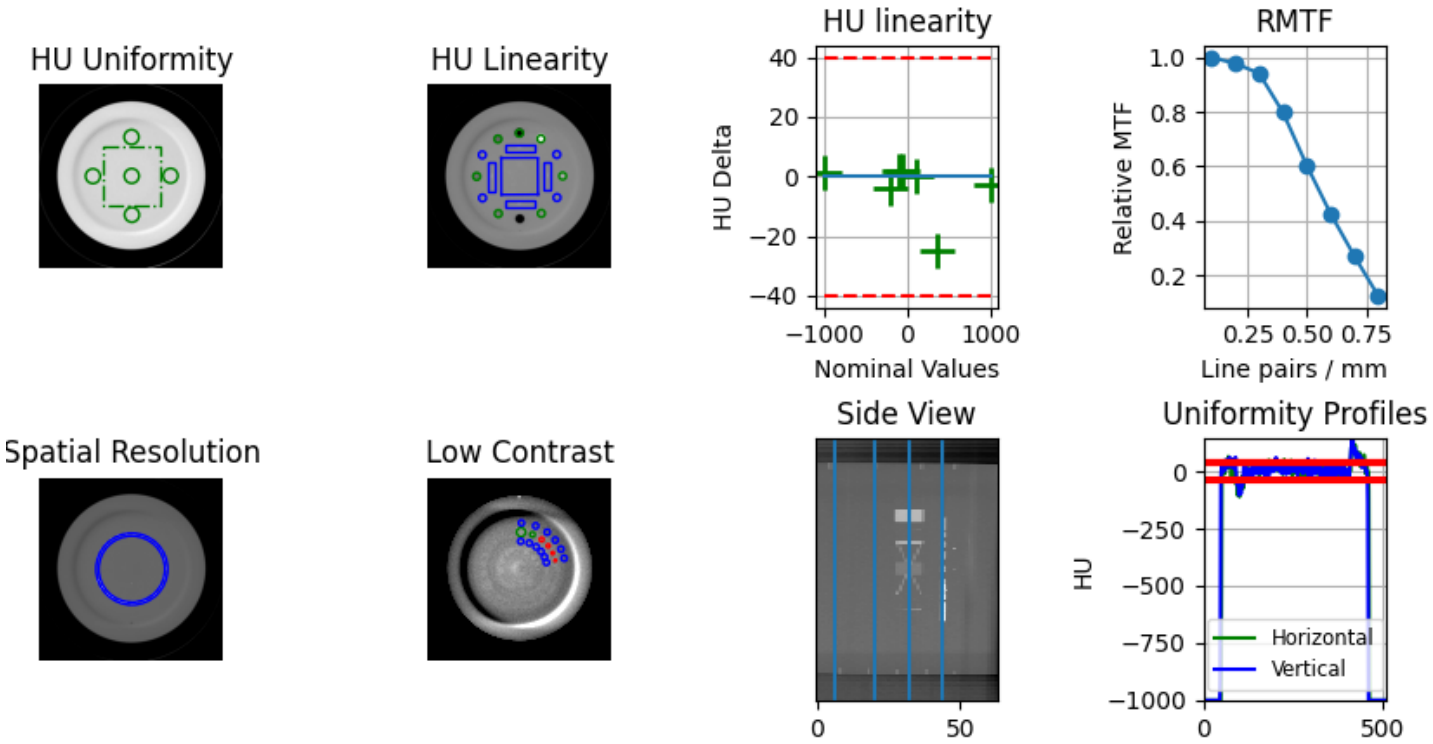


Figure 1: Visual results of the CT analysis showing relevant regions of interest for each module. Additionally, the relative MTF and HU linearity delta from baseline are plotted.

			Head			Thorax	
Module	**Region**	**Manual**	**Algorithm**	**Difference**	**Manual**	**Algorithm**	**Difference**
HU Linearity (HU)	Air	-992.9	-993.3	0.4	-999.8	-998.3	1.5
	Teflon	996.4	992.0	4.4	969.4	971.3	1.9
	Delrin	363.1	363.0	0.1	362.1	358.5	3.6
	Acrylic	128.6	127.0	1.6	120.9	123.0	2.1
	Poly	-41.2	-40.0	1.2	-34.1	-32.3	1.8
	LDPE	-96.5	-96.0	0.5	-90.6	-89.3	1.3
	PMP	-190.9	-184.7	6.2	-191.7	-182.0	9.7
Uniformity (HU)	Top	9.0	9.0	0.0	15.2	12.7	2.5
	Bottom	11.3	9.7	1.6	11.0	16.3	5.3
	Left	12.0	11.7	0.3	19.5	20.0	0.5
	Right	8.1	6.7	1.4	10.9	10.0	0.9
	Center	6.2	4.3	1.9	19.6	22.7	3.1
Geomtric Scaling (mm)	Mean	50.0	50.0	0.0	49.9	49.8	0.1
Slice Thickness (mm)	Mean	2.1	1.8	0.3	2.6	2.1	0.5

Table 1: CBCT measurement values comparing manual measurements of HU linearity, HU uniformity, geometric scaling, and slice thickness with the automatic algorithm analysis and difference between them for the “Head” and “Thorax” scan protocols.

implementation, and can contribute to improvements and enhancements, all of which adds to the robustness of the library.

The library is also extensible because of its modular and object-oriented design. As an example, several of the library's modules use digital imaging and communications in medicine (DICOM) images in their analysis. These modules all use a DICOM image class that itself can perform various data manipulations. Users can directly extend this class to do a specialized manipulation of their own DICOM images. As well, because Python has frameworks for web applications and databases, it is easy to use such libraries as this one for local or cloud-based web services.

Several investigators have also used and validated Pylinac within their own studies [2, 9, 10].

5 Conclusion

An open-source library has been built which allows users to easily analyze images and data for routine QA procedures. The CatPhan routine was validated in this study. The library has been benchmarked, agrees well with manual measurements, and has been tested using hundreds of image sets. The library can be used standalone or integrated with other services to provide any level of QA support for a clinic.

References

[1] D. Binny, T. Aland, B. R. Archibald-Heeren, et al. "A multi-institutional evaluation of machine performance check system on treatment beam output and symmetry using statistical process control". *Journal of Applied Clinical Medical Physics* 20 (3 2019). DOI: 10.1002/acm2.12547.

[2] J. Boudet, L. Aubignac, A. Beneux, et al. "Evaluation of QA software system analysis for the static picket fence test". *Journal of Applied Clinical Medical Physics* 23 (7 2022). DOI: 10.1002/acm2.13618.

[3] A. M. Fanou, N. Kollaros, G. Patatoukas, et al. "A step closer to automation: kilovoltage and Megavoltage Planar Imaging Quality Assurance, baseline, tolerance and action levels definition and exploration". *Physica Medica* 107 (2023). DOI: 10.1016/j.ejmp.2023.102536.

[4] *Task group 142 report: Quality assurance of medical accelerators*. 2009. DOI: 10.1118/1.3190392.

[5] J. Hanley, S. Dresser, W. Simon, et al. *AAPM Task Group 198 Report: An implementation guide for TG 142 quality assurance of medical accelerators*. Oct. 2021. DOI: 10.1002/mp.14992.

[6] K. R. Winston and W. Lutz. "Linear accelerator as a neurosurgical tool for stereotactic radiosurgery". *Neurosurgery* 22 (3 1988). DOI: 10.1227/00006123-198803000-00002.

[7] J. R. Kerns. "Pylinac: Image analysis for routine quality assurance in radiotherapy". *Journal of Open Source Software* 8 (92 Dec. 2023), p. 6001. DOI: 10.21105/joss.06001.

[8] P. Teams. *What is Python? Executive Summary*. 2023.

[9] "Assessment of cone beam computed tomography image quality and dose for commonly used pre-sets in external beam radiotherapy". *Radiation Physics and Chemistry* 199 (2022). DOI: 10.1016/j.radphyschem.2022.110287.

[10] R. C. Tegtmeier, W. S. Ferris, J. E. Bayouth, et al. "Characterization of imaging performance of a novel helical kVCT for use in image-guided and adaptive radiotherapy". *Journal of Applied Clinical Medical Physics* 23 (6 2022). DOI: 10.1002/acm2.13648.

A long short-term memory network-based proton dose calculation method in a magnetic field

Fan Xiao[1], Domagoj Radonic[1,2], Niklas Wahl[3,4], Luke Voss[3,4,5], Ahmad Neishabouri[4,6], Nikolaos Delopoulos[1], Stefanie Corradini[1], Claus Belka[1,7,8], George Dedes[2], Christopher Kurz[1], and Guillaume Landry[1]

[1]Department of Radiation Oncology, LMU University Hospital, LMU Munich, Munich, Germany
[2]Department of Medical Physics, LMU Munich, Munich, Germany
[3]Department of Medical Physics in Radiation Oncology, German Cancer Research Center (DKFZ), Heidelberg, Germany
[4]National Center for Radiation Oncology (NCRO), Heidelberg Institute for Radiation Oncology (HIRO), Heidelberg, Germany.
[5]Ruprecht Karl University of Heidelberg, Institute of Computer Science, Heidelberg, Germany
[6]Clinical Cooperation Unit Radiation Oncology, German Cancer Research Center (DKFZ), Heidelberg, Germany
[7]German Cancer Consortium (DKTK), partner site Munich, a partnership between DKFZ and LMU University Hospital, Munich, Germany
[8]Bavarian Cancer Research Center (BZKF), Munich, Germany

Abstract ***Objective*:** To present a long short-term memory (LSTM) network-based dose calculation method for magnetic resonance (MR)-guided proton therapy. ***Approach*:** 30 deformed planning computed tomography (CT) images of prostate cancer patients were collected for Monte Carlo (MC) dose calculation under a 1.5 T magnetic field. Proton pencil beams (PB) at 150, 175, and 200 MeV energies were simulated (6480 PBs at each energy). A 3D relative stopping power (RSP) cuboid covering the extent of the PB dose was extracted and flattened into the LSTM model, yielding a 3D predicted PB dose. Training and validation involved 25 patients; 5 patients were reserved for testing. ***Results*:** Test results of all PBs showed a mean gamma passing rate (2%/2 mm) of 99.89%, with an average center-of-mass discrepancy of 0.36 mm between predicted and simulated trajectories. The model inference time was 10 ms per beam. ***Conclusion*:** LSTM models for proton dose calculation in a magnetic field were developed and showed promising accuracy and efficiency for prostate cancer patients.

1 Introduction

Proton therapy has limitations in achieving high-precision dose delivery due to its inherent range uncertainty [1], which drives the need for online image guidance during proton dose delivery. Magnetic resonance imaging (MRI) has the advantages of high soft tissue contrast and non-ionizing radiation, and offers the potential for real-time anatomical imaging [2]. With the combination of real-time MRI and proton therapy, the treatment plan could be adapted or re-optimized based on the latest patient geometry before dose delivery, which would improve the target dose coverage and reduce the dose to surrounding normal tissues [3]. To make this online adaptive mode a reality, the plan adaptation and re-optimization require an accurate and fast proton dose calculation engine.
For proton dose calculation in magnetic fields, the Lorentz force causes protons to be deflected as they travel [4]. Monte Carlo (MC) methods, which compute the trajectories of individual particles by simulating particle transport physics, can consider the impact of magnetic fields on dose distributions but also require high computation times. To offer a trade-off between speed and accuracy, several analytical methods have been proposed to estimate the proton beam deflection and achieve accuracy close to MC methods in homogeneous tissues. Although the calculation time of recent MC and analytical methods can be accelerated by graphics processing units (GPUs) from several hours to several seconds [5, 6], the ultimate need for future real-time plan adaptation and re-optimization is still unmet, which requires dose engines to produce MC accuracy at millisecond speeds in the online iterative process.
Recently, several studies have shown the feasibility of deep learning techniques to realize millisecond proton dose calculations with MC accuracy. Some methods use computationally cheap physical inputs and computed tomography (CT) scans as inputs to U-Net convolutional models, which learn the correlation between physical inputs, CT scans and MC label dose [7, 8]; some other methods use long short term memory (LSTM) and transformer models, which are suitable for processing sequences with long dependencies, to learn the mapping of CT sequences to dose sequences in the beam's eye view (BEV) coordinate system [9, 10]. Although the two kinds of methods achieved good accuracy and high speed, to the best of our knowledge, they have not been applied to proton dose calculation scenarios under magnetic fields.
In this work, we applied a previously published LSTM proton dose calculation model [9] to dose calculation in a 1.5 T magnetic field, and confirmed its accuracy in patient cases with prostate cancer.

2 Materials and Methods

2.1 Proposed framework

The framework of this study is shown in Figure 1. First, based on the given initial energy and beam direction, a 3D relative stopping power (RSP) cuboid of interest containing the proton beam dose region is resampled from the patient's CT-derived RSP. Then, in the BEV coordinate system, the 3D cuboid of interest is treated as a sequence of 2D slices traveling from upstream to downstream. By flattening these

2D slices into 1D sequences, the directional LSTM unit can process the RSP sequences and generate the internal hidden states and outputs, where the former can be used as the input information for the subsequent slices. The outputs are passed to a fully-connected neural network to generate the predicted dose slices. Finally, the dose slices from our model are spliced back to the original 3D dose size and compared with the ground truth MC dose.

2.2 Dataset processing

We used a patient dataset which included deformed planning CT images after registration with MRI from 30 prostate cancer patients treated with a 0.35 T MR-linac (MRIdian, ViewRay, USA) at the Department of Radiation Oncology of the LMU University Hospital. All patients had no artificial implants and went through prior anonymization to comply with data protection law. The ground truth MC dose was generated through Geant4, and RSP maps with respect to water were then created within Geant4 and exported. To perform dose calculations in the presence of a magnetic field, a 1.5 T homogeneous magnetic field was set to be aligned with the z-axis (parallel to the superior direction) and act within a cylinder of 30 cm radius. The dose was scored on the same voxel geometry as the CT input, and all the proton beams were simulated with 10^6 histories, which made the maximal relative standard error in all regions above 10% of the maximum dose always below 1.15%.
Pencil beams (PBs) with energies of 150 MeV, 175 MeV and 200 MeV were simulated, with an energy spread of 0.83 MeV and laterally the beam was Gaussian with a standard deviation of 4.2 mm. PBs were isotropically sampled over a nearly cylindrical surface surrounding the prostate region and confined within a predefined angular sector to ensure no PBs were placed posterior or anterior to the patient. The propagation of the PBs was chosen to take place in the transverse plane, always orthogonal to the B field. Defining the right lateral direction as the 0° reference line, the sampling was done within $\alpha \in [40°, 140°] \cup [220°, 320°]$ region in $\Delta\alpha = 16.\dot{6}°$ steps, yielding 12 different angles per patient. At each of the 12 angles, a 3×6 PB grid was generated in the plane orthogonal to the PB's direction. In total, 216 PBs were simulated per patient and 6480 PBs for each energy.
After applying the body mask to RSP and dose maps, utilizing the transformation relationship between the patient and BEV coordinate systems, the 3D RSP-dose cuboid pairs were cropped and resampled according to the initial position and orientation of PBs. Cuboid sizes of $170 \times 43 \times 23$, $188 \times 49 \times 25$, and $205 \times 55 \times 25$ were used for the PBs with energies of 150 MeV, 175 MeV and 200 MeV, respectively. For each energy, 4320 PBs from 20 patients were used to train the model, 1080 PBs from 5 patients for validation and the remaining 5 patients with 1080 PBs for testing. Then, a global normalization was performed for the training, validation, and testing cuboid datasets using the maximum RSP and dose cuboid values of the training dataset. Finally, a reflection transformation across the xy plane was used, which doubled the original training dataset from 4320 PBs to 8640 PBs for each energy.

2.3 Model training

The network architecture used in this work corresponds to the LSTM model used in [9] for proton dose calculation in the absence of a B field in lung cancer patients. The LSTM features one layer with 1000 neurons as internal layer, followed by two fully connected layers with 100 neurons and ReLu activation layers.
For model training, we used the Adam optimizer to minimize the mean squared error (MSE) loss between the output dose and ground truth dose sequences. The initial learning rate was set to 10^{-5} and batch size to 8. The model was implemented in Python 3.8 and Pytorch 2.0.1 and trained with an NVIDIA RTX A6000 GPU (48 GB memory) for around 15000 epochs, which took about a week. The model with the best overall MSE validation loss was identified and used for the prediction in the test dataset. Each energy was trained separately.

2.4 Evaluation metrics

After obtaining model predictions, all dose voxels below a threshold of 0.01% of the global maximum were set to 0, similarly to [10]. Then, 198 PBs that overshot the patient were filtered out in the test dataset with energy of 200 MeV and all the output doses were reverse-normalized for the final evaluation.
Comparison metrics used in our study were according to the metrics in [11]. The 3D global gamma pass rate γ_{pr} was calculated to compare the 3D dose distributions, which was implemented within an open-source project, PyMedphys [12]. The distance-to-agreement threshold was set to 2 mm, the dose difference threshold to 2% and with a dose cutoff to 10% of the maximum dose. The relative error ε_{rel} was used to quantify the relative disagreement between two laterally integrated dose profiles at a certain depth. The integrated dose difference ε_{int} was evaluated to allow a comparison between PBs of different energies with each other. To compare the depth range of PBs, $\Delta R_{D_{80}}$ was used to calculate the difference of the depth of the 80% dose distal to the Bragg peak between two dose distributions. Finally, to analyze magnetic deflection modeling and the agreement of two PBs' lateral profiles at a given depth, the difference in the center of mass F_{COM} and full width at half maximum value FW_{50} of the PBs were calculated.

3 Results

Figure 2 (a) and (b) show a simulated and predicted PB from a 175 MeV energy model on a test patient. The LSTM can accurately predict the entry point and angle of the PB

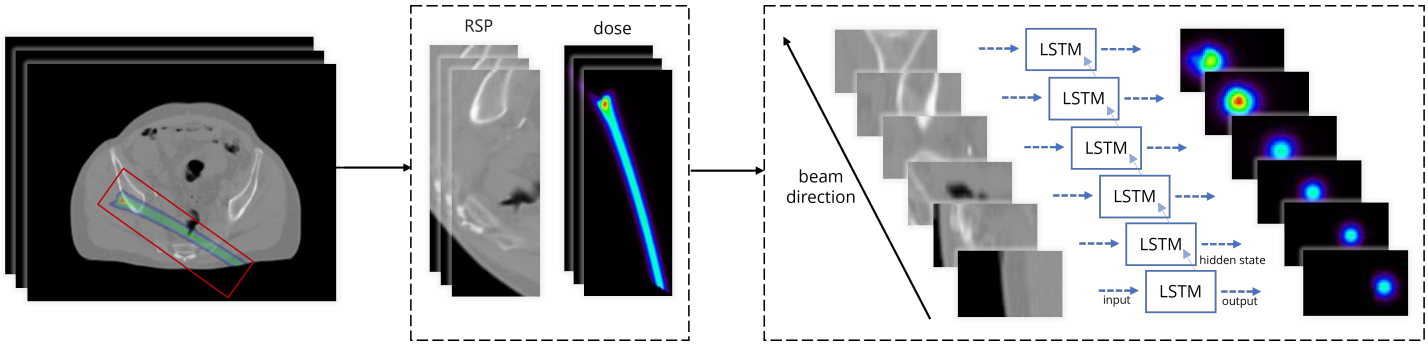
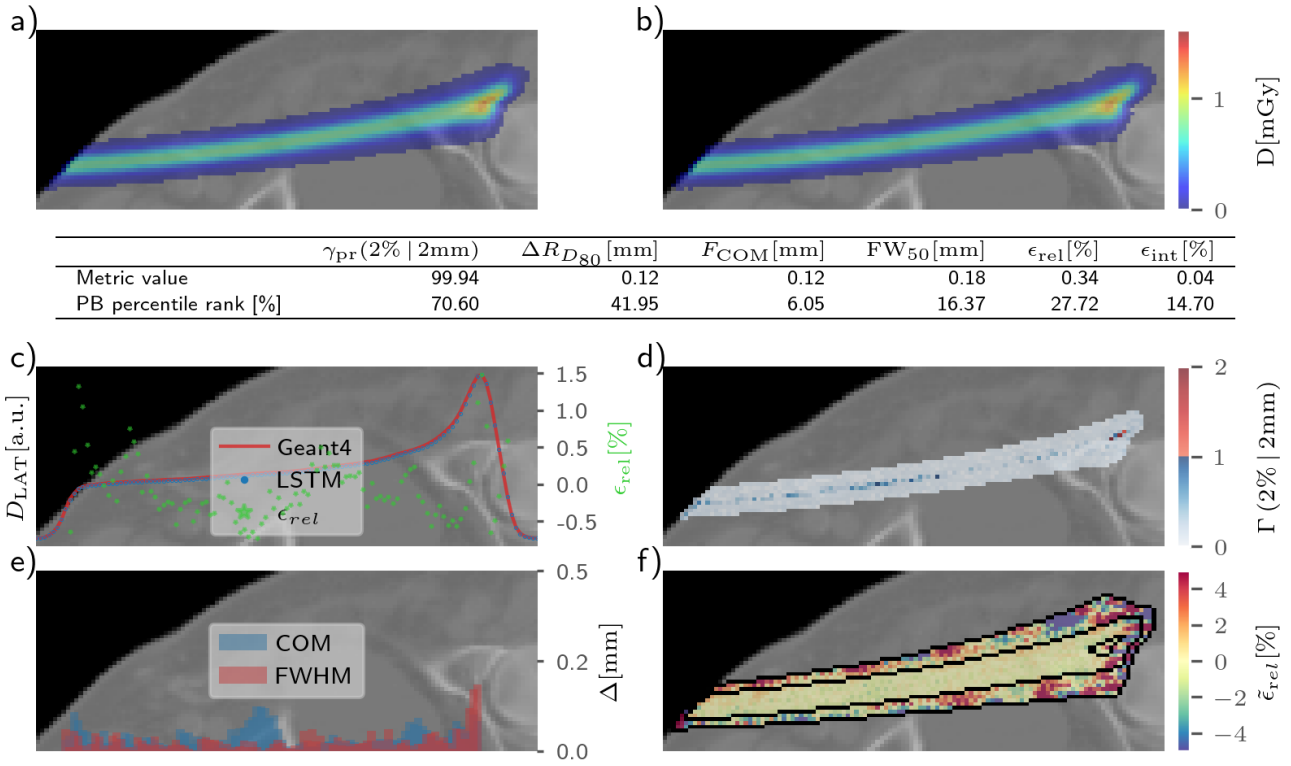


Figure 1: The framework of this study. (left) Dose distribution from a single pencil beam obtained from MC simulation with outline of 3D cuboid of interest. (middle) Resampling of input RSP and output dose for the cuboid of interest. (right) LSTM model converting sequence of RSP slices to a sequence of dose slices.



	$\gamma_{pr}(2\% \mid 2mm)$	$\Delta R_{D_{80}}$ [mm]	F_{COM} [mm]	FW_{50} [mm]	ϵ_{rel} [%]	ϵ_{int} [%]
Metric value	99.94	0.12	0.12	0.18	0.34	0.04
PB percentile rank [%]	70.60	41.95	6.05	16.37	27.72	14.70

Figure 2: An example of a predicted 175 MeV PB in test patient. An *xy* slice of a) Geant4 simulated and b) LSTM predicted dose distribution is plotted over the RSP map input. The table shows the PB's performance regarding different metrics and percentile ranks with respect to all other PBs from the test set. A percentile rank of 0% corresponds to the best- and 100% to the worst-ranked PB. c) Integrated lateral dose profiles over depth and ε_{rel} of each slice. d) Γ (2%|2mm) index. e) F_{COM} and FW_{50} at each depth. f) Voxel-based relative dose difference of dose distribution integrated over z capped at ±5%. Black lines denote 80%, 10% and 1% isodose lines with respect to z integrated dose in the Bragg peak.

in the patient and the deflection of the PB caused by the magnetic field. In Figure 2 (c), laterally integrated profiles of the simulated and predicted PB are shown and agree well. The $\Delta R_{D_{80}}$ and mean ε_{rel} were less than 0.5 mm and 0.5%, respectively. The Γ index (2%|2 mm) map is shown in Figure 2 (d) and its corresponding γ_{pr} was 99.94%. To analyze the predicted PB trajectory and its lateral profiles, F_{COM} difference and FW_{50} difference are shown in Figure 2 (e), and were both less than 0.5 mm. In Figure 2 (f), the 2D distribution of relative dose differences integrated over z is displayed.

To further analyze the overall model prediction performance for the whole test dataset, Table 1 shows the worst and average values of each evaluation metric for the models at the three energies. The worst γ_{pr} (2%|2 mm, $D > 10\%D_{max}$) was above 95.9%, the maximum $R_{D_{80}}$, F_{COM} and FW_{50} differences were 2.07%, 1.99 mm and 3.28 mm, respectively. High accuracy was achieved with respect to almost every metric we evaluated. For calculation time, the cuboid extraction and interpolation (not optimized) took 45 to 60 ms, while LSTM inference time was 10 ms; our method can thus achieve proton dose calculation per spot within 70 ms.

Table 1: The performance of all evaluated metrics summarized by worst and mean values across all five test patients combined and for each energy model individually. Red cells indicate the worst value.

	$\gamma_{(2\%\,\vert\,2\,\mathrm{mm})}$[%]		$R_{D_{80}}$[%]		F_{COM}[mm]		FW_{50}[mm]		$\vert\varepsilon_{\mathrm{rel}}\vert$[%]		$\varepsilon_{\mathrm{int}}$[%]
	min	mean	mean	max	mean	max	mean	max	mean	max	mean
E [MeV]											
150	95.90	99.90	0.27	1.60	0.40	1.99	0.45	2.62	0.60	4.55	0.26
175	96.64	99.88	0.25	1.85	0.37	1.46	0.46	3.28	0.54	2.75	0.24
200	98.16	99.89	0.17	2.07	0.31	1.76	0.45	2.54	0.46	2.30	0.19
Overall	95.90	99.89	0.23	2.07	0.36	1.99	0.45	3.28	0.54	4.55	0.23

4 Discussion

This study is an application of a previously developed architecture [9] to magnetic field dose calculation while using more extensive and rigorous criteria. To our knowledge, this work is the first one to utilize an AI architecture to speed up proton dose calculation in magnetic fields. Compared with recent transformer-based dose calculation methods [10], our study demonstrated that the LSTM architecture can also achieve state-of-the-art accuracy but has lighter trainable parameters; the GPU memory requirements of our model never increased above 1 GB and our network is easy to train and converged with a constant learning rate of 10^{-5}.

We can compare with recent analytical dose calculation methods for MRI-guided proton therapy [5], where maximum observed differences in water between their approach and MC ground truth were reported. For $\Delta R_{D_{80}}$, they reported 0.7 mm or 0.22% for the highest tested energy of 229 MeV. Furthermore, the maximum reported F_{COM} was 1.2 mm. In contrast, we reported for the highest energy model of 200 MeV a mean $\Delta R_{D_{80}}$ of only 0.17 mm and mean F_{COM} between 0.3 and 0.4 mm. This was in real test patients rather than in homogeneous water. Reference [6] presented a recent GPU-based MC algorithm for MRI-guided proton therapy called ARCHER. They report achieving a $\Delta R_{D_{80}} < 0.06$ mm and a mean $\varepsilon_{\mathrm{rel}}$ of 0.81% in water. For tissue and bone materials, which were also evaluated, they reported a mean $\varepsilon_{\mathrm{rel}}$ of 1.12% and $F_{\mathrm{COM}} < 0.05$ mm. Their 3D-gamma pass rate with a 2 mm/2% criterion in the region with dose greater than 10% of the maximum dose was over 99% for all tested cases. Although the computational time of ARCHER is found to range from 0.82 to 4.54 s for 10^7 proton histories, the potential advantage of the deep-learning method is that we can use more particles to simulate MC doses with lower statistical uncertainty as labels without increasing the prediction time. Finally, just like a MC engine has to be carefully adapted and commissioned for use in a specific treatment setting, a deep-learning model would have to be trained or at least fine-tuned. Unique settings and beam profiles of a specific machine necessitate a specific model per machine.

5 Conclusion

We successfully developed an LSTM network-based proton dose calculation method in a magnetic field and confirmed its accuracy for prostate patients. This sub-second dose engine has the potential to improve the efficiency of online plan optimization in the MR-guided proton therapy workflow.

References

[1] H. Paganetti. “Range uncertainties in proton therapy and the role of Monte Carlo simulations”. *Physics in Medicine & Biology* 57.11 (2012), R99.

[2] M. A. Schmidt and G. S. Payne. “Radiotherapy planning using MRI”. *Physics in Medicine & Biology* 60.22 (2015), R323.

[3] A. Hoffmann, B. Oborn, M. Moteabbed, et al. “MR-guided proton therapy: a review and a preview”. *Radiation Oncology* 15.1 (2020), pp. 1–13.

[4] J. Hartman, C. Kontaxis, G. Bol, et al. “Dosimetric feasibility of intensity modulated proton therapy in a transverse magnetic field of 1.5 T”. *Physics in Medicine & Biology* 60.15 (2015), p. 5955.

[5] A. Duetschler, C. Winterhalter, G. Meier, et al. “A fast analytical dose calculation approach for MRI-guided proton therapy”. *Physics in Medicine & Biology* 68.19 (2023), p. 195020.

[6] S. Li, B. Cheng, Y. Wang, et al. “A GPU-based fast Monte Carlo code that supports proton transport in magnetic field for radiation therapy”. *Journal of Applied Clinical Medical Physics* 25.1 (2024), e14208.

[7] G. Zhang, X. Chen, J. Dai, et al. “A plan verification platform for online adaptive proton therapy using deep learning-based Monte–Carlo denoising”. *Physica Medica* 103 (2022), pp. 18–25.

[8] C. Wu, D. Nguyen, Y. Xing, et al. “Improving proton dose calculation accuracy by using deep learning”. *Machine learning: science and technology* 2.1 (2021), p. 015017.

[9] A. Neishabouri, N. Wahl, A. Mairani, et al. “Long short-term memory networks for proton dose calculation in highly heterogeneous tissues”. *Medical physics* 48.4 (2021), pp. 1893–1908.

[10] O. Pastor-Serrano and Z. Perkó. “Millisecond speed deep learning based proton dose calculation with Monte Carlo accuracy”. *Physics in Medicine & Biology* 67.10 (2022), p. 105006.

[11] P. Lysakovski, A. Ferrari, T. Tessonnier, et al. “Development and benchmarking of a Monte Carlo dose engine for proton radiation therapy”. *Frontiers in Physics* (2021), p. 655.

[12] S. Biggs, M. Jennings, S. Swerdloff, et al. “PyMedPhys: A community effort to develop an open, Python-based standard library for medical physics applications”. *Journal of Open Source Software* 7.78 (2022), p. 4555.

Deep learning-based dose prediction for INTRABEAM

Mojahed Abushawish[1,2*], Arthur V Galapon [1,3*], Joaquín L Herraiz[1,4], José M Udías[1,4], Paula Ibáñez[1,4]

[1]University Complutense of Madrid, Nuclear Physics Group & IPARCOS, Madrid, Spain.
[2]CNRS, Délégation Rhône Auvergne, Lyon, France.
[3]University Medical Center Groningen, Department of Radiation Oncology, Groningen, Netherlands.
[4]Clinical hospital San Carlos, Health research institute, Madrid, Spain
*First two authors have made equal contributions in this work.

Intraoperative radiation therapy requires fast and accurate treatment planning, as these procedures involve tumor removal surgeries preceding irradiation which may alter the original treatment plan. While Monte Carlo calculations are considered the gold standard, they require significant computational resources and time. The use of deep learning algorithms may provide a faster alternative. In this study, we trained and tested two deep learning models, U-Net and MultiResUnet, to predict 3D dose distributions for INTRABEAM spherical and needle applicators. A cohort of 800 paired CTs and dose distributions (720 training, 80 validation) of partial breast irradiations with a spherical applicator and 41 (36 training, 5 validation) of kyphoplasty using the needle applicator were used separately for training and validation. Doses were calculated using a fast MC algorithm, serving as the ground truth. Additional tissue material information was incorporated into the input channel as a segmentation mask, and an image augmentation technique was introduced to produce an image by taking different parts from other images in the dataset. Gamma analysis was performed to compare the predicted dose to the simulated dose. In general, dose predicted conformed well with the reference dose. Consequently, deep learning models show great potential in predicting dose distributions for IORT, especially the MultiResUnet model.

1 Introduction

The use of Intra-Operative Radiation Therapy (IORT) with low-energy X-rays has been experiencing a notable surge in the clinical field. The INTRABEAM system, developed by Carl Zeiss Surgical Gmbh in Oberkochen, Germany, stands out as a miniature low-kV X-ray generator specifically designed for IORT applications. Various applicators may be attached to the probe for different treatment scenarios. However, this technique is currently limited by the inherent difficulties associated with the planning process, as the treatment plan is usually done *in situ* just after the tumor resection inside the operating room. Therefore, it requires fast and accurate treatment planning.

Deep learning algorithms, with their capacity to analyze vast datasets and discern intricate patterns, offer a promising avenue for expediting and enhancing IORT treatment planning. Unlike traditional Monte Carlo (MC) calculations, which demand substantial computational resources and time, deep learning algorithms can swiftly process complex medical imaging data to generate treatment plans that accommodate the dynamic changes occurring during tumor removal surgeries.

In this study, we trained and tested two deep learning models to predict 3D dose distributions for INTRABEAM spherical and needle applicators.

2 Materials and Methods

Two deep learning models, U-Net [1] and MultiResUnet [2], were trained to predict dose distributions of the spherical and needle applicators of the INTRABEAM system. Neural networks (NN) were trained and tested from existing CT images from random patients and the corresponding dose distribution.

2.1. Datasets

Two different datasets were used separately for training and validation:

1) Partial breast irradiation with spherical INTRABEAM applicators. A cohort of 800 paired CTs and dose distributions were used (720 images for training, 80 for validation).

2) Kyphoplasty IORT with needle INTRABEAM applicator. A cohort of 41 paired CTs and dose distributions were used (36 images for training, 5 for validation).

2.2. CT image extraction

The dataset used in this work was extracted from thoracic DICOM images of anonymous patients. CT images were interpolated to have a (0.5,0.5,0.5) mm voxel size and regions of (128,128,64) voxels were extracted from the original CTs. This corresponds to a 6.4 cm x 6.4 cm x 3.2 cm CT image. The regions selected were the left and right breasts for the spherical applicator dataset and the thoracic spine for the needle applicator dataset. Specific cases such as bone structures, tissue-to-lung interfaces and tissue-to air interfaces were selected to provide variety in the dose distribution configuration. Figure 1 shows examples of the central slices of some regions.

2.3. Calculation of dose distributions

Reference doses were calculated using the XIORT-MC algorithm [3], an in-house developed algorithm for kilovoltage x-rays that delivers statistical noise-free distributions in a fraction of time of a MC simulation. A set of phase space files was used for each applicator size – 1.5 cm, 2.5 cm, and 3.5 cm for the spherical applicators and 4.4 mm for the needle applicator. The position of the applicator was defined at the center of the CT image. Each HMC simulation was done for $1x10^7$ histories.

2.4. Input of the neural network

The extracted CT images and reference doses were used as the input and output training set. Additional tissue material

information was incorporated into the input channel as a segmentation mask to help the neural network differentiate between the tissues of the CT image. The training set was normalized.

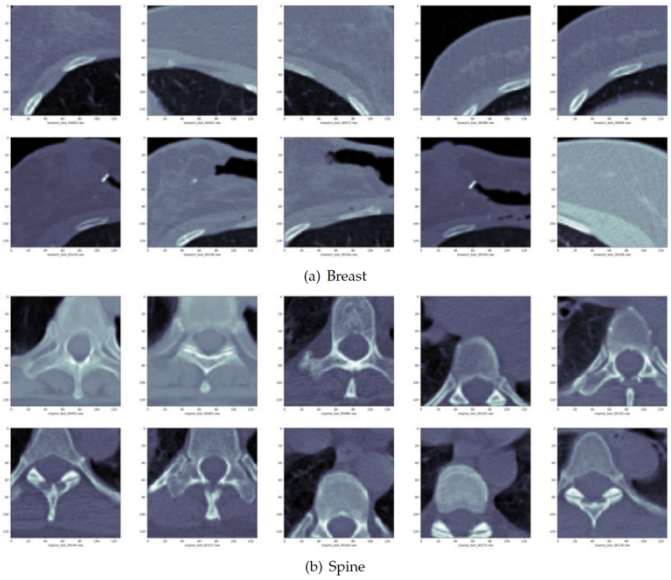


Figure 1. Sample CT central slice of the training set. (a) Breast (b) Spine

2.5. Image augmentation

We proposed an augmentation method based on the Culmix [4] data augmentation technique where a complete image made was built from different parts of other images. This method cannot be done for regular images in classification or segmentation tasks since the high-order features will be lost entirely, but our goal is to model the dose calculation, and this image will still represent the features of the dose calculation code. Moreover, this method will introduce structures that are generally not in the patient's anatomy, like air gaps between tissues, so the stability of the model will be increased. It will also allow us to produce a large and diverse dataset from few engineered images. An example of an artificial CT image generated with our augmentation method is shown in figure 2.

2.6. Model training

U-Net and MultiResUnet networks were used in this study. The models were optimized using the Adam optimizer [5] with a learning rate of 0.0001, beta1 was set at 0.9, beta2 at 0.999, and the epsilon value was equal to 1e-7. Each model was compiled using Keras, and the root mean squared error (RMSE) was the best performing loss function. A model checkpoint parameter was also implemented to save the model and weights of the best possible validation loss. The U-Net and MultiResUnet networks were trained using a subscription of Google Colab Pro running on a Tesla P100-16GB GPU for the model of the needle applicator and a V100-16GB GPU for the model of the spherical applicators. The training loss and validation loss values were saved for analysis. The training duration was also recorded for comparison.

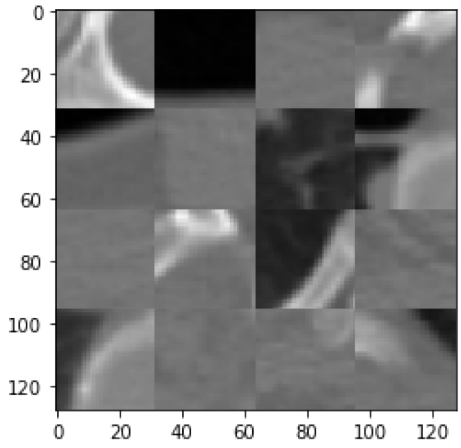


Figure 2. Artificial CT image generated from parts of different images.

2.7. Dose Analysis

The resulting predicted dose was compared to its corresponding ground truth (reference dose). The local dose profile, absolute and relative dose difference, and gamma index analysis were used to compare the ground truth to the predicted dose. The dose profiles located at the center, 5 mm and 15 mm from the center were extracted and compared for the spherical applicator doses. Similarly, the dose profiles at the center, 3 mm and 5 mm from the center were extracted and compared for the needle applicator doses. The absolute dose difference was calculated using pixel-by-pixel subtraction of the reference and predicted dose values. The absolute dose is divided by the reference dose for the relative dose difference calculation. Pixels resulting in a zero-value denominator were automatically set to zero to avoid run-time errors. The global gamma was calculated using 10%-0.5 mm criteria. The lower dose cut-off was set at 10%, and the maximum gamma was set to 2.

2.8. Model comparison

Both neural networks were trained with 2D CT images and doses of the partial breast irradiation dataset to compare between the U-Net and the MultiResUnet. A total of 1242 slices were extracted, 1118 were used for training and 124 for validation. The gamma index was calculated for all the validation images with dose difference criteria of 10%, 7%, 5% and 3%, with 0.5 mm distance-to-agreement (DTA).

3 Results

3.1. Model comparison

Figure 3 shows the box plots of the gamma passing rates obtained with the U-Net and with the MultiResUnet. The average of the gamma passing rate of the U-Net network was 97.9%, 97.14%, 96.51%, and 95.8% for a dose-difference criteria of 10%, 7%, 5%, and 3%, respectively, with 0.5 mm DTA, while for the MultiResUnet, the averages were 98.72 %, 98.12%, 97.59%, and 96.92%. From the averages and the box plot, we can see that the MultiResUnet outperformed the U-Net in all the criteria.

Therefore, the 3D models of the spherical and needle applicators were trained with the MultiResUnet network.

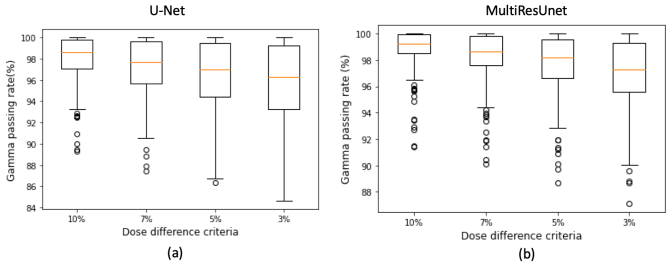


Figure 3. Box plot of the gamma index passing rate for the U-Net (a) and the MultiResUnet (b) with criteria of 0.5 mm and 10%, 7%, 5% and 3% dose difference.

3.2. Spherical applicator

The training of the neural network with the 3D images took 24 hours for 200 epochs on a V100-16GB GPU. An example of a predicted dose distribution for one of the cases is shown in figure 4. Looking at the dose profiles, one can see that the prediction conforms very well with the calculated dose. The network was able to predict the general shape and the gradient of the dose caused by the change in tissue.

3.3. Needle applicator

Training this model took 2 hours and 36 minutes on a P100-16GB GPU, with 500 epochs and a batch size of 2. An example of a predicted dose distribution obtained with the MultiResUnet for the needle applicator is shown in figure 5. In general, the prediction conforms well with the calculated dose, with a slight overestimation of the dose in the voxels next to the applicator.

4 Discussion

A deep learning application for predicting dose in intraoperative radiation therapy was presented in this study. Two convolutional neural network models were tested in this application – U-Net and MultiResUnet. As of the writing, there is no existing research on deep learning applications applied to dose prediction in intraoperative radiation therapy. A similar application developed by Mao et al. used a combination of deep learning algorithms to perform dose predictions for brachytherapy applications [6]. Other existing publications on the use of neural networks in radiation therapy are focused chiefly on external beam applications which several channel inputs such as dose-volume quantities for organs at risk and target dose for the primary target volume (PTV) [7-8]. The dataset used in these publications was based on actual treatment plans delivered to real patients. In this work, the dataset was arbitrarily generated from a set of CT images, and the ground truth for dose comparison was simulated using a Monte Carlo algorithm.

The use of the XIORT-MC algorithm [3] made the dose simulation faster, allowing for the generation of multiple applicator positions for both the spherical and needle applicators.

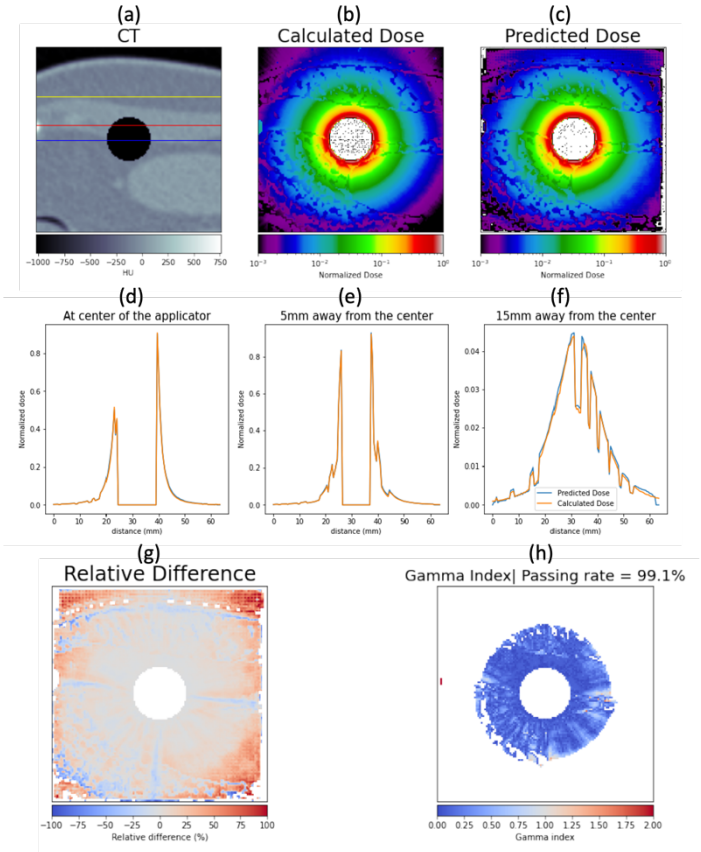


Figure 4: *An example of an equatorial CT slice (a) and the calculated (b) and predicted (c) dose distributions with the MultiResUnet for a spherical applicator. The horizontal lines that appear on the CT are the positions of line profiles (d-f). The relative difference (g) and gamma evaluation (h) are also shown.*

The U-Net architecture was developed initially for semantic segmentation of small dataset [1]. The small number of datasets forces the architecture to rely on data augmentation using deformation operations. The propagation of the filters for the U-Net model can generate a considerable number of trainable parameters, which might be computationally heavy, especially for datasets with multiple input channels. The number of the 3D U-Net structure parameters can reach up to 70 million parameters, requiring a huge video memory. The MultiResUnet is a more advanced structure and has a much lower number of parameters. For 2D, it has 7.2 million and 10.5 million for 3D. The MultiResUnet outperformed the U-Net by one point in average in the gamma passing rate, and it was more stable than the U-Net.

Normalization of the dataset transforms the data to a similar scale. A considerably extensive range in scales like that of the Hounsfield unit or the dose gradient can lead to the model not learning. Like most machine learning applications, data cleaning is essential to reduce extreme outliers.

In this study, the application of dose normalization and the use of a segmentation mask provides this option. Since the training set was directly generated using MC simulation, there is a higher chance of generating outliers due to the statistical nature of the algorithm. Treatment planning system (TPS) generated dose distribution may minimize this effect due to its existing built-in dose optimization algorithms. Nevertheless, the generated dose distribution using XIORT-MC is enough for the application in this study.

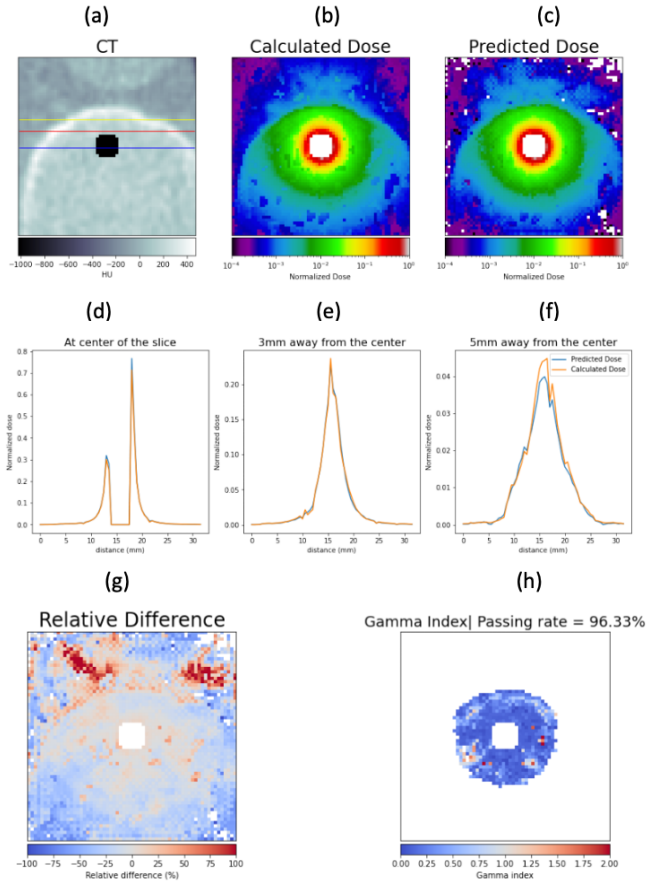


Figure 5: *An example of an equatorial CT slice (a) and the calculated (b) and predicted (c) dose distributions with the MultiResUnet for the needle applicator. The horizontal lines that appear on the CT are the positions of line profiles (d-f). The relative difference (g) and gamma evaluation (h) are also shown.*

We also proposed a new data augmentation technique where we produce an image by taking different parts from other images in the dataset. This technique will introduce a variety of structures around the applicator, consequently increasing the stability of the model and improving the ability to model the dose calculation code. It will also allow for the production of a large and diverse dataset from a small engineered data. The network was trained using the augmented dataset with the same spherical applicator to evaluate its effectiveness. The training took 24 hours. The average gamma passing rate increased by 2.54 points over the normal dataset; also, the prediction accuracy in lung tissue improved.

Our main objective was to produce a model that can calculate a treatment plan for IORT in real time. The XIORT-MC code takes 350 seconds to calculate the dose of a CT with a size of 128 ×128 × 64 voxels in a single-core CPU. Meanwhile, the inference time of the network is only 0.4 seconds, a factor of 875 faster.

5 Conclusion

Deep learning models show great potential in predicting dose distributions for INTRABEAM, especially the MultiResUnet model. This study has shown that 3D dose calculations can be accurately obtained with a trained neural network in less than a second, allowing for real-time treatment planning.

References

[1] Ronneberger, O., et al (2015). U-net: Convolutional networks for biomedical image segmentation. In Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18 (pp. 234-241).

[2] Ibtehaz, N., et al. (2020). MultiResUNet: Rethinking the U-Net architecture for multimodal biomedical image segmentation. Neural networks, 121, 74-87.

[3] Ibáñez, P., et al (2021). XIORT-MC: A real-time MC-based dose computation tool for low-energy X-rays intraoperative radiation therapy. Medical Physics, 48(12), 8089-8106.

[4] Yun, S., et al. (2019). Cutmix: Regularization strategy to train strong classifiers with localizable features. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 6023-6032).

[5] Kingma, D. P., et al. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

[6] X. Mao, et al. "Rapidbrachydl: rapid radiation dose calculations in brachytherapy via deep learning", International Journal of Radiation Oncology*Biology*Physics 108, 802 (2020).

[7] Mashayekhi, M., et al. (2022). Site-agnostic 3D dose distribution prediction with deep learning neural networks. Medical physics, 49(3), 1391-1406.

[8] T. I. Götz, et al. "A deep learning approach to radiation dose estimation", Physics in Medicine & Biology 65, 035007 (2020).

GPU-based Monte Carlo method for mixed beam radiation therapy electron beamlet calculation.

Veng Jean Heng[1], Marc-André Renaud[2], and Jan Seuntjens[1,3,4,5]

[1]Medical Physics Unit, McGill University, Montreal, Canada
[2]Gray Oncology Solutions, Montreal, Canada
[3]Princess Margaret Cancer Centre, Toronto, Canada
[4]Department of Radiation Oncology, University of Toronto, Toronto, Canada
[5]Department of Medical Biophysics, University of Toronto, Toronto, Canada

Abstract Calculation of robust electron beamlets for Mixed Beam Radiation Therapy (MBRT) is a computationally expensive process due to the need for Monte Carlo-based approaches. In this work, we use the GPU-based Pre-calculated Monte Carlo (PMC) algorithm to speed up electron dose calculation. PMC relies on pre-calculated electron tracks, generated ahead of time using the full Monte Carlo EGSnrc code. At run time, the tracks are used as look-ups to raytrace electrons. A photon transport method is added onto the PMC code to properly handle dose contributions due to bremsstrahlung and contamination photons. Excellent agreement in electron beamlets calculated with energies ranging from 6 to 20 MeV were found with an average latent uncertainty of 1.3%. Compared to single core EGSnrc calculations, a 12 MeV electron beamlet was calculated over 1800 times faster with PMC on the NVIDIA Titan RTX GPU.

1 Introduction

Mixed Beam Radiation Therapy (MBRT) is a novel technique that aims to combine the use of both external electron and photon beams in radiotherapy treatment plans [1–8]. For superficial tumours, it offers superior sparing of underlying normal tissue while providing equivalent target coverage to photon-only plans [1, 3–5, 8]. With MBRT, electron fields are collimated using the photon multileaf collimators (MLC), thus providing larger degrees of freedom in terms of deliverable aperture shapes and number. However, for treatment plan optimization purposes, beamlets for each beam angles must be pre-calculated. With robust optimization, the number of beamlets is then multiplied by the number of robust scenarios that are accounted for. To calculate robust electron beamlets for MBRT treatment planning, Renaud et al. [6] had used the Monte Carlo EGSnrc code, which required calculation times on the order of a week for all beamlets required in a typical plan. Although many analytical techniques have been developed for fast dose calculation of photon doses, this has not been the case for electron doses due to their erratic scattering. State-of-the-art treatment planning systems rely on fast implementations of the Monte Carlo method through usage of pre-calculated datas.

One such implementation of a fast Monte Carlo dose calculation algorithm was the Pre-calculated Monte Carlo (PMC) code first published by Jabbari et al. in 2009 [9]. The core idea of this technique was that the stochastic trajectory of an electron and of its secondary particles could be pre-sampled ahead of time and then used as look-ups at run time. Using the established EGSnrc Monte Carlo code, electrons would be generated in an identical initial direction and energy and transported through a homogeneous medium of interest. For each electron, in pre-determined step sizes, the position, direction and energy would be recorded until the electron reached a minimum threshold energy, called the cut-off energy. Any events along its path that would set in motion a new particle such as secondary electrons or bremsstrahlung photons would also be recorded. The compilation of these step information and events constitute an electron track. A sufficiently high number of these tracks would be generated for each medium and at differing energy intervals such as to cover the whole energy range of clinical electron beams. At run time, to transport an electron, a track is randomly selected at a suitable initial energy and used to transport the electron step by step through raytracing. Any secondary particles created along the track would be added to a stack to be transported subsequently. The dose deposited would be assumed to be uniform along an electron step. If an electron travels a pathlength r_v within a voxel v, then the dose deposited to the voxel e_{dep}, v can be calculated as:

$$e_{\text{dep},v} = e_{\text{dep}} \frac{r_v}{r}, \tag{1}$$

where r is the total length of the step of the electron track along which the electron travelled through v and e_{dep} is the total energy deposition along that step. As the electron is assumed to be going in a straight time along each step, a larger step size would induce a larger error in the location of the dose deposition. However, smaller step sizes would result in larger track sizes which must fit within the video random access memory (VRAM) at run time.

Although the generation of a track bank is done with EGSnrc, this time-consuming step is only performed once ahead of time and is independent of the treatment machine or the patient geometry. The time spent during this step can therefore be ignored for practical timing comparison. This method would provide a significant speed-up compared to conventional Monte Carlo methods as no costly interaction sampling

would then have to be performed during run time. The only significant computation time would be stemming from ray-tracing the electron voxel-by-voxel along the pre-calculated steps. Jabbari et al. observed speedups by a factor of 40 with PMC compared to EGSnrc calculations with dose discrepancies on the order of 2%. In 2015, Renaud et al. [10] published a graphics proccesing unit (GPU) implementation of the PMC technique which handled both electron and proton transport. They also quantified the uncertainty induced from the finite-size of the track bank as a latent uncertainty on PMC dose calculations. A second component of the PMC dose uncertainty is due to the statistical uncertainty arising from the finite number particles simulated at run time. While the latter can be reduced by increasing running more histories, the former is constant for a pre-generated track bank. If a PMC calculation is performed at high sufficiently histories N such that the statistical uncertainty can be ignored, then the latent uncertainty σ_L can be estimated as the root mean square deviation of local residuals between PMC doses D^{PMC} and benchmark doses D^b, calculated using a conventional Monte Carlo code such as EGsnrc:

$$\sigma_L = \lim_{N\to\infty} \sqrt{1/N \sum_V \left(\frac{D_V^{\text{PMC}} - D_V^b}{D_V^b} \right)^2}, \tag{2}$$

where the summation is performed over all voxels V.
Although the PMC code had been validated to provide accurate electron dose calculations when compared with EGSnrc by Renaud et al. [10], it was done without accounting for photon transports. However, in megavoltage external electron beam, bremsstrahlung photons and contamination photons generated in the gantry head account for a significant portion of the patient dose and must be corrected for. Therefore, to be adequate for beamlet calculations, in this work, a photon transport method was added to the PMC code.

2 Method

2.1 Photon transport

Bremsstrahlung photon creation events in electron tracks were recorded at pre-generation time in EGSnrc. During live calculations, each GPU thread loads up a particle from a phase space source file. If the particle is an electron, it is transported according to its track data until they either reach E<`ecut` or cross to a different medium. For all calculations presented in this paper, an electron total energy cutoff of `ecut`=0.7 MeV was used. If secondary electrons or bremsstrahlung photons are generated along the track, they are added to a stack of particles for the thread to transport subsequently. For photon transport, a method similar to the ones used in the EGS4 report [11] and by Fippel [12] is applied. The mass attenuation coefficient for photoeletric, Compton and pair production interactions in water are obtained from NIST [13] and are initialized on the GPU as texture objects. For each photon of energy E, its corresponding total mass attenuation coefficient $\mu(E)/\rho$ is fetched from the texture objects based on its energy E. Let $z(E)$ be the distance to be travelled by a photon. The probability that a photon interaction occurs within this distance $z(E)$ can be written as:

$$P(z(E)) = 1 - e^{-\mu(E)z(E)}. \tag{3}$$

Using the direct sampling method, we let $P(z(E))$ be represented by a uniformly distributed random number ξ between 0 and 1. By directly inverting Eq. 3, we obtain:

$$z(E) = -\frac{1}{\mu(E)} \ln(1-\xi). \tag{4}$$

As $1-\xi$ is also uniform between 0 and 1 for $\xi \in (0,1)$, the density-normalized distance $z(E)/\rho$ to be travelled by the photon before an interaction occurs can then be sampled as:

$$z(E)/\rho = -\frac{1}{\mu(E)/\rho} \ln(\xi'), \tag{5}$$

where ξ' is a uniformly distributed random number between 0 and 1. The photon is then ray-traced by accounting for each voxel's density ρ until the distance $z(E)$ has been fully travelled. Photons are transported as long as their energy is higher than `pcut`=0.01 MeV and while they remain within the phantom geometry. Photons below the cutoff energy `pcut` are forced to deposit all their energy in the voxel they are found in. Once the photon has travelled the full distance $z(E)$, one of the three interaction types is sampled from the interaction-specific mass attenuation coefficients. The methods described in EGSnrc [14] to sample energy and scattering angle of the secondary particles generated by each interaction type are applied. For pair production, secondary positrons are handled as if they were electrons.

Each thread is responsible for transporting one particle sampled from the initial source. Any ensuing secondary particle is appended to the thread's stack of particles and transported subsequently. For radiotherapy applications, all body tissue is treated as density scaled water. However, as the initial particle source must be transported from outside the patient body, the PMC code must also handle transporting particles in air. For electrons, the tracks are pre-generated in air with a larger maximum step size of 1 cm and lower history of 4000 tracks per energy. This is in contrast to tracks in water being generated with a maximum step size of 0.5 mm and with 40,000 tracks per energy [10]. This reduces the memory size of the air track bank such that it can be loaded onto the GPU simultaneously. If an electron is found to be outside the patient geometry (i.e. in air), the track segment is first verified to intersect the patient's bounding box [15] before performing more costly voxel-by-voxel raytracing. Any particle found to exit the phantom geometry is immediately discarded.

For beamlet generation, at each energy, a Varian-provided phase space file is transported through a TrueBeam model in BeamNRC to generate an intermediate phase space file at the MLC plane. The intermediate phase space is then further divided into 1080 sub-phase space files by binning particles into a regular square grid. Each sub-phase space file thus consists of particles found in a square area at the MLC plane that projects to 1×1 cm^2 at isocenter. Each beamlet is thus calculated by using its corresponding sub-phase space file as its particle source.

2.2 Validation and benchmark

For beamlet generation, at each energy, a Varian-provided phase space file is transported through a TrueBeam model in BeamNRC to generate an intermediate phase space file at the MLC plane. The intermediate phase space is then further divided into 1080 sub-phase space files by binning particles into a regular square grid. Each sub-phase space file thus consists of particles found in a square area at the MLC plane that projects to 1×1 cm^2 at isocenter. Each beamlet is thus calculated by using its corresponding sub-phase space file as its particle source.

To validate the accuracy of PMC dose calculations, percentage depth dose (PDD) curves were calculated for 5 electron energies (6, 9, 12, 16, and 20 MeV) in a homogeneous water slab of dimension $25.6\times25.6\times25.6$ cm^3. A profile curve is also traced at a depth of 1.5 cm. For each energy, one such aforementioned sub-phase space file was used as the particle source. The calculations are compared to Monte Carlo calculations using the same sub-phase space files as particle sources on EGSnrc. The latent uncertainty is estimated by using the formalism presented by Renaud et al. [10]. All simulations are configured such that the water phantom is located at 100 cm source-to-surface distance (SSD). The same phantom file (`.egsphant`) is used for both calculation methods with a voxel size of $0.2\times0.2\times0.2$ mm^3. For reference, a PMC calculation without photon transport is also included. The computation time for a beamlet calculation to a type-A uncertainty of 1% was recorded with PMC on 1 NVIDIA TitanRTX GPU and compared to the same calculation with EGSnrc on a single Intel(R) Xeon(R) Gold 6140 CPU core as well as compared to using all 160 CPU cores available on our computer cluster in Fig. 3.

3 Results and Discussion

Both the PDD (Fig. 1) and off-axis profile (Fig. 2) plots showed excellent agreement between PMC and EGSnrc. Sufficient histories were used with both calculation methods to reach a type-A uncertainty of less than 0.2% on voxels receiving at least 50% of the maximum dose. The latent uncertainty as averaged over the 5 electron beamlet energies

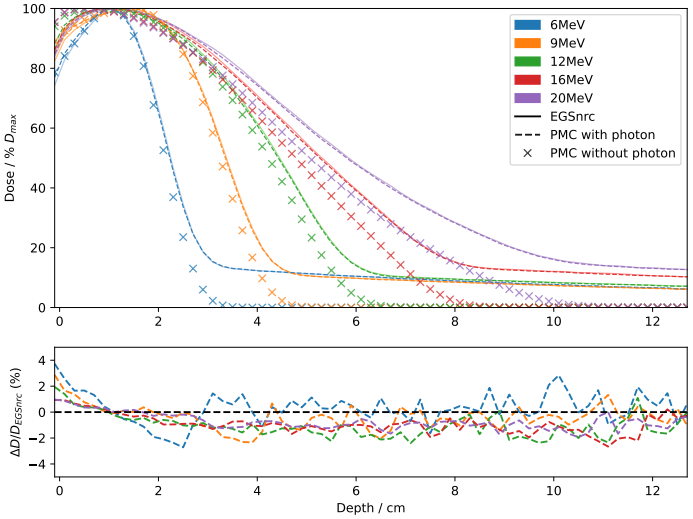


Figure 1: Validation of PMC PDD against EGSnrc in a homogeneous water slab. Both EGSnrc and PMC type-A uncertainties are below 0.2% in voxels receiving at least 50% of the maximum dose. The PMC dose as calculated without photon transport is included for comparison. Local residuals are plotted for the PMC dose with photon transport. Error bars were omitted for clarity as the type-A uncertainties are too small to be visually rendered.

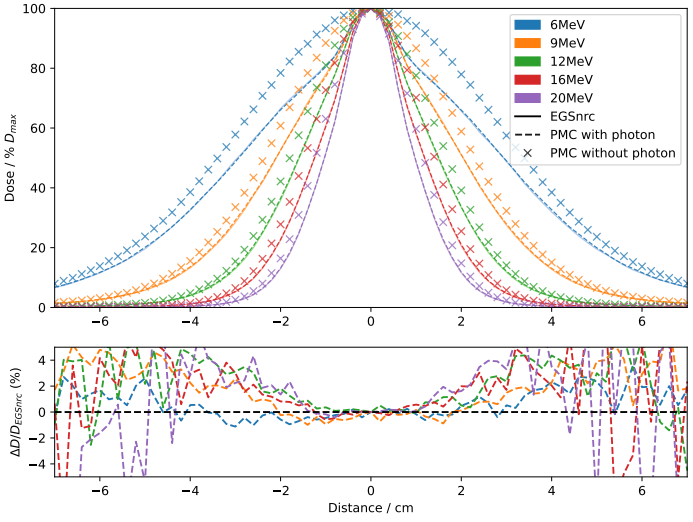


Figure 2: Validation of PMC profile against EGSnrc in a homogeneous water slab at 1.5 cm depth.

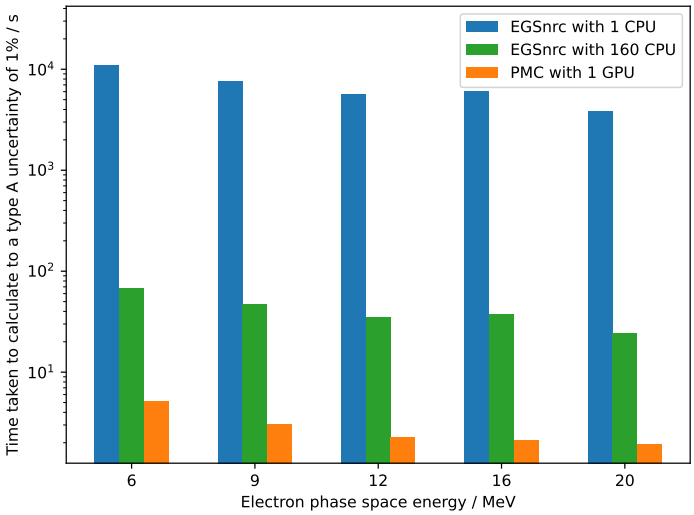


Figure 3: Comparison of time taken to calculate to 1% type-A uncertainty by EGSnrc and PMC. For a more practical comparison, the calculation time when using our entire computer cluster (160 CPUs) is also included.

was 1.3% as evaluated in voxels receiving at least 20% of the maximum dose [10]. The handling of photon transport in PMC was critical in the rapid dose falloff region and in the bremsstrahlung tail of electron beamlet PDDs. The total GPU memory usage was around 10 GB, mainly due to the track bank size.

To reach the same type-A uncertainty for a 12 MeV beamlet calculation, PMC is more than 1800 times faster than EGSnrc on a single CPU. Even when compared to our small computer cluster, PMC calculations on a single commercially available GPU outperforms EGSnrc by over a factor of 10. This speed-up in the electron beamlet computation time is crucial for the clinical translation of MBRT. Indeed, for robust optimizations, beamlets must be calculated in each robust scenarios, at each control points, and at each electron energy. For a typical MBRT plan, this can total to over 30 000 beamlets. With EGSnrc, this beamlet calculation step took on the order of 1 week to compute while PMC does it on the order of 1 day. Furthermore, as consumer GPUs are much more affordable to purchase than computer clusters with hundreds of CPU cores, PMC allows for a wider and more accessible implementation of MBRT in the clinic.

4 Conclusion

To speed up MBRT treatment planning times, a GPU-based Monte Carlo dose calculation method was improved to be used for MBRT electron beamlet calculations. A simplified photon transport method was added to PMC to handle bremsstrahlung photons and contamination photons. PMC allowed for faster Monte Carlo dose calculation of MBRT electron beamlets compared to EGSnrc by over 3 order of magnitude.

References

[1] B. A. Palma, A. U. Sánchez, F. J. Salguero, et al. “Combined modulated electron and photon beams planned by a Monte-Carlo-based optimization procedure for accelerated partial breast irradiation”. *Physics in Medicine and Biology* 57.5 (2012), pp. 1191–1202. DOI: 10.1088/0031-9155/57/5/1191.

[2] C. Míguez, E. Jiménez-Ortega, B. A. Palma, et al. “Clinical implementation of combined modulated electron and photon beams with conventional MLC for accelerated partial breast irradiation”. *Radiotherapy and Oncology* 124.1 (2017), pp. 124–129. DOI: 10.1016/j.radonc.2017.06.011.

[3] M.-A. Renaud, M. Serban, and J. Seuntjens. “On mixed electron-photon radiation therapy optimization using the column generation approach”. *Medical Physics* 44.8 (2017), pp. 4287–4298. DOI: 10.1002/mp.12338.

[4] S. Mueller, M. K. Fix, A. Joosten, et al. “Simultaneous optimization of photons and electrons for mixed beam radiotherapy”. *Physics in Medicine and Biology* 62.14 (2017), pp. 5840–5860. DOI: 10.1088/1361-6560/aa70c5.

[5] S. Mueller, P. Manser, W. Volken, et al. “Part 2: Dynamic mixed beam radiotherapy (DYMBER): Photon dynamic trajectories combined with modulated electron beams”. *Medical Physics* 45.9 (2018), pp. 4213–4226. DOI: https://doi.org/10.1002/mp.13085.

[6] M.-A. Renaud, M. Serban, and J. Seuntjens. “Robust mixed electron-photon radiation therapy optimisation”. *Medical Physics* 0.ja (2019). DOI: 10.1002/mp.13381.

[7] E. Heath, S. Mueller, G. Guyer, et al. “Implementation and experimental validation of a robust hybrid direct aperture optimization approach for mixed-beam radiotherapy”. *Medical Physics* 48.11 (2021), pp. 7299–7312. DOI: https://doi.org/10.1002/mp.15258.

[8] V. J. Heng, M. Serban, M.-A. Renaud, et al. “Robust mixed electron-photon radiation therapy planning for soft tissue sarcoma”. *Medical Physics* 50.10 (2023), pp. 6502–6513. DOI: https://doi.org/10.1002/mp.16709.

[9] K. Jabbari, P. Keall, and J. Seuntjens. “Considerations and limitations of fast Monte Carlo electron transport in radiation therapy based on precalculated data”. *Medical Physics* 36.2 (2009), p. 530. DOI: 10.1118/1.3058480.

[10] M. A. Renaud, D. Roberge, and J. Seuntjens. “Latent uncertainties of the precalculated track Monte Carlo method”. *Medical Physics* 42.1 (2015), pp. 479–490. DOI: 10.1118/1.4903502.

[11] W. R. Nelson, H. Hirayama, and D. W. Rogers. “EGS4 code system” (Dec. 1985).

[12] M. Fippel. “Fast Monte Carlo dose calculation for photon beams based on the VMC electron algorithm”. *Medical Physics* 26.8 (1999), pp. 1466–1475. DOI: https://doi.org/10.1118/1.598676.

[13] J. Hubbell and S. Seltzer. *Tables of X-Ray Mass Attenuation Coefficients and Mass Energy-Absorption Coefficients 1 keV to 20 MeV for Elements Z = 1 to 92 and 48 Additional Substances of Dosimetric Interest*. en. 1995-01-01 1995.

[14] I. Kawrakow, D. Rogers, E. Mainegra-Hing, et al. *EGSnrc toolkit for Monte Carlo simulation of ionizing radiation transport*. doi:10.4224/40001303 [release v2021]. 2000.

[15] A. Williams, S. Barrus, R. K. Morley, et al. “An Efficient and Robust Ray-Box Intersection Algorithm”. *Journal of Graphics Tools* 10.1 (2005), pp. 49–54. DOI: 10.1080/2151237X.2005.10129188.

Revisiting Over Three Years of Clinical Application of Monte Carlo Based Patient Quality Assurance at the Maastro Proton Therapy Center

Ilaria Rinaldi[1], Jan Gajewski[2], Angelo Schiavi[3], Nils Krah[4], Antoni Rucinski[2], Vincenzo Patera[3], Sebastiaan Nijsten[1]

[1]Department of Radiation Oncology (Maastro), GROW School for Oncology, Maastricht University Medical Centre+, Maastricht, The Netherlands
[2]Institute of Nuclear Physics PAN, Krakow, Poland
[3]University of Rome, Sapienza, Rome, Italy
[4]CREATIS, CNRS, UMR5220, Centre Léon Bérard, INSA Lyon, France

Abstract
In our proton facility, the Maastro Proton Therapy Center in Maastricht, we have entirely supplanted traditional measurements with machine logfile-based patient specific quality assurance (PSQA). This innovative approach is underpinned by the independent GPU-accelerated Monte Carlo code Fred. This study presents a comprehensive report on the clinical application of Monte Carlo based PSQA over a period exceeding three years at our proton therapy center.

1 Introduction

In the context of proton beam therapy, maintaining concordance between the planned and delivered doses is imperative. Ensuring the absence of disparities arising from factors such as planning intricacies, data transfer nuances, and delivery errors is paramount. Consequently, a rigorous validation process is adopted for each clinical plan in proton beam therapy. This validation, commonly conducted as part of patient-specific quality assurance (PSQA), involves the measurement of absolute dose distributions for each individual treatment field before the initiation of treatment, followed by a meticulous comparison with the predicted dose distributions. In an effort to streamline the beam-time required for PSQA measurements and enhance overall PSQA quality, the utilization of machine log files has been explored.

Machine log files can be procured either through a dedicated dry-run irradiation or inherently during the delivery of a treatment fraction. These log files encapsulate crucial information regarding spot position, dose, and energy, which can be restructured into a treatment plan and used to reconstruct the actual delivered dose. This reconstructed dose can then be compared against the originally prescribed planning dose.

For the generation of an independent dose distribution based on log files, the adoption of a Monte Carlo dose engine distinct from the clinical Treatment Planning System (TPS) is imperative.

At the Maastro Proton Therapy Center, we opted for the utilization of the GPU-accelerated Monte Carlo code Fred [1] owing to its efficient computational performance, precision, and capacity to accommodate the idiosyncrasies of our proton machine [2]. Fred can recompute a proton treatment plan using 1% of the total particles within a matter of minutes.
In the scope of this study, we present our endeavors to transition from a measurement-based approach to a simulation-based one for PSQA at the Maastro Proton Therapy Center. This paradigm shift involves the integration of Fred as an independent dose calculation engine, coupled with machine log files.

2 Materials and Methods

We devised a hybrid methodology for the integration of our proton machine system's beam model into Fred. The validation process of our beam model successfully met all clinical acceptance criteria.

The beam model and machine geometry of the therapeutic Mevion S250i proton accelerator were meticulously incorporated into the independent Monte Carlo code Fred, as extensively detailed in [3]. Given the Mevion S250i's unique delivery system, encompassing a range modulation system and an adaptive aperture in the beam path, special consideration was given during implementation.

A validation process, involving extensive comparisons with commissioning measurements [2] and alignment with the clinical TPS, was undertaken. The pristine proton beam is described by a phase space model situated in a plane at the scanning magnet position, while the dynamically extendable nozzle housing the beam monitor system, range modulation system, and adaptive aperture is explicitly simulated. Validation of Fred met all established clinical acceptance criteria [3].

For a comprehensive patient simulation, input data can be extracted either from the treatment plan or from the machine log files of the accelerator. An automated pipeline adeptly extracts CT images, RTstruct, and RTplan DICOM files from the clinical database system, linking them to the corresponding measured files of the proton machine. Subsequently, a new RTplan file is generated based on the log files and serves as input to Fred. The ensuing dosimetric validation report, inclusive of gamma index and dose volume histograms, is automatically generated. Each time a

new plan or log file in the clinical database, A comprehensive report, featuring visualizations and a quantitative dosimetric summary, is automatically produced.

A fully operational pipeline has been established, and the routine review of automatic verification reports has become part of our daily clinical practices. The recalculation of patient dose distributions takes mere minutes, enabling swift on-the-fly execution of the pipeline. During the initial months following the introduction of this pipeline in our clinical routine in the summer of 2020, full 3D Monte Carlo dose recalculations for QA plans complemented the conventional 2D patient-QA measurements. In the contemporary setting, Monte Carlo based QA has entirely supplanted patient-QA measurements. Nowadays, patient-QA measurements are only performed in the rare situation that a recalculation results in a failed verification.

3 Results

An exhaustive analysis of the generated data will be presented at the ICCR. To date, our pipeline has been used to recalculate over 4000 clinical plans, with 2812 of them originating from machine logfiles. The chronological record of the passing rate and its distribution per clinical indication are depicted in Figure 1. The Gamma index pass rate in percentage (%) was calculated using a gamma criterion of 3 mm and 3%. Notably, in December 2022 (indicated by the black vertical dotted line in the left panel), the scoring of dose to water in Fred was initiated. Subsequent to this implementation, the pass rate has consistently reached 99.7%

Sparring the measurements of more than 4000 clinical plans implies a reduction of approximately one working month (21 days) in routine QA annually, considering an average of 30 minutes for a PSQA measurement. Notably, this method enables the identification of cases that might have eluded detection through standard PSQA measurements in homogeneous media and that could potentially exert a non-negligible impact on clinical patient treatments, as elucidated in Figure 2.

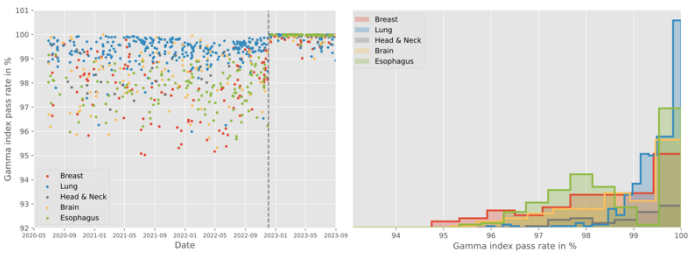


Figure 1: Gamma index pass rate in % (gamma criterion of 3 mm and 3%) for log file based PSQA as a function of time (left) and histogram per clinical indication (right). In December 2022 (black vertical dotted line), we started scoring dose to water in Fred. Since then, the pass rate is 99,7%.

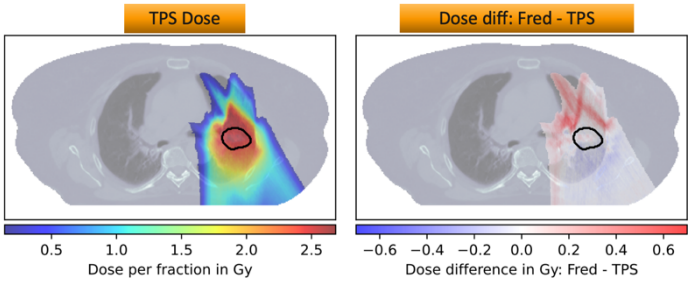


Figure 2: Example of a clinical case which would not have been detected with standard PSQA measurements in homogeneous media and which would have had a non-negligible impact in clinical patient treatments.

4 Discussion

This work delineates the establishment of a comprehensive pipeline for the automated recalculation of patient plans n proton beam therapy, leveraging both the clinical Treatment Planning System (TPS) plan and machine log files derived from the proton accelerator. The automated workflow adeptly extracts requisite patient information, including CT images and treatment plans, from the clinical database system. Recalculations performed on the clinical plan serve to automatically verify the integrity of the planning process and plan transfer to the proton accelerator. On the other hand, recalculations on the machine log files ascertain the accuracy of treatment delivery by the proton accelerator.
The dose recalculation pipeline followed by the creation of a dosimetric validation report alongside a quantitative dosimetric summary is triggered automatically by the upload of each new plan or log file in the clinical database. We are currently investigating a dedicated gamma index analysis tailored to and optimized for every clinical indication. The dose recalculation and verification pipeline could also be extended in the future to include considerations based on dose volume histograms.

5 Conclusion

The PSQA pipeline developed at Maastro significantly streamlines daily operations by furnishing invaluable and comprehensive 3D dosimetric information, eliminating the necessity for dedicated beam time.

References

[1] Schiavi A, Senzacqua M, Pioli S, Mairani A, Magro G, Molinelli S, Ciocca M, Battistoni G and Patera V. Fred: A GPU-accelerated fast-Monte Carlo code for rapid treatment plan recalculation in ion beam

therapy. Physics in Medicine and Biology 62 (2017) 7482–7504. doi:10.1088/1361-6560/aa8134.

[2] Vilches-Freixas G, Unipan M, Rinaldi I, Martens J, Roijen E, Almeida IP, Decabooter E and Bosmans G. Beam commissioning of the first compact proton therapy system with spot scanning and dynamic field collimation. The British journal of radiology 93 (2020) 20190598. doi:10.1259/bjr.20190598

[3] Gajewski J, Schiavi A, Krah N, Vilches-Freixas G, Rucinski A, Patera V, and Rinaldi I. Implementation of a compact spot-scanning proton therapy system in a GPU Monte Carlo code to support clinical routine. Frontiers in Physics 8 (2020). doi:10.3389/fphy.2020.578605

A stochastic model to estimate dose to circulating lymphocytes during brain radiotherapy

François de Kermenguy[1], Nathan Benzazon[1], Pauline Maury[1,2], Rémi Vauclin[3], Meissane M'hamdi[1], Vjona Cifliku[1], Elaine Limkin[1,2], Ibrahima Diallo[1], Daphné Morel[1], Candice Milewski[2], Céline Clémenson[1], Michele Mondini[1], Eric Deutsch[1,2], Charlotte Robert[1,2]

[1]Université Paris-Saclay, Gustave Roussy, Inserm U1030, Radiothérapie Moléculaire et Innovation Thérapeutique, 94800, Villejuif, France
[2]Gustave Roussy, Département de radiothérapie, 94800, Villejuif, France
[3]TheraPanacea, Paris, France

Abstract Severe radiation-induced lymphopenia occurs in 40% of patients treated for primary brain tumors and is an independent risk factor of poor survival outcomes. We developed an in-silico model that estimates the radiation doses received by lymphocytes during brain irradiation. The simulation consists of two interconnected compartmental models describing the slow recirculation of lymphocytes between lymphoid organs ($\mathcal{M}_1$) and the bloodstream ($\mathcal{M}_2$). We used dosimetry data from 33 patients treated for glioblastoma to compare three scenarios: H1) lymphocytes circulation only in the bloodstream; H2) lymphocytes recirculation between lymphoid organs H3) lymphocytes recirculation between lymphoid organs and deep-leaning computed out-of-field dose to head-and-neck lymphoid structures. According to H3 case, almost all lymphocytes were irradiated after treatment, at a mean dose of 265.6 ± 48.5mGy. Our study is the first to take into account out-of-field doses and recirculation in lymphocyte dose assessment during brain irradiation. Our results demonstrate the need to clarify the indirect effects of irradiation on lymphopenia, in order to potentiate the combination of radio-immunotherapy or the abscopal effect.

1 Introduction

Severe radiation-induced lymphopenia (sRIL) occurs in 40% of patients treated for primary brain tumors with radiochemotherapy and is an independent risk factor for inferior overall survival outcomes [1]. One hypothesis put forward to explain the occurrence of sRIL after brain irradiation is the cytotoxic doses received by the blood lymphocyte pool. Therefore, several research groups have been interested in developing models to evaluate the doses received by the peripheral blood [2]. However, none has yet considered the out-of-field (OOF) dose to the head and neck (H&N) region rich in lymphoid structures, or the slow recirculation and homing of lymphocytes between lymphoid organs such as the spleen or lymph nodes and blood (which only contains about 5% of the total lymphocyte pool at any given time [3]). To tackle these shortcomings, this work proposes a new model to assess more accurately the doses received by recirculating lymphocytes during brain irradiation.

2 Materials and Methods

To assess the doses received by the lymphocyte pool during irradiation, we modeled their dynamics within the organs. We distinguished two distinct but interdependent processes: (i) the slow recirculation of lymphocytes between lymphoid organs, mediated by specific biological process [4], and (ii) the rapid circulation of lymphocytes through all organs via the bloodstream. These 2 phenomena were modelled in this work by implementing 2 interconnected compartmental models, named $\mathcal{M}_1$ and $\mathcal{M}_2$ respectively and shown in Figure 1. Both models are associated with a stochastic continuous-time semi-Markov chain process to simulate lymphocyte paths between compartments during the whole treatment. In practice, the implementation was based on a Kinetic Monte Carlo algorithm. Each compartment of the $\mathcal{M}_1$ and $\mathcal{M}_2$ models were associated with 2 parameters: 1) a continuous distribution of lymphocyte residence times, which enables a residence time to be stochastically determined for each lymphocyte passage through that compartment via inverse transform sampling, and 2) a vector of transition probabilities to the connected compartments, in order to stochastically determine which compartment the lymphocyte will migrate to.

Model $\mathcal{M}_1$ (Figure 1) is a modified model based on the work of Ganusov et al. [5], which describes lymphocyte recirculation via *circulating blood* between *spleen*, *subcutaneous lymph nodes (SCLNs*), *mesenteric lymph nodes (MLNs)* and *Peyer's Patches (PPs)*, using rats experimental data. In accordance with [5], residence times were distributed according to an exponential law with a parameter $\lambda_{M1} = \frac{1}{Mean\,residence\,time}$ for the *Spleen*, *PPs* and *Blood Circulation* compartments, and according to a gamma distribution with a shape $k'_{M1} = 2$ and a scale $\theta_{M1} = \frac{Mean\,residence\,time}{k'_{M1}}$ for the *NCLNs* and *MLNs* compartments.

Model $\mathcal{M}_2$ (Figure 1) was based on the ICRP Publication 89 [6], which describes a transition matrix for blood flow between 28 organs in the male and female human body. Mean residence time values were calculated considering a total blood volume of 5.3 L and 3.9 L and a cardiac output of 6.5 L.min^{-1} and 5.9 L.min^{-1} for male and female patients respectively. In accordance with the data given in [7], residence times were distributed according to a Weibull distribution with a shape $k_{M2} = 2$ and a scale $\lambda_{M2} = \frac{Mean\,residence\,time}{\Gamma(1+\frac{1}{k_{M2}})}$, with Γ the gamma function.

A cohort of 33 patients with glioblastoma treated after 2018 with conventional VMAT (Volumetric Modulated Arc Therapy) irradiation was created. To simulate a treatment, each patient was characterized by 1) his gender, which defined which variant of the $\mathcal{M}_2$ model (male or female) to apply, and 2) his irradiation parameters. For a given patient, a conventional VMAT treatment was defined

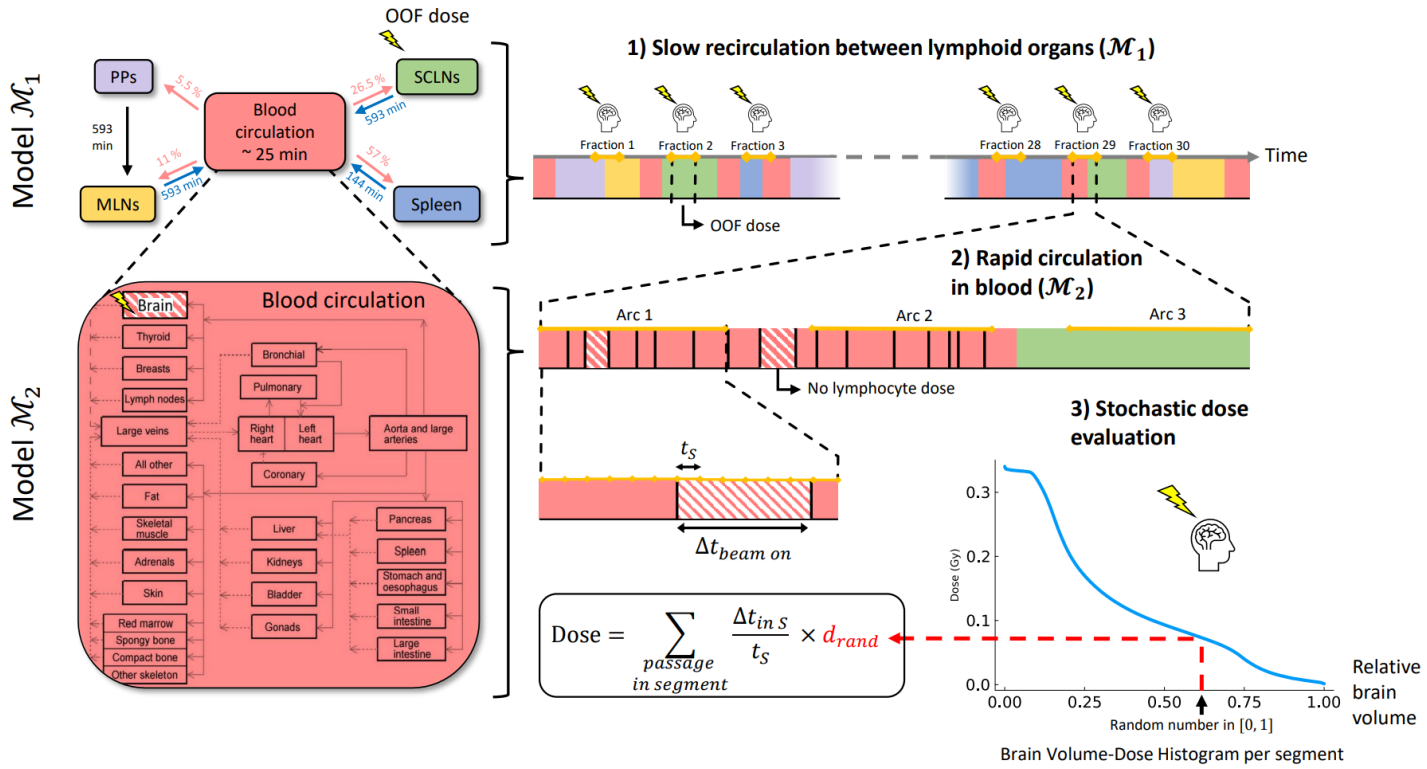


Figure 1 *On the left, compartment models $\mathcal{M}_1$ and $\mathcal{M}_2$ describing lymphocyte recirculation between lymphoid organs and blood flow, respectively. On the right, a schematic representation of a recirculating lymphocyte path and the dose it receives in the continuous time semi-Markov chain simulation SCLNs: subcutaneous lymph nodes, MLNs: mesenteric lymph nodes, PPS: Peyer's Patches (PPs), OOF: Out-of-field*

by a succession of N_F fractions F_i ($i \in [1, N_F]$) spaced 24 hours apart. Each fraction F_i was composed of a succession of N_A arcs $A_{i,j}$ ($j \in [1, N_A]$), each lasting $t_{A_{i,j}}$ seconds and spaced 5 seconds apart. Each arc $A_{i,j}$ was itself subdivided into N_S successive segments $S_{i,j,k}$ ($k \in [1, N_S]$), with the assumption that its duration was $t_{S_{i,j,k}} = \frac{t_{A_{i,j}}}{N_S}$. For each segments $S_{i,j,k}$, a whole brain dose-volume histogram $DVH_{i,j,k}$ were recalculated using a Collapsed-Cone Convolution dose engine provided by TheraPanacea company (Paris, France).

In order to evaluate the doses received by lymphocytes using the models described above, the logic is as follows (Figure 1): since the brain contains few lymphoid organs, a given lymphocyte circulating in $\mathcal{M}_1$ is likely to be irradiated in the field if it is in the $\mathcal{M}_1$ *Blood circulation* compartment during a segments $S_{i,j,k}$. In this case, the lymphocyte enters $\mathcal{M}_2$ via the *Large Veins*. The latter is then effectively irradiated in the field if it is in the *Brain* during the beam-on time, and the received dose is stochastically evaluated using an inverse transform sampling on the whole brain $DVH_{i,j,k}$.

Because of the high radiosensitivity of lymphocytes [8], the contribution of low doses outside the brain irradiation field and delivered to the H&N region, rich in lymphoid structures, was also considered. To do this, the right and left lymph nodes areas, including Ia, Ib, II, III, IVa, V, VIIa et VIIb regions [9], were automatically contoured using ART-Plan software v1.11.5 (Therapanacea, Paris, France), and an in-house deep learning neural network [10] was used to generate "whole body" dose map from the patient's external contour and the in-field dose map from the TPS. Together, lymphoid structures contour and H&N OOF dose map allow calculation of the average $\overline{OOF}$ dose to lymphoid structures in the H&N region for each patient: $\overline{OOF} = \frac{\sum_i d_i \times v_i}{\sum_i v_i}$, with v_i and d_i the volume and the mean dose received by the lymphoid structure i, respectively. Finally, based on the analysis of 1,200 lymph nodes performed on the Visible Human Data Set [11], having shown that 31.5 ± 0.7% of subcutaneous lymph nodes (simply defined as non-mesenteric as in [5]) are located in the H&N region, we considered that a lymphocyte had a probability of 0.315 of receiving an $\overline{OOF}$ dose if it is in the SCLNs during an irradiation fraction.

To compare different physical and biological assumptions, assumptions, the dose to lymphocytes has been computed with three cases of the model: H1) circulation only in the bloodstream i.e., lymphocytes circulate in $\mathcal{M}_2$ only; H2) with recirculation between lymphoid organs i.e. $\mathcal{M}_1$ and $\mathcal{M}_2$ interconnected; H3) with recirculation between lymphoid organs and with OOF dose to H&N lymphoid structures.

The code was developed in Julia language version 1.9.3 and simulations were performed considering 10^5 particles.

3 Results

Case H1 resulted in the irradiation of almost all the circulating blood ($V_{>0Gy}$ = 99.8 ± 0.7%) at the end of VMAT treatment, at a mean dose of 309.9 ± 74.7 mGy (Figure 2). Taking into account lymphocytes recirculation in H2 case, the fraction of lymphocytes receiving a non-zero dose dropped to 40.4 ± 10.2%, with a lower mean dose estimated at 52.6 ± 21.1 mGy, especially because irradiated lymphocytes passed through the irradiation field only 1.58 ± 0.91 times on average due to slow recirculation between lymphoid organs in the $\mathcal{M}_1$ model. The most realistic scenario, including the OOF dose component (case H3), showed that almost all the lymphocytes (97.6 ± 2.5%) were irradiated at the end of the treatment, mostly due to the OOF dose (mean dose equal to 265.6 ± 48.5 mGy).

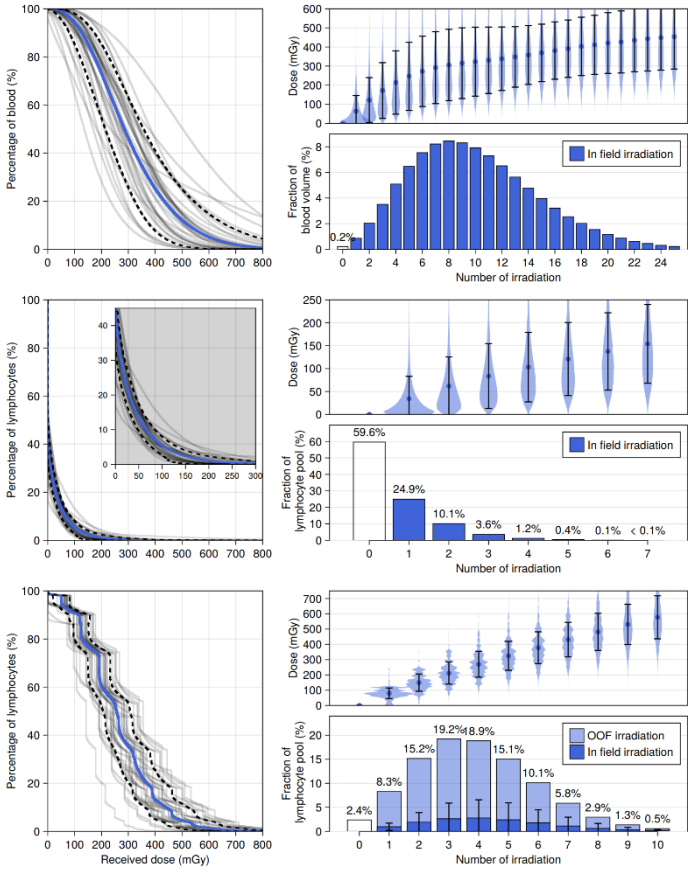


Figure 2 The left column represents Dose Volume Histograms (DVH) of blood (H1) and of the recirculating lymphocyte pool without (H2) and with (H3) consideration of the low out-of-field (OOF) dose in the lymph nodes of the head and neck region at the end of treatment. In each figure, the mean DVH over the whole cohort is shown in solid blue, the associated standard deviation in dotted line and the DVHs for each patient in the cohort in solid light grey. On the right column, the average number of lymphocyte irradiations (bottom), and the associated dose (top), averaged over the whole cohort are described

4 Discussion

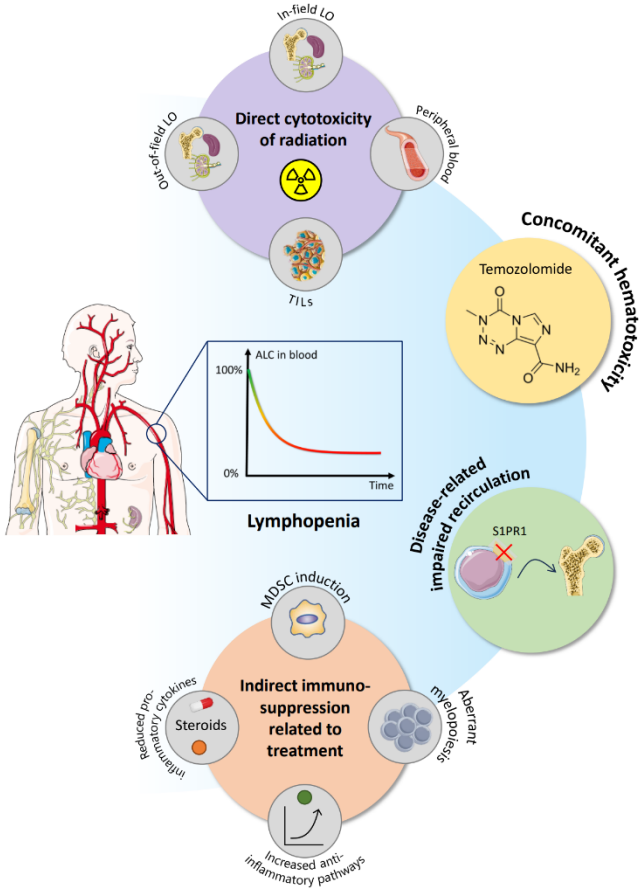


Figure 3 Several immunomodulatory factors can alter both quantitatively and qualitatively the patient blood lymphocyte composition and contribute to lymphopenia. Thus, a decrease in blood ALC, used to characterize lymphopenia, can have several different causes. The cytotoxic effect of irradiation on lymphocytes is one of these factors, but our results suggest that its relative contribution to the development of lymphopenia might be limited, especially in the context of brain irradiation. LO: Lymphoid organs, TILs: Tumor infiltrating lymphocytes, ALC: Absolute Lymphocyte Count, S1PR1: Sphingosine-1-phosphate receptor 1.

In this work, a new in-silico model, efficient in terms of calculation time and closer to biological reality, was proposed. Compared to previously published studies [2], out model highlighted the major role of OOF doses and recirculation for the correct calculation of the dose received by lymphocytes during brain irradiation. Despite the high level of radiosensitivity of lymphocytes [8], our results (regardless of the H1, H2 or H3 cases used) hardly explain the 40% sRIL rates observed clinically in patients with brain tumours treated with radiochemotherapy. These results call into question the relative contribution of the direct cytotoxic effect of irradiation on lymphocytes to the induction of RIL, and several other immunosuppressive effects will certainly have to be taken into account in order to potentiate treatments (Figure 3). First, lymphopenia can be influenced

by the concomitant exposure to systemic medications, including temozolomide and corticosteroids [12]. Then, it has been shown that intracranial tumors are intrinsically characterized by a S1P1-mediated sequestration of T cells in the bone marrow [13]. In addition, radiation can trigger an aberrant myelopoiesis in the bone marrow leading to an increase of lymphotoxic myeloid-derived suppressive cells (MDSC) in the peripheral blood [14]. All these direct and indirect mechanism could contribute to the worsening of lymphopenia upon radiotherapy treatment.

Several perspectives are open for future research. Firstly, clinical and pre-clinical data are essential to experimentally validate our results and to complete the modelling. Secondly, the recirculation model ($\mathcal{M}_1$) used is based on rats data. Data in humans would be valuable, as the immune systems of the 2 species are patiently different [15]. Moreover, several different lymphocyte subpopulations with different recirculation patterns coexist in the blood [16], and therefore models different from the $\mathcal{M}_1$ model are required. Thirdly, average residence times used in the $\mathcal{M}_2$ model, for the *Brain* compartment in particular, are those of circulating blood. However, it could be significantly higher for the lymphocytes it contains, in particular because of their ability to adhere to endothelial cells, or simply because their diameter (~7µm) is of the same order of magnitude as diameters of some cerebral capillaries [17]. Finally, Neural Network used to evaluate OOF dose was trained on pediatric data. It therefore probably overestimates OOF doses, and experimental measurements are required to validate these results

5 Conclusion

We have developed an in-silico model based on continuous time semi-Markov chain process to evaluate the dose to peripheral blood and to circulating lymphocytes after brain radiotherapy treatment. The model consider the rapid circulation of lymphocytes in blood, the recirculation of lymphocytes between lymphoid organs and the OOF dose to lymphoid structure of the H&N region. It is an important preliminary step towards a better understanding of the radiobiological mechanisms underlying RIL, and the model could be applied to other irradiation locations in the future. From a clinical point of view, reducing the incidence of RIL could not only improve the response to radiotherapy and chemoradiation, but also enhance the potential of radioimmunotherapy combinations or abscopal effect.

This work has benefited from the grant ANR-21-RHU5-0005 within the FRANCE2030 investment plan.

References

[1] F. de Kermenguy *et al.*, « Radio-induced lymphopenia in the era of anti-cancer immunotherapy », in *International Review of Cell and Molecular Biology*, Academic Press, 2023. doi: 10.1016/bs.ircmb.2023.03.002.

[2] C. Beekman *et al.*, « A stochastic model of blood flow to calculate blood dose during radiotherapy », *Phys. Med. Biol.*, oct. 2023, doi: 10.1088/1361-6560/ad02d6.

[3] V. V. Ganusov et R. J. De Boer, « Do most lymphocytes in humans really reside in the gut? », *Trends Immunol.*, vol. 28, nº 12, p. 514-518, déc. 2007, doi: 10.1016/j.it.2007.08.009.

[4] V. V. Ganusov et M. Tomura, « Experimental and mathematical approaches to quantify recirculation kinetics of lymphocytes ». bioRxiv, p. 268326, 20 août 2018. doi: 10.1101/268326.

[5] V. V. Ganusov et J. Auerbach, « Mathematical Modeling Reveals Kinetics of Lymphocyte Recirculation in the Whole Organism », *PLOS Comput. Biol.*, vol. 10, nº 5, p. e1003586, mai 2014, doi: 10.1371/journal.pcbi.1003586.

[6] ICRP Publication 89, « Basic Anatomical and Physiological Data for Use in Radiological Protection Reference Values », 2002, [En ligne]. Disponible sur: https://www.icrp.org/publication.asp?id=icrp%20publication%2089

[7] J. Shin *et al.*, « HEDOS—a computational tool to assess radiation dose to circulating blood cells during external beam radiotherapy based on whole-body blood flow simulations », *Phys. Med. Biol.*, vol. 66, nº 16, p. 164001, août 2021, doi: 10.1088/1361-6560/ac16ea.

[8] H. Paganetti, « A review on lymphocyte radiosensitivity and its impact on radiotherapy », *Front. Oncol.*, vol. 13, 2023, Consulté le: 31 août 2023. [En ligne]. Disponible sur: https://www.frontiersin.org/articles/10.3389/fonc.2023.1201500

[9] V. Grégoire *et al.*, « Delineation of the neck node levels for head and neck tumors: a 2013 update. DAHANCA, EORTC, HKNPCSG, NCIC CTG, NCRI, RTOG, TROG consensus guidelines », *Radiother. Oncol. J. Eur. Soc. Ther. Radiol. Oncol.*, vol. 110, nº 1, p. 172-181, janv. 2014, doi: 10.1016/j.radonc.2013.10.010.

[10] N. Benzazon *et al.*, « PD-0169 An innovative and versatile deep learning approach to estimate the out-of-field dose. », *Radiother. Oncol.*, vol. 182, p. S129-S130, mai 2023, doi: 10.1016/S0167-8140(23)08813-8.

[11] S. M. Qatarneh, I.-C. Kiricuta, A. Brahme, U. Tiede, et B. K. Lind, « Three-dimensional atlas of lymph node topography based on the visible human data set », *Anat. Rec. B. New Anat.*, vol. 289, nº 3, p. 98-111, mai 2006, doi: 10.1002/ar.b.20102.

[12] A. A. Brandes *et al.*, « Temozolomide 3 weeks on and 1 week off as first-line therapy for recurrent glioblastoma: phase II study from gruppo italiano cooperativo di neuro-oncologia (GICNO) », *Br. J. Cancer*, vol. 95, nº 9, Art. nº 9, nov. 2006, doi: 10.1038/sj.bjc.6603376.

[13] P. Chongsathidkiet *et al.*, « Sequestration of T cells in bone marrow in the setting of glioblastoma and other intracranial tumors », *Nat. Med.*, vol. 24, nº 9, Art. nº 9, sept. 2018, doi: 10.1038/s41591-018-0135-2.

[14] S. Ghosh *et al.*, « Radiation-induced circulating myeloid-derived suppressor cells induce systemic lymphopenia after chemoradiotherapy in patients with glioblastoma », *Sci. Transl. Med.*, vol. 15, nº 680, p. eabn6758, janv. 2023, doi: 10.1126/scitranslmed.abn6758.

[15] G. Wildner, « Are rats more human than mice? », *Immunobiology*, vol. 224, nº 1, p. 172-176, janv. 2019, doi: 10.1016/j.imbio.2018.09.002.

[16] A. J. Young, W. L. Marston, M. Dessing, L. Dudler, et W. R. Hein, « Distinct Recirculating and Non-Recirculating B-Lymphocyte Pools in the Peripheral Blood Are Defined by Coordinated Expression of CD21 and L-Selectin », *Blood*, vol. 90, nº 12, p. 4865-4875, déc. 1997, doi: 10.1182/blood.V90.12.4865.

[17] M. Marín-Padilla, « The human brain intracerebral microvascular system: development and structure », *Front. Neuroanat.*, vol. 6, p. 38, sept. 2012, doi: 10.3389/fnana.2012.00038.

Deep Learning dose engine for electron FLASH radiotherapy

L. Arsini[1,2], J. Humphreys[3], C. White[3], B. Caccia[4], A. Ciardiello[1,2], A. De Gregorio[1,2], G. Franciosini[1,2], S. Giagu[1,2], A. Muscato[5], A. Schiavi[6,2], M. Hagenbuchner[3], S. Guatelli[3], and C. Mancini-Terracciano[1,2]

[1]Department of Physics, Sapienza University of Rome, Rome, Italy
[2]National Institute of Nuclear Physics, INFN Rome, Italy
[3]University of Wollongong, Wollongong, NSW, Australia
[4]Istituto Superiore di Sanità, Rome, Italy
[5]Scuola post-laurea in Fisica Medica, Sapienza University of Rome, Rome, Italy
[6]Basic and Applied Sciences for Engineering, Sapienza University of Rome, Rome, Italy

Abstract The aim of our work is to develop a Deep Learning (DL) algorithm capable of efficiently generating dose distributions for particle beams within heterogeneous mediums, such as the human body. This model computes dose as a function of the beam parameters, as entry point, energy, and fluence with an accuracy comparable to Monte Carlo algorithms but at an exceptionally rapid pace (0.01 seconds per beam on a single GPU). The resulting dose maps hold great potential for integration into the cost function of forthcoming Radiotherapy (RT) auto - planning algorithms, particularly those based on differentiable optimization. The core of our DL architecture is a U-Net based on Graph Neural Networks, with a purpose-developed pooling technique. Trained on Geant4 Monte Carlo simulations using a patient CT scan and beam characteristics as inputs, our model achieve satisfactory performances through rigorous training and testing on diverse beam orientations and energies.

1 Introduction

External beam RT is a widely recognized and effective cancer treatment [1]. In order to deliver the correct amount of dose which satisfies the medical prescription, a "treatment plan optimization" which considers beams orientation, energy and fluence is needed. At the basis of a reliable treatment plan optimization there should be a robust method to estimate the dose delivered by the therapeutic beams. Nonetheless, while MC simulations provide precise dose estimation, their time-consuming nature often leads to reliance on faster deterministic algorithms, which may sacrifice accuracy, especially near organ boundaries.

Recent interest in high-dose-rate therapies, particularly electron FLASH therapy, offers ultra-fast radiation delivery with many potential benefits. FLASH-RT, characterized by ultra-high dose rates (>40 Gy/s) and approximately 400 times faster than traditional radiotherapy, exploits the "FLASH effect", sparing healthy tissues while effectively targeting tumors [2, 3]. Very High Energy Electrons (VHEE) are currently considered one of the best candidates to exploit the FLASH effect on deep seated tumors, also benefiting from the greater simplicity in changing beam's orientation and energy [4, 5].

At present, building a treatment plan optimization strategy for FLASH electrons which simultaneously considers beams orientation, energy and fluence is still a challenge, also because of the lack of quick algorithms for dose estimation.

In this context, DL can emerge as a promising tool to overcome present limitations. DL models can learn to emulate MC simulations, facilitating the fast generation of precise dose distribution maps. Their versatility, makes DL models applicable to various RT treatments, possibly leading to a broad impact on the RT treatment field.

In the existing literature which explores DL applications in dose deposition estimation and treatment planning, there is a focus on replicating entire treatment plans for clinically established treatments [6, 7]. Only a limited number of studies focuses on the bare prediction of dose maps [8, 9]. We propose a DL model which aims to replicate MC-simulated dose distributions for VHEE beams, utilizing a voxelated cylindrical volume strategy for efficient and accurate predictions. Employing Graph Neural Networks, specifically a U-Net-like model based on Graph Convolutional Networks, our model provides an ultra-fast (~ 0.01 s) estimate of the deposited energy distribution within a cylindrical volume around the beam.

2 Materials and Methods

We trained a DL model to emulate the energy deposition distribution of an electron beam on a human head. The DL model takes the beam's energy and organ densities in a cylindrical volume around the beam as input, providing the deposited dose distribution in the same cylindrical volume as output.

We utilized a Geant4 simulation with a CT scan of a patient's head coming from the HaN-Seg dataset [10]. The CT scan, comprising 116 slices with a thickness of 2 mm and 512×512 pixels per slice, served as the detector geometry. Materials were assigned based on Hounsfield Unit (HU) values, and nominal densities were assigned using the standard calibration curve from [11].

VHEE beams were simulated as Gaussian pencil beams with a full width at half maximum (FWHM) of 0.5 cm. A cylindrical scorer was used to collect dose deposition data, and a dataset of 10,000 low statistics examples (10,000 primaries) was generated for DL model training, varying beam energies between 70 and 130 MeV. A small set of 10 examples with

high statistics (1 million primaries) was also generated for the final evaluation of the DL model. The cylindrical scorers are 40 *cm* long, with a radius of 8 *cm*. Each cylindrical voxel is 4 *mm* long along the z and r axis. Inner voxels have an angular length of ~ 1 *mm*, while the outer ones have an angular length of ~ 2 *cm*. In this way we can provide a higher resolution near the beamline, where most of the dose is deposited, keeping the number of voxels, and thus the computational requirements, relatively low.

Density values in cylindrical voxels were computed through interpolation using $N = 100$ random 3D points inside each voxel. This process ensured reliable density estimates for input to the DL model. The cylindrical voxels were considered as nodes of cylindrical graphs and each node was connected to its two nearest neighbours in each dimension: z, θ, r.

The DL model, to which we will refer as ReNN GU-Net, is based on Graph Neural Networks with a U-Net structure and comprises three modules: encoder, embedder, and decoder. The encoder processes cylindrical graphs containing density information through a series of GraphConv layers and regular pooling operations [12], while the embedder handles the beam energy. The outputs of the encoder and the embedder are summed and pased to the decoder. The decoder reconstructs the dose distribution using GraphConv layers and un-pooling operations with skip connections from the encoder. The model is based on the one already presented in [13], with some improvements.

The main improvement regards increasing the connectivity in the graphs basing on a simple prior related to the physics of the dose deposition: the dose deposited in the "deeper" voxels in the cylinder depends on the dose deposited at smaller depth. Thus we wired additional direct links between each node and its 4 neighbours in the positive z direction of the cylinder. In order to inject other nodes' spatial information in the model, we also concatenated to the input graphs a 3D positional encoding on the cylindrical coordinates. Overall, this modifications led to a $\sim 10\%$ improvement in the model performance on the dose prediction task.

The model was trained with Binary Cross-Entropy loss function.

3 Results and discussion

We divided our 10,000 examples into train, validation, and test sets. The test set, comprising 275 examples, met specific conditions for beam energy, angle θ, and angle ϕ. The remaining data was split into a training set (8,752 examples) and a validation set (973 examples), maintaining a 1:10 ratio. This test set configuration aimed to rigorously assess the ReNN GU-Net model's ability to interpolate between samples and generalize to unseen configurations, considering both patient anatomy and beam energy.

The 10 high-statistics examples for the final evaluation and plots were generated in the parameter range of the test set.

MAE $[\mu Gy]$	(0.26 ± 0.05)
$\delta < 3$	$(99.8 \pm 0.2)\%$
MRE z	$(0.6 \pm 0.3)\%$
MRE r	$(0.4 \pm 0.2)\%$
MRE E	$(0.2 \pm 0.2)\%$

Table 1: We report the results for all the metrics used to compare the prediction of our DL model with the MC simulation. Reported values of mean and standard deviation are computed on the high-statistics evaluation set.

The model was trained with a learning rate of 0.005, a batch dimension of 16, and a learning rate scheduler (ReduceOnPlateau) with a patience of 20 epochs. Training stopped at the convergence of the validation loss after 104 epochs. Training was performed on a Tesla V100 GPU and lasted approximately 4 hours.

3.1 Evaluation Metrics

We employed four metrics to evaluate model performance:

1. Mean Absolute Error (MAE): Voxel-wise mean absolute error between MC and DL dose deposition.
2. δ-index: Introduced by Mentzel [14], it quantifies voxel-wise absolute error differences between MC and DL dose depositions normalized by the maximum deposited dose. We considered the δ-index 3% passing rate, which measures the percentage of voxels with δ-index less than 3%.
3. Mean Relative Error (MRE) on Dose Profiles: Mean Relative Error for integrated dose deposition profiles along the cylindrical axes z and r.
4. Mean Relative Error (MRE) on Total Dose: Mean Relative Error on the full integral of the dose deposition, representing the total deposited dose.

Results for all metrics, on the evaluation high-statistics set, are presented in Table 1. They show a high level of agreement between the MC simulated and DL emulated dose with a δ-index 3% passing rate of 99.8% and a MRE on dose profile around 0.5%.

In Figure 1 we illustrate the comparison between the integrated dose deposition profiles along the z and r axis of the cylinder. In the top plots, we report in blue, the dose profile computed with MC simulation, in orange, the one computed with our DL model and with dashed grey line, the average density in g/cm^3. In the bottom plots, the relative error between the two profiles is reported. As it can be seen, the relative error is always bounded between -2% and 2%.

The similar level of agreement is also found locally. In Figure 2 we show the same plot computed considering only one voxel in the r dimension. the dose profiles are thus computed considering only the inner cylinder with radius

equal to 4 *mm*, where most of the dose is deposited. In this case, the relative error is bounded between -4% and 2%. However, it is possible to note that the relative error increase in the tail of the distribution, where also the MC simulations typically present greater variance, due to the decreased statistics.

Overall, while maintaining a satisfactory precision level, our DL model allows to generate dose distributions in an ultra-fast way. While the high-statistics MC simulations takes around 18 *min* in multi-threaded mode on a 16 cores CPU (HP Z2 Tower G5 Workstation), our DL model can generate the same dose map in around 0.02 *s* on the same device -i.e. on a CPU- leading to a speed up factor of approximately 5×10^4 times. Moreover, making the model run on GPU devices shorten even more -up to 6 orders of magnitude in batch mode- the execution time.

Finally, the greater advantage of a DL dose map generation resides in the fact the dose maps predicted by DL models are, by construction, differentiable with respect to the beam parameters, which are model inputs. This characteristic will allow us to build a treatment planning optimization strategy completely based on exact gradient-based optimization.

4 Conclusion

In conclusion, this study addresses the critical challenge of developing a treatment plan optimization strategy for VHEE beams, specifically in the context of FLASH therapy. The reliance on MC simulations for dose estimation poses time constraints, necessitating the exploration of alternative methods. Leveraging the potential of DL, we introduced the ReNN GU-Net model based on Graph Neural Networks, aiming to efficiently and accurately predict MC-simulated dose distributions for VHEE beams.

Through extensive training and evaluation using a Geant4 simulation with a CT scan of a patient's head, our model demonstrated remarkable performance. The evaluation metrics, including Mean Absolute Error (MAE), δ-index passing rate, and Mean Relative Error (MRE) on dose profiles, revealed a high level of agreement between MC-simulated and DL-emulated dose distributions. The model's ability to generalize to unseen configurations, considering both patient anatomy and beam energy, was rigorously assessed.

Furthermore, our improvements to the model, involving increased connectivity in the graphs and the incorporation of spatial information, resulted in a notable enhancement in performance. The ReNN GU-Net achieved an ultra-fast estimate of the deposited energy distribution within a cylindrical volume around the beam, with an execution time approximately 0.02 seconds on a CPU.

The significance of this DL model extends beyond its speed and accuracy. The differentiability of the predicted dose maps with respect to beam parameters opens up avenues for exact gradient-based optimization in treatment planning.

In essence, the successful development and validation of the ReNN GU-Net model present a promising leap forward in the integration of AI, specifically DL, in the optimization of treatment plans for advanced radiation therapies, showcasing potential transformative impacts on the field of radiotherapy.

References

[1] G. Delaney, S. Jacob, C. Featherstone, et al. "The role of radiotherapy in cancer treatment: estimating optimal utilization from a review of evidence-based clinical guidelines". en. *Cancer* 104.6 (Sept. 2005), pp. 1129–1137.

[2] V. Favaudon, L. Caplier, V. Monceau, et al. "Ultrahigh dose-rate FLASH irradiation increases the differential response between normal and tumor tissue in mice". eng. *Sci. Transl. Med.* 6.245 (July 2014), 245ra93.

[3] M.-C. Vozenin, P. De Fornel, K. Petersson, et al. "The Advantage of FLASH Radiotherapy Confirmed in Mini-pig and Cat-cancer Patients". eng. *Clin. Cancer Res.* 25.1 (Jan. 2019), pp. 35–42.

[4] E. Schüler, K. Eriksson, E. Hynning, et al. "Very high-energy electron (VHEE) beams in radiation therapy; Treatment plan comparison between VHEE, VMAT, and PPBS". eng. *Med. Phys.* 44.6 (June 2017), pp. 2544–2555.

[5] A. Sarti, P. De Maria, G. Battistoni, et al. "Deep Seated Tumour Treatments With Electrons of High Energy Delivered at FLASH Rates: The Example of Prostate Cancer". en. *Front. Oncol.* 11 (Dec. 2021), p. 777852.

[6] V. Kearney, J. W. Chan, T. Wang, et al. "DoseGAN: a generative adversarial network for synthetic dose prediction using attention-gated discrimination and generation". en. *Sci. Rep.* 10.1 (July 2020), p. 11073.

[7] P. J. Jensen, J. Zhang, B. F. Koontz, et al. "A Novel Machine Learning Model for Dose Prediction in Prostate Volumetric Modulated Arc Therapy Using Output Initialization and Optimization Priorities". en. *Front Artif Intell* 4 (Apr. 2021), p. 624038.

[8] X. Zhang, Z. Hu, G. Zhang, et al. "Dose calculation in proton therapy using a discovery cross-domain generative adversarial network (DiscoGAN)". en. *Med. Phys.* 48.5 (May 2021), pp. 2646–2660.

[9] C. Kontaxis, G. H. Bol, J. J. W. Lagendijk, et al. "DeepDose: Towards a fast dose calculation engine for radiation therapy using deep learning". en. *Phys. Med. Biol.* 65.7 (Apr. 2020), p. 075013.

[10] G. Podobnik, P. Strojan, P. Peterlin, et al. "HaN-Seg: The head and neck organ-at-risk CT and MR segmentation dataset". *Medical Physics* 50.3 (2023), pp. 1917–1927. DOI: https://doi.org/10.1002/mp.16197.

[11] W. Schneider, T. Bortfeld, and W. Schlegel. "Correlation between CT numbers and tissue parameters needed for Monte Carlo simulations of clinical dose distributions". *Physics in Medicine Biology* 45.2 (Feb. 2000), p. 459. DOI: 10.1088/0031-9155/45/2/314.

[12] L. Arsini, B. Caccia, A. Ciardiello, et al. "Nearest Neighbours Graph Variational AutoEncoder". en. *Algorithms* 16.3 (Mar. 2023), p. 143.

[13] L. Arsini, B. Caccia, A. Ciardiello, et al. "Fast and precise dose deposition estimation for FLASH electron radiotherapy with Graph Neural Networks". en. *Submitted to Frontiers in Physics* 0.0 (Feb. 2024), p. 0.

[14] F. Mentzel, K. Kröninger, M. Lerch, et al. "Fast and accurate dose predictions for novel radiotherapy treatments in heterogeneous phantoms using conditional 3D-UNet generative adversarial networks". en. *Med. Phys.* 49.5 (2022), pp. 3389–3404.

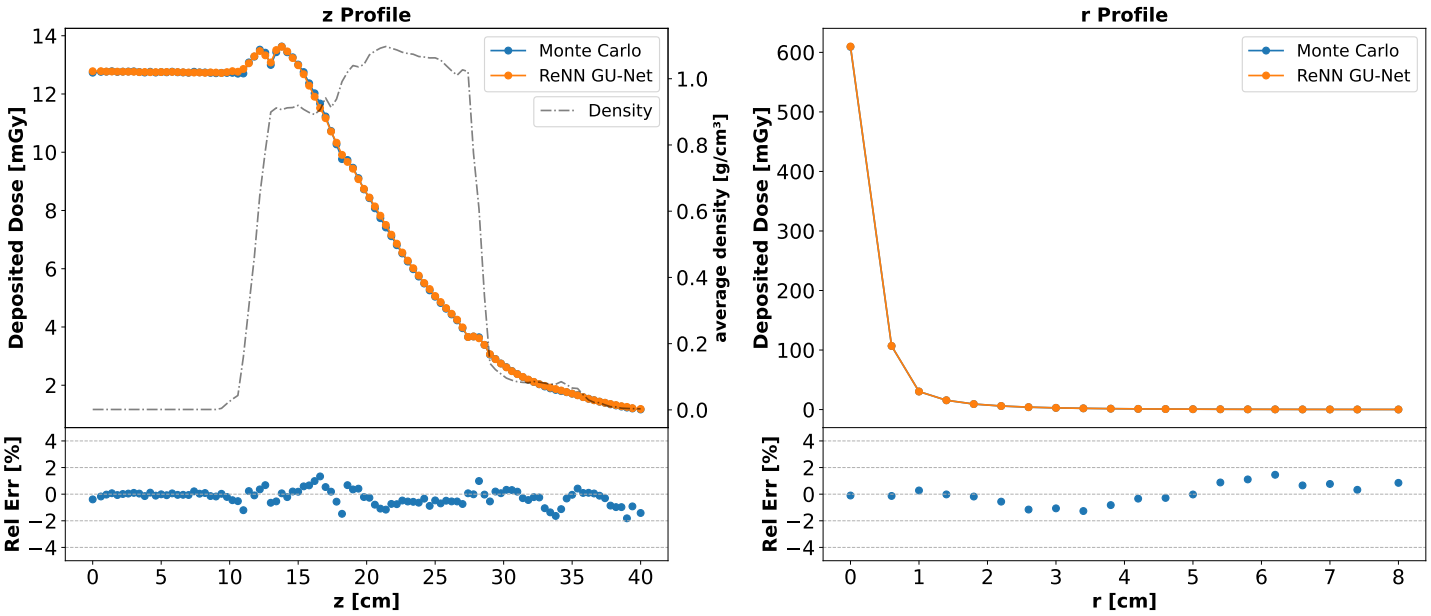


Figure 1: Integrated dose deposition profiles. Top: Comparison between dose deposition integrated profiles along the z and r axis of the cylinder computed with MC simulation (in blue) and our DL model (in orange) for an example drawn form the high-statistics evaluation set. With a grey dashed line, it is reported the average CT scan density computed from the HU using a calibration curve. Bottom: Relative error in percentage between the MC and DL computed dose profiles.

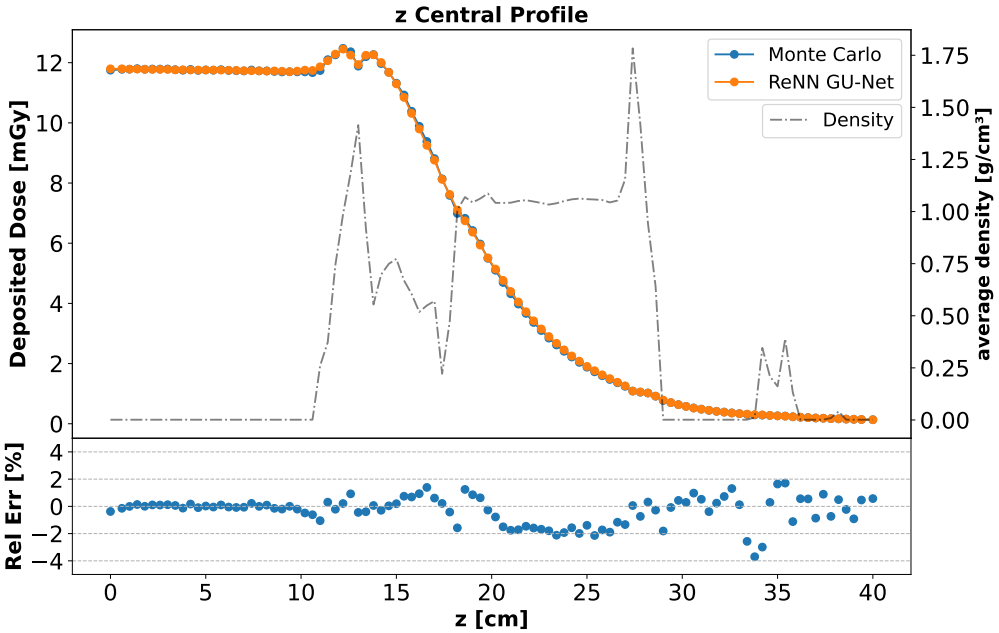


Figure 2: Inner dose depostion profiles. Top: Comparison between dose deposition z profile computed for the core cylinder around the beam (with $r < 4$ *mm*) with MC simulation (in blue) and our DL model (in orange) for an example drawn form the high-statistics evaluation set. With a grey dashed line, it is reported the average CT scan density computed from the HU using a calibration curve. Bottom: Relative error in percentage between the MC and DL computed dose profiles.

Skin dose in breast radiation therapy: Monte Carlo calculations from deformed vector fields (DVF)-driven CT images

N. Arbor[1], L. Bartolucci[2], B. Dai[3], H. Elazhar[2], P. Galmiche[3], D. Jarnet[2], M. De Mathelin[3], P. Meyer[2], H. Seo[3]

[1]Université de Strasbourg, CNRS, IPHC UMR 7178, F-67000 Strasbourg, France
[2]Département de Radiothérapie, ICANS, Strasbourg, France
[3]Laboratoire ICube, UMR7357, Strasbourg, France

Abstract Skin dose in radiation therapy is a key issue for reducing patient side effects, but dose calculations in this high-gradient region is still a challenge today. To support radiation therapists and medical physicists in their decisions, a computational tool has been developed to systematically recalculate the skin dose for each treatment session, taking into account changes in breast geometry. This tool is based on Monte Carlo skin dose calculations from deformed vector fields (DVF)-driven CT images, using non-ionizing 3D-scanner data. The application of deformation fields to the initial patient CT, as well as the validation of Monte Carlo skin dose calculations and comparison with TPS algorithm, are presented. The tool is currently being evaluated using a dataset of 60 patients from the MorphoBreast3D clinical study.

1 Introduction

Breast cancer is the most common cancer among women in France, with around 60,000 new cases and over 10,000 deaths per year. Some 85% of patients are currently treated with radiotherapy [1]. Despite continuous improvements in techniques to optimize dose distribution, i.e. increase the precision of radiation targeting while minimizing exposure to normal tissues, a significant number of patients still suffer from acute toxicity (burns) to the skin and subcutaneous tissues. In order to reduce these side effects, one needs to better estimate the exact dose delivered to the skin during irradiations. However, calculating the dose distribution in this very high-gradient region is complicated by the combination of the lack of precision of treatment planning systems (TPS) at the air-tissue interface [2], the compromise to be made on the calculation grid size in order to remain within acceptable calculation times, and the inevitable changes in breast geometry during and throughout treatment (breathing, breast swelling between fractions) [3].

Radiotherapy centers have developed protocols to limit the possible side-effects of inaccurate surface dose estimation, such as modifying the patient's target or body volume margins [4], or using a bolus during the optimization process to spare dose build-up to the skin. But the main difficulty encountered by medical physicists is the lack of tools available for determining the optimal solution for each patient, especially when considering intensity modulated techniques with regard to changes in breast geometry during treatment.

In this context, the Strasbourg Europe Cancer Institute (ICANS), the ICube laboratory, and the Hubert Curien Interdisciplinary Institute (IPHC) are collaborating on the development of a computational framework dedicated to skin dose calculation in breast radiation therapy. This framework combines a Monte Carlo dose calculation tool and a breast deformation algorithm by using 3D surface data recorded during treatment sessions.

This framework allows:
1) to evaluate the necessity for recalculating the treatment plan, based on the production of deformed CT images that reproduce the inter-fraction evolution of breast geometry;
2) to exploit the precision of the Monte Carlo calculation to estimate the dose delivered in the 1st millimeter of skin. This complements and evaluates the calculations of the TPS algorithms, thereby improving the understanding of the induction of side effects introduced by the breast deformation.

After introducing the general framework architecture, we present experimental measurements and Monte Carlo calculations of skin dose, the method to compute CT images of deformed breast by using 3D surface measurements, as well as the results we obtained on the first patients as part of a clinical study.

2 Materials and Methods

The skin is a complex organ, characterized by a layered structure with different cell types. The thickness of the epidermis varies between 0.05 and 1.5 mm depending on the anatomical region. The International Commission on Radiological Protection (ICRP) and the International Commission on Radiological Units and Measurements (ICRU) recommend a thickness of 0.07 mm for the practical study of the dose to the skin.

70 μm remains a precision that is difficult to access with current clinical dose calculation algorithms, thus motivating the use of Monte Carlo calculation to address the problem of skin dose in breast radiotherapy. Our dose calculation tool is based on the GATE Monte Carlo software [5], and has been previously validated on a 6 MV TrueBeam STX accelerator [6]. This tool uses the phase spaces provided by the manufacturer and the information available in the DICOM-RT files (image, plan, structure, dose) to calculate the dose delivered to the patient. A digital phantom of the patient is created from the scanner images and the HU-materials conversion method proposed by Schneider et al. [7].

The dose to the skin is due to low energy photons scattered from the accelerator head, photons backscattered into the medium, and contamination electrons produced in the accelerator head and the head-patient interface (air, accessories). Significant validation work was therefore carried out to ensure the good accuracy of the Monte Carlo dose calculation in the first millimeters of tissue. Contamination electrons, which can contribute up to nearly 50% of the dose delivered to the skin depending on the configurations, have been particularly studied from Varian phase spaces.

A clinical study carried out at ICANS in 2019-2020 enabled the acquisition of 60 patient datasets (*MorphoBreast3D*). These data contain CT images of the patients used for planning (with contouring of the organs including the skin), to dose maps calculated by the TPS, and a set of 3D surfaces meshes of the patient bust acquired at different treatment sessions using Artec Eva 3D hand-held scanner (Figure 1).



Figure 1. Artec Eva 3D scanner

The objective is to use the 3D surfaces recorded during a session in order to deform the initial CT image of the patient and to obtain a new CT image representing the exact breast geometry during the session. A per-vertex correspondence has been found between a 3D surface pair, one acquired from the CT scan (M_{init}), and another during a session (M_s). Then a deformation vector field is calculated by aligning the two surfaces and connecting the vertices in correspondence. The displacement vector field is then transferred to the CT image by rigidly aligning the surface M_{init} and the CT image. We finally obtain the deformed CT image via the use of radial basis functions (RBF), whose point specifications are defined by the displacement vector. The latter step uses voxelization and interpolation techniques to ensure the overall consistency of the distorted CT image which will serve as input for dose calculations.

CT images containing deformed breast geometry serve as input to both TPS global dose calculations and GATE Monte Carlo skin dose calculations. The use of TPS to recalculate the dose map on the exact geometry of the breast during the session should enable medical physicists to estimate whether a replanning of the treatment is necessary, by examining deviations from the initial dose plan. The calculation of dose to the skin via our Monte Carlo tool is used to provide additional information to the TPS calculations in the first millimeter of tissue, and to precisely study the impact of deformations on the distribution of dose to the skin, such as under and over-estimation or hot spots.

3 Results

Figure 2 shows an example of a deformation vector field calculated on a patient's 3D surface mesh M_{init}.

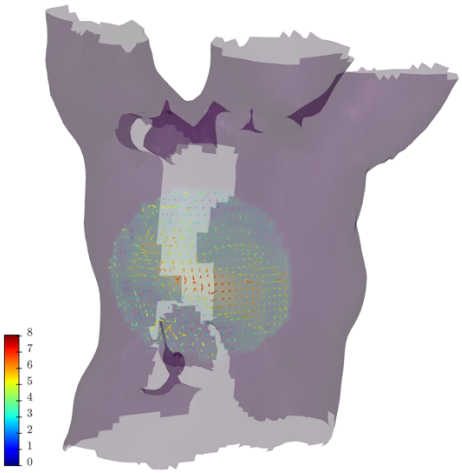


Figure 2. Deformation vector field

From this deformation field, our algorithm produces a CT image representing the new geometry of the deformed breast for a given treatment session (Figure 3).

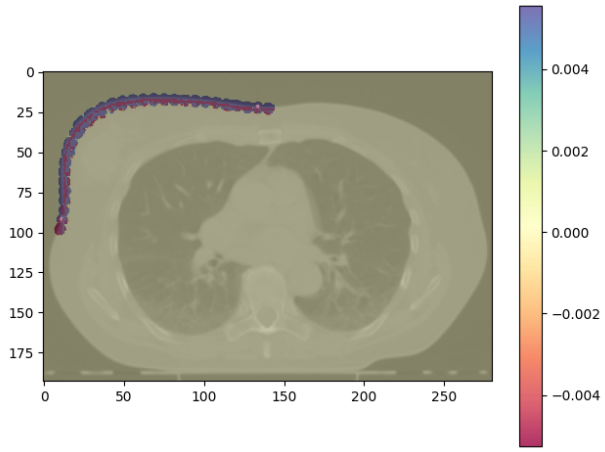


Figure 3. Breast deformation (colored voxels (a.u.)) from pre-treatment patient CT

Before applying Monte Carlo calculations to initial and deformed CT images, experimental validations of skin dose have been carried out on PMMA phantoms for various treatment plans. Monte Carlo dose calculations in the first millimeter have been compared to EBT3 (Ashland, Bridgewater, NJ, USA) Gafchromic films [8], demonstrating the good calculation performance of GATE Monte Carlo algorithm at this millimeter scale (Figure 4).

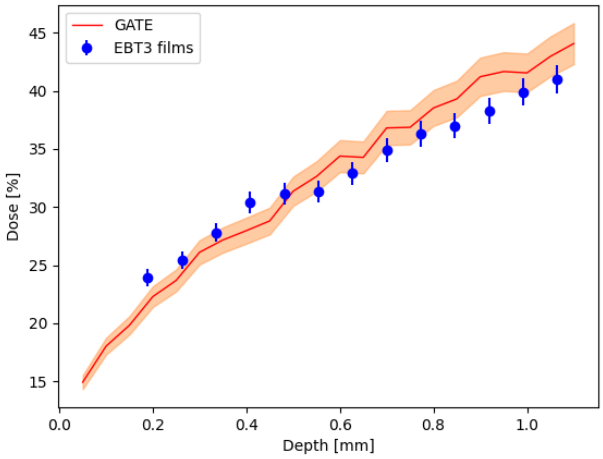


Figure 4. Depth dose comparison between GATE (Monte Carlo) and EBT3 film measurements (PMMA phantom)

We have currently tested our method on two patients of the clinical trial. For these patients, a mean breast deformation field length of about 5 mm has been obtained. For each patient, dose maps were calculated by using the Acuros TPS and GATE MC tool for five deformed CT images (D+5, D+10, D+15, D+20, and D+25), and subsequently compared to the dose map from the original treatment plan.

4 Discussion

These first tests of a skin dose calculation framework, carried out as part of this work, are very encouraging. The study will continue its systematic application to all 60 patients who participated in the clinical study. The ultimate goal is then to exploit the developed framework in several complementary applications:
a) verification of the dose calculation after each session, taking into account the exact geometry of the breast, offering both to the radiation therapists and medical physicists a decision support tool to determine the necessity of treatment replanning;
b) set up research work on the precision of surface dose calculations from TPS algorithms (contour, thresholding) by a systematic comparison with MC calculations
c) provide data for a new clinical study on understanding the side effects of breast radiotherapy depending on the type of in-therapy deformation (shrink, inflation, folds)

Other tools exist today that allow recalculation of the treatment plan based on breast deformations at each session. These tools are mainly based on the use of CBCT images taken at the beginning of the session for the repositioning of the patient. One of the specificities of the proposed framework compared to these tools consists of using 3D surfaces obtained from an optical scanner. Its application does not require an additional X-ray image, relying solely on the initial CT, thus avoiding extra imaging doses for the patient and adhering to the ALARA principle for patient radiation protection. Another advantage of the algorithm developed in this work is its applicability to exploit images provided by the SGRT (Surface Guided Radiation Therapy) systems, which are increasingly present in treatment centers.

5 Conclusion

Skin dose calculation in breast radiation therapy is a key issue for reducing patient side effects. A framework has been developed to enable systematic recalculation of the skin dose at each session, taking into account breast deformations from 3D-scanner images. This framework enables Monte Carlo skin dose calculations based on deformed vector fields (DVF)-driven CT images. The first encouraging results on patients, as well as the increasing availability of SGRT images, point to the future clinical use of this non-ionizing method. This tool will support radiation therapists and medical physicists in their replanning decisions during treatment, and provide useful data for improving treatment precision and better understanding the induction of skin toxicity.

References

[1] Institut National du Cancer, Panorama des cancers en France, Edition 2023

[2] Y. Cao et al., Superficial dose evaluation of four dose calculation algorithms. Radiat Phys Chem 2017 137:23–8.
DOI: 10. 1016/j.radphyschem.2016.02.032

[3] J. Seppälä et al., Breast deformation during the course of radiotherapy: The need for an additional outer margin, Physica Medica 2019 65:1-5.
DOI: 10.1016/j.ejmp.2019.07.021

[4] L. Wang, A. J. Cmelak and G. X. Ding, A simple technique to improve calculated skin dose accuracy in a commercial treatment planning system, J Appl Clin Med Phys 2018 19:2:191–197.
DOI: 10.1002/acm2.12275

[5] D. Sarrut et al., A review of the use and potential of the GATE Monte Carlo simulation code for radiation therapy and dosimetry applications. Medical Physics 2014 41(6):064301.
DOI: 10.1118/1.4871617

[6] N. Arbor, J. Gasteuil, C. Noblet, M. Moreau, P. Meyer, A GATE/Geant4 Monte Carlo toolkit for surface dose calculation in VMAT breast cancer radiotherapy, Physica Medica 2019 61.
DOI: 10.1016/j.ejmp.2019.04.012

[7] W. Schneider W, T. Bortfeld T, W. Schlegel, Correlation between CT numbers and tissue parameters needed for Monte Carlo simulations of clinical dose distributions, Phys. Med. Biol. 2000 45(2):459.
DOI: 10.1088/0031-9155/45/2/314

[8] S. Devic et al., Accurate Skin Dose Measurements Using Radiochromic Film in Clinical Applications: Skin Dose Measurements Using Radiochromic Film, Medical Physics 2006 33(4): 1116-24.
DOI: 10.1118/1.2179169

Peeking under the Hood: Addressing Biomarker-driven Uncertainty in Machine Learning-based Radiotherapy Personalization

Ali Ajdari[1], Zhongxing Liao[2], and Thomas Bortfeld[1]

[1]Department of Radiation Oncology, Division of Radiation BioPhysics, Massachusetts General Hospital & Harvard Medical School, Boston, USA
[2]Department of Thoracic Radiation Oncology, Division of Radiation Oncology, University of Texas' MD Anderson Cancer Center, Houston, USA

Abstract Radiotherapy (RT) has long been designed using a set of population-wide dose prescriptions that do not fully reflect the individual patients' biological makeup. Although powerful machine learning (ML) models are available for predicting patient-specific outcomes of RT, these models remain disconnected from the realm of dose personalization. Here, we propose a novel strategy for bridging the gap between ML-based outcome modeling and dose-based RT treatment planning. The novelty of our approach lies in directly addressing data uncertainty (from predictive biomarkers) during the outcome modeling and optimization. On the modeling side, we train *adversarially robust* neural networks that are immune to reasonable perturbation of the underlying data. On the computational side, we propose efficient algorithms for embedding the resulting robust NNs directly within conventional treatment planning for personalized dose adjustment. We test our methodology on a cohort of non-small cell lung cancer patients, with the endpoint of minimizing radiation-induced cardiac toxicity.

1 Introduction

Radiotherapy (RT) treatment plan optimization (TPO) has long adhered to population-wide guidelines for setting tumor dose prescriptions and "safe" tolerances for normal tissues. In ignoring considerable individual differences in tumor and normal tissue radiosensitivity, the resulting treatment plans can lead to insufficient tumor dose or excessive normal tissue damage in patients who are farther from the average population. Variability in patient response highlights the need for outcome-centric approaches that reflect patients' unique characteristics and therapeutic responses. In recent years, power machine learning (ML) models have been proposed for predicting patient-specific RT outcomes. These models are capable of synthesizing conventional dosimetric factors with more patient-specific biological information derived from patients' medical history, imaging, and genetic or molecular assays and predict individualized outcomes [1]. Despite this promise, current ML-based outcome models remain largely disconnected from clinical RT planning. Outside of simple patient stratification strategies, there is no clear path toward translating the predictions of these models into "clinically actionable decisions", i.e., to truly personalize the dose for patients at higher risk of toxicity and/or disease progression. We have recently introduced a novel methodology for the direct embedding of ML-based outcome models within conventional (dose-based) RT TPO, enabling ML-driven RT planning [2]. Guided by powerful ML outcome models such as Random Forests, Gradient Boosting Machines, and Neural Networks (NN), the methodology is capable of automatically steering the RT plan of high-risk patients towards desired (predicted) clinical outcome, thereby bridging the gap between ML-based outcome modeling and RT plan optimization.

An important issue in outcome-driven TPO is the *uncertainties* associated with the outcome models' predictions, specifically those stemming from data and the model uncertainties. The former (hereafter called *biomarker uncertainty*) is especially relevant in RT outcome modeling, as the underlying data, particularly those derived from imaging, genetic, or other biological assays, are subject to measurement uncertainties and thus have the potential to severely hamper the performance of outcome-driven RT planning.

The goal of this study is to address biomarker uncertainties in ML-driven RT planning. Extending our previous work, here we will (i) build on the emerging field of Adversarial Robustness [3, 4] to train RT outcome models whose predictions are robust against (reasonable) perturbations in the underlying biomarker data, and (ii) offer mathematical strategies for embedding these models and deriving "robust" (with respect to biomarker uncertainties) ML-driven plans. We will test our approach on a retrospective cohort of non-small cell lung cancer (NSCLC) patients who had undergone chemoradiation, focusing on minimizing the risk of radiation-induced cardiotoxicity (RICT) for high-risk NSCLC patients.

2 Methods

Our methodology consists of three interconnected steps: first, we will introduce the concept of adversarial robustness (AR) to design NN outcome models that are less sensitive to reasonable perturbations in the underlying training data (biomarkers); we then introduce an efficient methodology for embedding the resulting NN models as *patient-specific* planning constraints within the conventional TPOs; finally, we will derive fully outcome-driven plans that are robust *w.r.t.* biomarker uncertainties.

2.1 Defining biomarker uncertainty

Throughout this work, we will define "biomarker uncertainty" as *how accurately the biomarker reflects the 'true' (unknown) state of the variable being measured.* This is tied to the inherent accuracy of the measuring instrument. In the case of

imaging uncertainty, this could be interpreted as the difference between the imaging signals and the true underlying anatomy (e.g., tumorous vs. normal cells in pathology slices), whereas for liquid biopsies, it could be the difference between the biomarker expression (e.g., circulating tumor DNA [5]) in the sample vs. in the body.

2.2 ML-driven treatment plan optimization under biomarker uncertainty

The general ML-based outcome-driven TPO under biomarker uncertainty can be expressed as:

$$\max_{\mathbf{d},\mathbf{x}\geq 0} \Gamma(\mathbf{d},\tilde{\phi}) \tag{1}$$

$$\text{subject to: } A\mathbf{x} = \mathbf{d}, , \tag{2}$$

$$\sup \Psi(\mathbf{d},\tilde{\phi}) \in \mathscr{H},\ \tilde{\phi} \in \Phi_P. \tag{3}$$

where $\Gamma(.)$ and $\Psi(.)$ are the (ML-based) outcome models predicting, respectively, the tumor control and normal tissue toxicity probability. The vectors $\mathbf{d}$ and $\tilde{\phi}$ denote the vectors of physical dose and patient-specific (biological) characteristics (biomarkers), and $\sup(.)$ denotes the supremum of a set. Constraint (3) ensures that the predicted toxicity profile of the patient belongs to a set $\mathscr{H}$ of 'safe' states (e.g., no severe toxicity), as long as the biomarker measurement $\tilde{\phi}$ belongs to a known confidence interval Φ_P.

The above formulation can incorporate instances where the mathematical expressions of $\Gamma(.)$ and $\Psi(.)$ are unknown, such as in the case of complex ML models, in which case their functional forms have to be directly learned from data. Without loss of generality, we will assume that only the normal tissue outcome model $\Psi(.)$ has to be learned from data and replace $\Gamma(.)$ with a known surrogate such as the conventional dose fidelity term, i.e., $\max \Gamma(\mathbf{d},\phi) \approx \min \|\mathbf{d}-\mathbf{d}_{\text{ref}}\|$, where $\|.\|$ is an arbitrary norm that penalizes the deviation from a set of given dose prescription $\mathbf{d}_{\text{ref}}$.

In the following sections, we will first provide a way for obtaining outcome models $\Psi(.)$ which are less sensitive to the perturbation in $\tilde{\Phi}$, focusing on the case of neural networks, and follow that with the method for embedding the learned model within RT TPO.

2.3 Designing robust neural networks using Adversarial Robustness

For a given set of input-label pairs $\mathbf{x} \in \mathbb{R}^d$ and $y \in \mathbb{R}$ and a NN classifier $f(\mathbf{x},\theta)$, the conventional training algorithms seek to find a set of parameters $\theta \in \mathbb{R}^p$ that minimize a given loss function $\mathscr{L}(f(\mathbf{x},\theta),y)$. Traditionally, such loss functions do not account for the cases where the input $\mathbf{x}$ could be contaminated or noisy, therefore, oftentimes, the trained networks are not robust against even slight perturbations of x. This can pose significant challenges in the noisy context of medical data, where various sources of measurement uncertainty (arising due to image registration, equipment inaccuracy, for instance) exist. Therefore, it is desirable to *immunize* the model against *adversarial* perturbations of the input x.

We define such perturbations as $\text{Adv}(x) = \{x' : \|x'-x\|_\infty \leq \delta\}$, where $\|.\|_\infty$ is the l_∞ norm. An *adversarially robust* (AR) model can then be trained by solving the following optimization problem:

$$\min_\theta \left\{ \max_\delta \mathscr{L}(f(\mathbf{x},\theta),y) : x \in \text{Adv}(x) \right\}. \tag{4}$$

While the inner maximization aims to find an adversarial version of a given data point x to incur a high loss, the outer minimization seeks to minimize the resulting "adversarial loss" by adjusting the model parameter θ [3, 4]. Note that this is an instance of the *min-max* robust optimization (RO) [6]. Unfortunately, due to the complexity (nonconvexity) of the resulting loss function, conventional RO algorithms either do not apply or are prohibitively expensive to run. We therefore attempt to solve the inner maximization problem (the "adversarial attack" problem) by the Fast Gradient Sign Method (FGSM) heuristic [3]. At each iteration t of the outer minimization problem, FGSM produces the most adversarial example by linearizing the loss function around the current value of θ^t, that is, a linear perturbation of x is given by $x' = x + \delta \mathbf{sgn}(\nabla_x \mathscr{L}(\mathbf{x},y,\theta^t))$, where $\mathbf{sgn}$ is the sign function. This approach essentially expands the set of training data points by the adversarial examples and "prepares" the model for encountering such perturbation.

2.4 Embedding NN models within RT treatment plan optimization

Once the network parameters θ are learned using the AR training algorithm, the resulting model could be embedded within the ML-TPO problem. Embedding a trained NN with the weight and bias parameters $\theta^* = (\mathbf{W}^*, \mathbf{b}^*)$ and a rectified linear unit (ReLu) activation function ($\text{ReLu}(\mathbf{x}) = \max\{\mathbf{0},\mathbf{x}\}$) can be achieved by the following set of linear mixed-integer constraints per neurons [2]:

$$v \geq x, \tag{5}$$

$$v \leq x - M_L(1-z), \tag{6}$$

$$v \leq M_U z, \tag{7}$$

$$v \geq 0. \tag{8}$$

where $v \in \mathbb{R}$ and $z \in \{0,1\}$ are the auxiliary variables used for linearizing the ReLu output and M_L and M_U are the lower and upper bound on all possible values of the decision vector $\mathbf{x} = (\mathbf{d},\phi)$. The constraints for embedding NNs can be recursively generated, starting from the innermost input layer, thus allowing the embedding of trained networks with an arbitrary number of hidden layers. The use of mixed-integer constraints imposes significant algorithmic and computational challenges due to the imposed nonconvexity. We use a simple column generation approach [7] for handling the size of the integer variables.

2.5 NSCLC dataset

To test the performance of our approach, we test it, retrospectively, on an NSCLC cohort [8] with the endpoint of limiting cardiotoxicity. The dataset contains 154 patients (photon and proton) NSCLC patients and includes the RT dose distribution, clinicopathological information (e.g., age, sex, pre-existing heart disease), as well as pre- and mid-treatment measurements of high-sensitivity cardiac troponin T (hs-cTnT), an important cardiac-specific biomarker secreted by the cardiac muscle upon injury. Prior works [9] have associated baseline hs-cTnT level with a higher risk of RICT (hazard ratio = 4.06, p = 0.005). We, therefore, treat *baseline* measurement of hs-cTnT as our *uncertainty-prone* biomarker ϕ. We use this dataset both to train our NN model for predicting the risk of RICT grade ≥ 2 (the endpoint) and to test the result of ML-TPO dose adjustment after embedding the predictive model.

3 Results

All experiments were coded in MATLAB v2022a and run on a PC with 3.70GHz CPU and 32GB of RAM. We used Gurobi for solving the optimization problem and matRad [10] for dose calculation/ visualization.

3.1 Characterizing hs-cTnT measurement uncertainty

In the absence of ground truth (the true state of the hs-cTnT expression in the body), to characterize the amount of expected uncertainty in the hs-cTnT measurement, one needs repeated measurements taken at the same time point. Unfortunately, such data did not exist in our dataset. Therefore, we used a surrogate by taking the variation across different time-points during the RT for patients who (i) had received limited heart doses (MHD<=5Gy) and (ii) did not have prior heart disease. The assumption is that the hs-cTnT expressions for these patients are less (negligibly) affected by the treatment, and any changes in its repeated measurements are likely due to other (non-treatment) factors (which we use as a surrogate for the measurement uncertainty).
Our analyses revealed that the average hs-cTnT variation across the eligible patients (measurement uncertainty) was $30.7\% \pm 37\%$ (standard deviation).

3.2 Robust NN training

We used a compact (shallow) neural network with one fully connected layer with 50 neurons and ReLu activation function as our benchmark NN model. Input variables to the model included mean heart dose, mean lung dose, baseline hs-cTnT, pre-existing heart disease (Yes/No), along with standard clinicopathological factors (e.g., age, sex, performance status, tumor size, etc.).
We then ran the AR NN training with three different levels of uncertainty: $\delta = 10\%, 20\%$, and 30%. Table 1 shows the results. The reported accuracy (in %) is normalized to the highest accuracy achieved under the conventional NN training, corresponding to the 5-fold cross-validation area under the curve of 62.5%. The column 'Adversarial attack' reports the misclassification rates caused by perturbing the biomarker measurement by the amount of δ. All results were averaged after five training runs. Despite the increase in the training time and the lower accuracy under the perfect condition compared to the conventional training, as expected, the AR-trained models show significantly higher stability under slight (and expected) perturbation in the biomarker measurement.

	Time	Accuracy	Adversarial attack (δ)		
			0.1	0.2	0.3
Conventional NN	23 s	100%	61%	53%	45%
AR NN - $\delta = 10\%$	587 s	95%	89%	78%	67%
AR NN - $\delta = 20\%$	605 s	91%	87%	85%	68%
AR NN - $\delta = 30\%$	497 s	87%	81%	83%	76%

Table 1: The results of the Adversarial Training for constructing the NN model. The column 'Time' refers to the training time (in seconds) under each training strategy. AR=adversarially robust, NN=neural networks.

3.3 Embedding NN models within conventional TPO

We ran the NN-embedded TPO algorithm on 5 NSCLC patients who were classified as high-risk according to the NN model (based on the clinical dose distribution). The algorithm automatically adjusted the dose distribution for these patients to ensure a favorable risk profile (i.e., no predicted RICT). To ensure clinical deliverability, in addition to the NN-learned constraint, we enforced conventional planning criteria [11]. We chose the AR-trained model with 10% uncertainty consideration, as it delivered the best combination of accuracy (under normal conditions) and stability (under perturbation). Overall, there was a 1.78Gy reduction (range=0.54-3.42Gy) in MHD as the result of the NN-guided dose adjustment. For two patients, the algorithm failed to reclassify the patient into the low-risk group, as the other (non-controllable) risk factors for these patients had an outsize effect on the predicted risk (both had pre-existing heart disease with high baseline hs-cTnT level). Figure 1 shows the result for one of the patients.

4 Discussions

In this work, we have built upon our recent effort to bridge the gap between ML-based outcome modeling and dose-driven radiotherapy planning. Specifically, we have attempted to address an important bottleneck in many outcome-driven strategies, namely, the impact of (biomarker) data uncertainty on the resulting outcome predictions. We have employed an emerging technique in the field of machine learning for

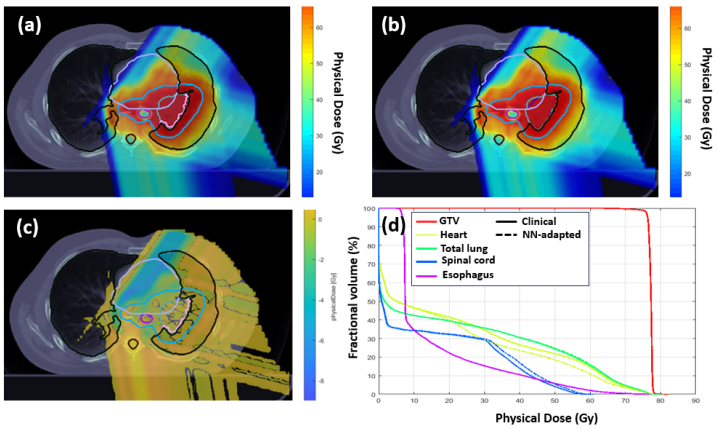


Figure 1: The results of ML-guided adaptation for an exemplary NSCLC case, to minimize the risk of cardiac toxicity. (a)-(c) Original, NN-guided, and difference dose distribution; (d) DVH plots for the original (solid) and NN-guided (dashed) plans.

designing outcome models (neural networks) that are reasonably robust against normal perturbations (adversarial attacks, as they are known in the ML community) in the underlying data. By directly accounting for these perturbations in the loss function during the training phase, the resulting models show superior (up to 30%) stability (i.e., lower misclassification rates) compared to the normally trained model.

We further embedded these robust models within the conventional dose-based treatment planning for a fully outcome-driven RT planning. The results show that in 3/5 cases, it was possible to safely *de-escalate* the heart dose to avoid excessive cardiac damage, without a significant compromise in tumor coverage (Figure 1). Although this improvement naturally comes at the expense of extra dose to other organs (higher mean dose to the spinal cord, SC, in Figure 1), these did not constitute a constraint violation (max-dose to the SC stays below 54Gy).

A limitation of our approach was the relatively small sample size used in the model training. With the sample size of ~ 150, the performance of the AR training algorithm (Table 1) should be taken with a grain of salt (the relatively low AUC of 0.63 is likely the result of the insufficient data points). Moreover, the introduction of integer variables for embedding large NN models causes computational issues during the optimization, an issue that we had previously reported [2]. Although we tried to counter the added complexity using column generation, more work has to be done before the resulting ML-embedded models can be safely deployed in clinics and used to solve large, complex clinical cases.

Nevertheless, the works presented here should be interpreted as a proof-of-concept and not a full-on clinical demonstration. By harnessing the power of ML and mathematical optimization, the algorithms and methodologies developed here will lay the foundation for facilitating the transition towards a fully outcome-driven RT treatment planning and provide the necessary tools for a future clinical trial to test these concepts.

5 Conclusions

A novel approach is proposed for personalized radiotherapy by (i) combining machine learning (ML)-based outcome modeling and dose-based treatment plan optimization, and (ii) addressing an important bottleneck in outcome-driven RT, namely, the data-driven uncertainties caused by biomarker measurement errors. The new uncertainty-aware ML models exhibit superior stability in the face of biomarker uncertainty, and the resulting ML-driven treatment plans carry the potential for improving the therapeutic outcome of patients through selective dose personalization.

References

[1] L. J. Isaksson, M. Pepa, M. Zaffaroni, et al. “Machine Learning-Based Models for Prediction of Toxicity Outcomes in Radiotherapy”. *Frontiers in Oncology* 10 (June 2020), p. 790. DOI: 10.3389/FONC.2020.00790/FULL.

[2] D. Maragno, G. Buti, I. Birbil, et al. “Embedding Machine Learning Based Toxicity Models within Radiotherapy Treatment Plan Optimization”. *Physics in Medicine Biology* (under revision) (2024).

[3] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, et al. “Generative adversarial networks”. *Commun. ACM* 63.11 (2020), pp. 139–144. DOI: 10.1145/3422622.

[4] A. Madry, A. Makelov, L. Schmidt, et al. “Towards Deep Learning Models Resistant to Adversarial Attacks”. *International Conference on Learning Representations*. 2018.

[5] A. Ono, A. Fujimoto, Y. Yamamoto, et al. “Circulating Tumor DNA Analysis for Liver Cancers and Its Usefulness as a Liquid Biopsy”. *CMGH* 1 (5 Sept. 2015), pp. 516–534. DOI: 10.1016/j.jcmgh.2015.06.009.

[6] A. Ben-Tal, L. E. Ghaoui, and A. Nemirovski. *Robust Optimization*. Ed. by J. K. L. and Endre Süli Ingrid Daubechies Weinan E. Princeton University Press, 2009, p. 568. DOI: 10.1007/s10957-013-0421-6.

[7] C. Barnhart, E. L. Johnson, G. L. Savelsbergh, et al. “Branch-and-Price: Column Generation for Solving Huge Integer Programs”. *https://doi.org/10.1287/opre.46.3.316* 46 (3 June 1998), pp. 316–329. DOI: 10.1287/OPRE.46.3.316.

[8] Z. Liao, J. J. Lee, R. Komaki, et al. “Bayesian adaptive randomization trial of passive scattering proton therapy and intensity-modulated photon radiotherapy for locally advanced non–small-cell lung cancer”. *Journal of Clinical Oncology* 36 (18 2018), pp. 1813–1822. DOI: 10.1200/JCO.2017.74.0720.

[9] T. Xu, Q. H. Meng, S. C. Gilchrist, et al. “Assessment of Prognostic Value of High-Sensitivity Cardiac Troponin T for Early Prediction of Chemoradiation Therapy-Induced Cardiotoxicity in Patients with Non-Small Cell Lung Cancer: A Secondary Analysis of a Prospective Randomized Trial”. *International journal of radiation oncology, biology, physics* 111 (4 Nov. 2021), pp. 907–916. DOI: 10.1016/J.IJROBP.2021.07.035.

[10] H. P. Wieser, E. Cisternas, N. Wahl, et al. “Development of the open-source dose calculation and optimization toolkit matRad”. *Medical physics* 44 (6 June 2017), pp. 2556–2568. DOI: 10.1002/MP.12251.

[11] L. B. Marks, E. D. Yorke, A. Jackson, et al. “Use of Normal Tissue Complication Probability Models in the Clinic”. *International Journal of Radiation Oncology Biology Physics* 76 (3 SUPPL. Mar. 2010), S10. DOI: 10.1016/j.ijrobp.2009.07.1754.

pyanno4rt: a toolkit for machine learning (N)TCP model-based inverse radiotherapy treatment plan optimization

Tim Ortkamp[1,2,3], Patrick Salome[2,4], Oliver Jäkel[2,3,5,6], Martin Frank[1,3], and Niklas Wahl[2,5]

[1]Karlsruhe Institute of Technology (KIT), Scientific Computing Center, Karlsruhe, Germany
[2]Department of Medical Physics in Radiation Oncology, German Cancer Research Center (DKFZ), Heidelberg, Germany
[3]Helmholtz Information and Data Science School for Health (HIDSS4Health), Karlsruhe/Heidelberg, Germany
[4]Clinical Cooperation Unit Radiation Oncology, German Cancer Research Center (DKFZ), Heidelberg, Germany
[5]National Center for Radiation Research in Oncology (NCRO), Heidelberg Institute for Radiation Oncology (HIRO), Heidelberg, Germany
[6]Heidelberg Ion Beam Therapy Center (HIT), Heidelberg, Germany

Abstract Machine learning (ML) models on tumor control probability (TCP) or normal tissue complication probability (NTCP) are on the verge of replacing classical radiotherapy treatment outcome models. In spite of this, solving the inverse problem in intensity-modulated radiotherapy (IMRT) treatment planning still relies on mathematically simple, non-model-based surrogate functions with empirical dose prescription and tolerance parameters, rather than directly optimizing (N)TCP based on such outcome models. In this paper, we present the open-source Python package *pyanno4rt*, a toolkit which implements a technical framework for ML outcome model-based IMRT optimization. *pyanno4rt* leverages transformations of the probability prediction functions as (N)TCP objectives and backward propagation of the gradients from these objectives to the fluence vector, both of which could then be embedded in any gradient-based solver. To demonstrate *pyanno4rt*, we first fitted ML outcome models (logistic regression, neural network, support vector machine) internally, using two outcome data sets from a head-and-neck patient cohort. Second, we optimized photon treatment plans on a head-and-neck patient by adding the corresponding (N)TCP objectives to the conventional plan. Our results show that *pyanno4rt* is able to generate ML model-based treatment plans with appealing dose distributions, while allowing for improved outcome predictions.

1 Introduction

Machine learning models are increasingly applied in radiotherapy treatment outcome prediction, and have already been shown to outperform classical outcome models [1]. One of the main advantages is the ability to handle high-dimensional and mixed inputs, which allows them to exploit large amounts of treatment and patient data and therefore provide a highly personalized treatment outcome prediction. This makes machine learning outcome models interesting for application in intensity-modulated radiotherapy (IMRT), where the inverse problem is commonly solved by using surrogate functions with empirical dose prescription and tolerance parameters, e.g. squared deviation from a reference dose, rather than directly generating outcome-oriented treatment plans.
One possible approach for the integration of inverse treatment plan optimization and machine learning outcome models is at the level of the optimization components (objectives and constraints). However, this poses some challenges, e.g. in terms of model stability with changing input over the optimization iterations or the dependence on the efficiency of the forward objective and backward gradient calculations.
In this paper, we present *pyanno4rt* [2], a toolkit which implements a technical framework for addressing these issues. We first introduce the software framework and the underlying mathematical concepts to derive feasible (N)TCP objectives and gradients from the machine learning model prediction functions, then outline the datasets used for model fitting and provide details on treatment planning. To showcase *pyanno4rt*, we integrate the best-fit models into photon treatment plan optimization on some independent head-and-neck patient. An analysis of the results concludes this paper.

2 Materials and Methods

In this chapter, we present the software framework and the underlying mathematical concepts, as well as the datasets used for model fitting and some treatment planning details.

2.1 Software Framework

pyanno4rt is an open-source Python package for conventional and outcome prediction model-based inverse photon and proton treatment plan optimization. It encompasses an extensive and scalable number of optimization components, feature definitions, different solution methods and algorithms, plan evaluation and visualization tools, as well as a graphical user interface in addition to the code-based interface (see Figure 1). The package is under continuous development and currently being actively maintained. All results in this paper were generated from *pyanno4rt*.

2.2 Inverse Treatment Plan Optimization with Machine Learning Model-based Components

We now explain how *pyanno4rt* combines inverse planning and machine learning outcome prediction models.
First, we define the weighted-sum optimization problem for the conventional intensity-modulated treatment plan as

$$\begin{aligned} \min_{\mathbf{w}} \; & f(\mathbf{w}) = \sum_r \omega_r f_r(\mathbf{D}_r \mathbf{w}) \\ \text{s.t.} \; & c_m^l \leq c_m(\mathbf{w}) \leq c_m^u \,. \end{aligned} \tag{1}$$

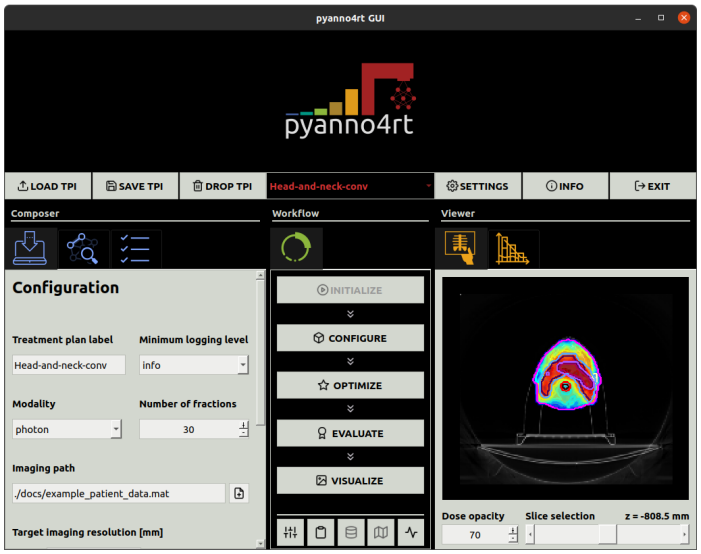


Figure 1: pyanno4rt GUI interface

with the fluence vector $\mathbf{w} \in \mathbb{R}^J_+$, the dose-influence matrix $\mathbf{D}_r \in \mathbb{R}^{I_r \times J}_+$, the objective function f_r and the weight ω_r for the segment r, as well as the (optional) constraints c_m.
The novelty of our approach lies in the interpretation of the machine learning outcome prediction functions as objectives or constraints (where we only show the objective case below). Throughout, we denote by $\mathbf{x} \in \mathbb{R}^p$ the p-dimensional feature vector, and by y the corresponding (binary) class label. Additionally, let n with index $i = 1,\dots,n$ be the number of samples. Under the assumption that $\mathbf{x}$ can be grouped into dosiomic ($\mathbf{x}^{dose}$), radiomic ($\mathbf{x}^{rad}$) and demographic features ($\mathbf{x}^{dem}$), all possibly with a transformation t, the generalized outcome prediction function becomes

$$(N)TCP_{ML} = p_\theta\left(t(\mathbf{x}^{dose})|t(\mathbf{x}^{rad}),t(\mathbf{x}^{dem})\right). \tag{2}$$

In principle, $(N)TCP_{ML}$ could be used directly for optimization. However, it may show unfavorable mathematical properties, e.g. non-convexity or vanishing gradients. We therefore propose an equivalent formulation $f_{ML}(p_\theta)$ as objective and the backward gradient of f_{ML} with respect to $\mathbf{w}$ as

$$\nabla_{\mathbf{w}} f_{ML}(p_\theta) = \nabla_{\mathbf{w}}\mathbf{d}^\top \nabla_{\mathbf{d}}\mathbf{x}^{dose}\nabla_{\mathbf{x}^{dose}} t(\mathbf{x}^{dose})\nabla_{t(\mathbf{x}^{dose})} f_{ML}(p_\theta). \tag{3}$$

pyanno4rt handles both reformulation of $(N)TCP_{ML}$ and differentiation of $f_{ML}(p_\theta)$ as well as the iterative recalculation of $\mathbf{x}^{dose}$ efficiently through JIT compilation and automatic differentiation using *JAX* [3], with the following machine learning model-based objective and gradient functions:

Logistic regression (LR)

Fitting the logistic regression model yields the coefficient vector $\beta \in \mathbb{R}^p$ and the outcome prediction function reads

$$(N)TCP_{LR} = p_\beta(y=1|t(\mathbf{x})) = \sigma\left(\beta^\top t(\mathbf{x})\right). \tag{4}$$

Inversion of the enclosing sigmoid function $\sigma(z) = (1-\exp(-z))^{-1} \in (0,1)$ yields the objective

$$f_{LR}(p_\beta) = \beta^\top t(\mathbf{x}). \tag{5}$$

This function is convex in $t(\mathbf{x})$ with the constant gradient

$$\nabla_{t(\mathbf{x})} f_{LR}(p_\beta) = \beta. \tag{6}$$

Neural networks (NN)

Let $\mathbf{W}_l$ be the weight matrix, b_l the bias, and g_l the activation function in layer $l = 1,\dots,L$, the outcome prediction function can be described by the recursive rule

$$(N)TCP_{NN} = p_{(\mathbf{W},\mathbf{b})}(y=1|t(\mathbf{x})) = \mathbf{h}_L, \tag{7}$$

with the input $\mathbf{h}_0 = t(\mathbf{x})$ in the first layer and

$$\mathbf{h}_l = g_l\left(\mathbf{W}_l\mathbf{h}_{l-1} + b_l\cdot\mathbf{1}\right), \quad l = 1,\dots,L. \tag{8}$$

In the output layer L, we use the sigmoid activation function $g_L(z) = \sigma(z)$ for projecting onto the $(0,1)$ interval.
By the same reasoning as before, the objective after inverting the sigmoid activation function in the output layer reads

$$f_{NN}(p_{(\mathbf{W},\mathbf{b})}) = \mathbf{W}_L\mathbf{h}_{L-1} + b_L\cdot\mathbf{1}. \tag{9}$$

The gradient function can therefore be expressed as

$$\nabla_{t(\mathbf{x})} f_{NN}(p_{(\mathbf{W},\mathbf{b})}) = \nabla_{t(\mathbf{x})}\mathbf{h}_0 \prod_{l=1}^{L-1} \nabla_{\mathbf{h}_{l-1}}\mathbf{h}_l \nabla_{\mathbf{h}_{L-1}} f_{NN}(p_{(\mathbf{W},\mathbf{b})}). \tag{10}$$

Tensorflow's GradientTape API provides the gradient function via automatic input differentiation. We implemented both standard feedforward and input-convex neural networks [4, 5] in *pyanno4rt* to enable the comparison between the less restrictive but non-convex feedforward architecture and the more restrictive but input-convex architecture.

Support vector machines (SVM)

With the learnable dual parameters $\alpha \in \mathbb{R}^n$ and $b \in \mathbb{R}$, the outcome prediction function can be formulated as

$$(N)TCP_{SVM} = p_\alpha(y=1|t(\mathbf{x})) = \sigma_P\left(\sum \alpha_i K(t(\mathbf{x}_i),t(\mathbf{x})) + b\right), \tag{11}$$

where $\sigma_P(z) = (1-\exp(-Az-B))^{-1}$ denotes the Platt scaling function with extra parameters $A,B \in \mathbb{R}$ [6] and K is a kernel. Table 1 shows the kernels implemented in *pyanno4rt*. By inverting the Platt scaling function, we end up with

$$f_{SVM}(p_\alpha) = \sum \alpha_i K(t(\mathbf{x}_i),t(\mathbf{x})) + b \tag{12}$$

as the objective, so that the gradient function is simply obtained by differentiation of the kernel, i.e.,

$$\nabla_{t(\mathbf{x})} f_{SVM}(p_\alpha) = \sum \alpha_i \nabla_{t(\mathbf{x})} K(t(\mathbf{x}_i),t(\mathbf{x})). \tag{13}$$

Kernel	Definition
linear	$K_{lin}(\mathbf{x}_i,\mathbf{x}) = \mathbf{x}^\top \mathbf{x}_i$
rbf	$K_{rbf}(\mathbf{x}_i,\mathbf{x}) = \exp(-\gamma\|\mathbf{x}-\mathbf{x}_i\|^2)$
poly	$K_{poly}(\mathbf{x}_i,\mathbf{x}) = (\gamma\mathbf{x}^\top \mathbf{x}_i + c_0)^p$
sigmoid	$K_{sig}(\mathbf{x}_i,\mathbf{x}) = \tanh(\gamma\mathbf{x}^\top \mathbf{x}_i + c_0)$

Table 1: Kernel functions for the SVM [7].

Kernel	Definition
linear	$\nabla_\mathbf{x} K_{lin}(\mathbf{x}_i,\mathbf{x}) = \mathbf{x}_i$
rbf	$\nabla_\mathbf{x} K_{rbf}(\mathbf{x}_i,\mathbf{x}) = -2\gamma(\mathbf{x}-\mathbf{x}_i)K_{rbf}(\mathbf{x}_i,\mathbf{x})$
poly	$\nabla_\mathbf{x} K_{poly}(\mathbf{x}_i,\mathbf{x}) = \gamma p\mathbf{x}_i(\gamma\mathbf{x}^\top \mathbf{x}_i + c_0)^{p-1}$
sigmoid	$\nabla_\mathbf{x} K_{sig}(\mathbf{x}_i,\mathbf{x}) = \gamma\mathbf{x}_i \cosh(\gamma\mathbf{x}^\top \mathbf{x}_i + c_0)^{-2}$

Table 2: Kernel gradients for the SVM [7].

Table 2 lists the kernel gradients implemented in *pyanno4rt*.

Putting everything together, we can expand the weighted-sum objective function from Equation 1 to yield

$$f(\mathbf{w}, p_\theta) = f(\mathbf{w}) + \sum_s \omega_s f_s(p_{\theta_s}), \tag{14}$$

where we add the machine learning model-based objectives f_s. Note that this approach also allows to (partially) substitute $f(\mathbf{w})$, or integrate outcome model-based constraints c_{ML}.

2.3 Datasets, Modeling and Treatment Planning

Machine learning modeling in this paper is based on two outcome data sets from a retrospective cohort of 153 head-and-neck patients treated with IMRT at Heidelberg University Hospital between the years 2010 and 2015.
The first data set includes features of the parotids to predict the probability of long-term grade 2+ xerostomia, and we refer to [1] for details on data structure and handling. To prevent multicollinearity, we focus on 6 dosiomic (mean, standard deviation, skewness, gradients) and 3 radiomic features (volume, sphericity, eccentricity) for each parotid.
The second data set has 14 dosiomic and radiomic features of the PTV to predict the probability of progression-free survival. From these, we removed 3 dosiomic features (maximum, minimum, quantile at 10% volume), which we found to cause over- or underdosing of the PTV after optimization.
Internal to *pyanno4rt*, the outcome models from Section 2.2 are fitted with *scikit-learn* [8] and *tensorflow* [9], and equipped with individual preprocessing, hyperparameter tuning, evaluation and inspection units, where we currently restrict the preprocessing to feature standardization and the hyperparameter tuning to sequential model-based global optimization using tree-structured Parzen estimators from *hyperopt* [10], modified by a robust "worst-case" variant of stratified k-fold cross-validation to prevent overfitting.

Segment	$\mathbf{d_{mean}}$			$\mathbf{d_{std}}$		
	S1	S2	S3	S1	S2	S3
Par. (L)	0.72	0.66	0.55	0.28	0.23	0.31
Par. (R)	0.70	0.61	0.22	0.28	0.22	0.21
PTV63	2.09	2.09	1.56	0.20	0.25	0.75
PTV70	2.30	2.31	2.30	0.10	0.12	0.10

Table 3: Dose mean and deviation for the scenarios S1–S3.

Segment	$\mathbf{d_{max}}$			$\mathbf{d_{min}}$		
	S1	S2	S3	S1	S2	S3
Par. (L)	1.66	1.52	1.87	0.24	0.33	0.16
Par. (R)	1.73	1.36	1.35	0.15	0.14	0.01
PTV63	2.48	2.78	3.07	0.25	0.24	0.11
PTV70	2.48	2.66	2.51	0.46	0.46	0.47

Table 4: Dose maximum and minimum for the scenarios S1–S3.

Treatment planning relies on an independent head-and-neck patient from the CORT dataset [11]. The segments of interest are left/right parotid, skin, PTV63 and PTV70. To evaluate the impact of machine learning outcome models for optimization, we investigate three optimization scenarios:

(**S1**) Dose objectives only (squared overdosing on left/right parotid and skin, squared deviation on PTV63/PTV70)

(**S2**) S1 with additional machine learning model-based objectives on left/right parotid and PTV63

(**S3**) S2 with only machine learning model-based objectives on left/right parotid and PTV63

Moreover, we opt for photon irradiation with 30 fractions and the L-BFGS-B algorithm from *scipy* [12] for optimization.

3 Results

In this chapter, we report the most important findings on the resulting dose distributions and outcome predictions.

3.1 Dose Distributions

The optimal 2D dose distributions are shown in Figure 2, with the corresponding dosimetrics in Table 3 and 4. S1 achieves good coverage of the PTV63/PTV70 and sparing on the left/right parotid. Adding the best-fit models, the LR models, in S2 (center) implies a minor shift of dose away from the parotids and more pronounced hotspots in the PTVs. The impact of the LR models is most evident for S3, where the heterogeneity of the dose distribution in the PTV63 increases, while the dose distribution in the PTV70 compares well with S1. Due to the overall lower dose in the PTV63, there is also a larger dose reduction in the parotids. We attribute the changes in dose to the most contributing features, i.e., the mean dose in the parotids or the median dose in the PTV63.

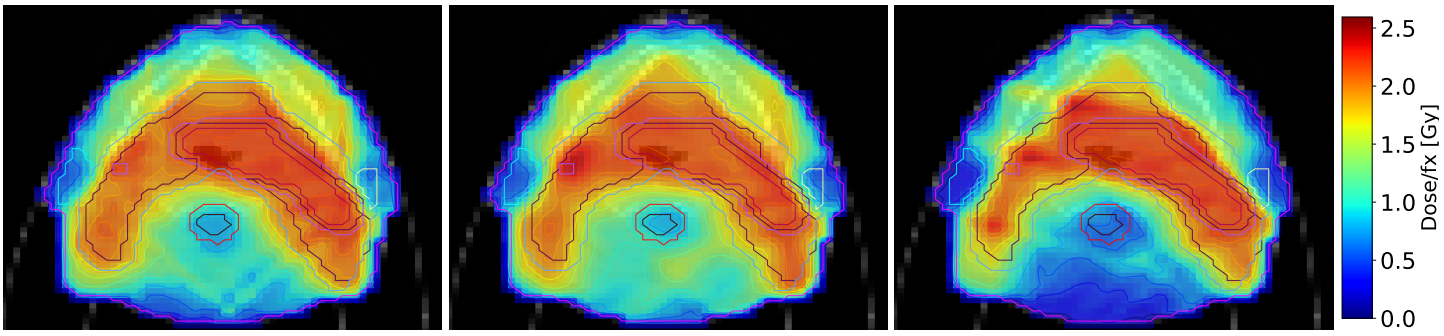


Figure 2: Optimal 2D dose distributions for the scenarios S1–S3 (from left to right).

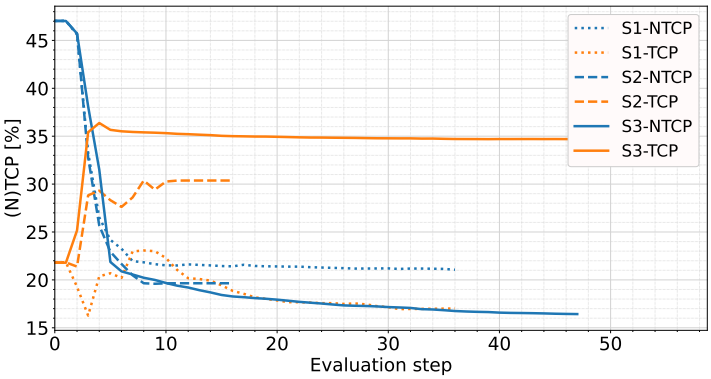


Figure 3: (N)TCP curves for the scenarios S1–S3. Our optimization approach allows for improved outcome predictions, even though the range of the outcome values seems fallacious.

3.2 Outcome Predictions

Figure 3 displays the (N)TCP curves for the three planning scenarios S1–S3. We observe both an increase in TCP (S1: 17%, S2: 30.4%, S3: 34.7%) and a decrease in NTCP (S1: 21%, S2: 19.7%, S3: 16.3%) as the LR models become more influential. This meets our expectation that, on the one hand, the mere inclusion of the (N)TCP objectives in S2 should already improve the outcome predictions, and on the other hand, that the removal of the parallel dose objectives in S3 resolves some trade-offs, allowing the LR outcome models to shape the dose distribution in a less restricted way.

4 Discussion

pyanno4rt shows the ability to optimize machine learning outcome model-based treatment plans with low additional overhead and generate appealing dose distributions with improved outcome predictions. We are also able to (partially) interpret some effects of the fitted outcome models on the optimal dose distribution. However, there is still a lack of sufficient data and high-quality prediction models, which can be observed from the fact that the (N)TCP values from Figure 3 are in a clinically unacceptable range. In addition to that, convexity of the derived (N)TCP objectives only holds under conditions, with no absolute guarantees for a gradient-based algorithm to find a reasonably good optimal point.

5 Conclusion

pyanno4rt is an open-source toolkit for conventional and outcome prediction model-based inverse planning, written purely in Python. It efficiently implements machine learning model-based treatment plan optimization by utilizing the outcome prediction functions. This has a great potential to support the future use of machine learning models in IMRT.

References

[1] H. S. Gabryś, F. Buettner, F. Sterzing, et al. “Design and Selection of Machine Learning Methods Using Radiomics and Dosiomics for Normal Tissue Complication Probability Modeling of Xerostomia”. *Frontiers in Oncology* 8 (2018), p. 35.

[2] T. Ortkamp, O. Jäkel, M. Frank, et al. *pyanno4rt: python-based advanced numerical nonlinear optimization for radiotherapy*. http://github.com/pyanno4rt/pyanno4rt. Version 0.0.1. 2024.

[3] J. Bradbury, R. Frostig, P. Hawkins, et al. *JAX: composable transformations of Python+NumPy programs*. http://github.com/google/jax. Version 0.3.13. 2018.

[4] B. Amos, L. Xu, and J. Z. Kolter. *Input Convex Neural Networks*. 2017.

[5] S. Sivaprasad, A. Singh, N. Manwani, et al. *The Curious Case of Convex Neural Networks*. 2021.

[6] J. Platt. “Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods”. *Adv. Large Margin Classif.* 10 (June 2000).

[7] J. E. Johnson, V. Laparra, A. Pérez-Suay, et al. “Kernel methods and their derivatives: Concept and perspectives for the earth system sciences”. *PLoS ONE* 15.10 (2020).

[8] F. Pedregosa, G. Varoquaux, A. Gramfort, et al. “Scikit-learn: Machine Learning in Python”. *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[9] Martín Abadi, Ashish Agarwal, Paul Barham, et al. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. https://www.tensorflow.org/. Software available from tensorflow.org. 2015.

[10] J. Bergstra, R. Bardenet, Y. Bengio, et al. “Algorithms for hyper-parameter optimization”. *Advances in neural information processing systems* 24 (2011).

[11] D. Craft, M. Bangert, T. Long, et al. “Shared Data for Intensity Modulated Radiation Therapy (IMRT) Optimization Research: The CORT Dataset”. *GigaScience* 3.1 (2014).

[12] P. Virtanen, R. Gommers, T. E. Oliphant, et al. “SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python”. *Nature Methods* 17 (2020), pp. 261–272.

Intelligent Automatic Treatment Planning for Radiation Therapy

Yin Gao[2], YangKyun Park[2], Daniel Yang[2], and Xun Jia[1]

[1]Department of Radiation Oncology and Molecular Radiation Sciences, Johns Hopkins University, Baltimore, MD 21287, USA
[2]Department of Radiation Oncology, University of Texas Southwestern Medical Center, Dallas, TX 75390, USA

Abstract The current workflow of treatment planning in radiation therapy (RT) heavily relies on human planners, leading to variable plan quality and low efficiency due to complex interactions among physicians, planners, and the treatment planning system (TPS). We propose an Intelligent Automatic Treatment Planning (IATP) framework including two artificial intelligence agents to emulate the roles of human physicians and planners. Based on a Virtual Physician Network, constructed with the deep neural network to model physicians' preferences for plan approval for prostate cancer Stereotactic Body Radiation Therapy (SBRT), we developed a Physician Preference-guided Virtual Treatment Planner Network (PgVTPN) via an end-to-end deep reinforcement learning approach. PgVTPN was able to operate the Eclipse TPS autonomously, like a human planner under physician's guidance, to generate prostate SBRT plans that meet dosimetric requirements and closely align with physicians' clinical desires. We connected PgVTPN to Eclipse TPS via ESAPI. Statistical analysis revealed that PgVTPN-generated plans exhibited a significantly higher total score of 252.99 ± 15.57 compared to clinical plans with a score of 237.29 ± 18.72. PgVTPN demonstrated superiority over human planners in key metrics, including target conformality, $V_{Rectum}[36Gy]$, $D_{Rectum}[40\%]$, $D_{Urethra}[20\%]$, $D_{Bowel}[1cc]$, and $D_{RFemoralHead}[max]$, with statistical significance. The plans were deliverable and successfully passed quality assurance with a gamma passing rate (3%/3mm) of 96.3%. Blind evaluation by experienced physician and physicists favored PgVTPN-generated plans on average 90% over corresponding clinical plans.

1 Introduction

Radiation therapy (RT) uses high-energy radiation beams generated by a medical linear accelerator (LINAC) to deliver a potent amount of radiation dose to the cancerous target and kill cancer cells. Patient-specific treatment planning, the process of determining LINAC control parameters to generate a desired 3D dose distribution based on the patient's anatomy, is crucial for RT's success, as it ensures optimal coverage of the tumor target with a prescription dose while minimizing exposure to nearby healthy organs. Advanced delivery techniques such as Intensity-modulated RT (IMRT) and Volumetric-modulated RT (VMAT) are planned by human operators using a commercial treatment planning system (TPS). This process requires the planner to optimize a multi-criterion problem to achieve the desired dose distribution under physician guidance. The trial-and-error parameter tuning process in TPS and the iterative planner-physician evaluations contribute to a complex planning workflow. Heavy reliance on human intervention results in variable plan quality and low workflow efficiency, emphasizing the need for a more streamlined approach. The fundamental issue is that TPS can only solve an optimization problem defined by the human planner, but lacks the human intelligence to decide on plan acceptability and steer a plan toward that.

To overcome this problem, we proposed an Intelligent Automatic Treatment Planning (IATP) framework (Fig. 1) that generally follows the current human-centered planning workflow but with two artificial intelligence (AI) agents to emulate the roles of human physicians and planners in this process, hence achieving automatic treatment planning. Specifically, a Virtual Physician Network (VPN) was trained to assess the likelihood of clinical plan approval based on patient anatomy and identify areas for improvement in achieving the desired dose. A Virtual Treatment Planner Network (VTPN), developed via deep reinforcement learning (DRL), autonomously operates the TPS to refine treatment planning parameters to generate high-quality RT plans, similar to a skilled human planner.

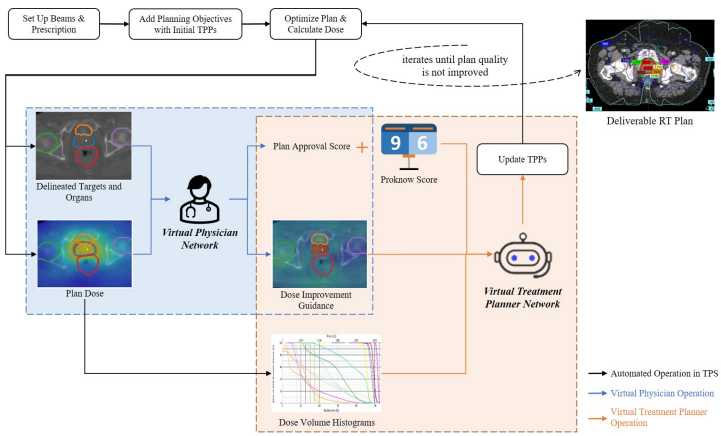


Figure 1: Proposed IATP framework, consisting of two AI agents, to emulate the roles of human physicians and planners in the treatment planning process using a commercial TPS.

In a previous study for prostate cancer Stereotactic Body Radiation Therapy (SBRT), we successfully developed a VPN[1]. This study reports the incorporation of VPN into the development of a Physician Preference-guided VTPN (PgVTPN) that can operate a clinical TPS as a virtual agent to produce treatment plans aligning with the physician's clinical desire. In addition to employing the VPN's output of the likelihood of clinical plan approval as the reward function in the DRL training process, our study also used the VPN's output of identified areas for improvement to guide the PgVTPN model training, effectively overcoming the low-efficiency burden of DRL training.

2 Materials and Methods

2.1 Virtual Physician Network

In a previous work[1], we have developed a VPN with decision-making capabilities on plan approval of prostate SBRT. The input to VPN included binary masks of PTV and eight OARs and a dose distribution to be evaluated (shown in blue in Fig. 1). The VPN consists of two sub-networks. One employs a hierarchically densely connected U-Net (HD U-Net) to generate a dose distribution highlighting desired changes from the input. The second sub-network has the structure of HD U-Net's encoder, followed by fully connected layers and a sigmoid activation function, to output the probability of approval for the given plan based on the anatomy. Since the clinical routine only stores approved plans in the TPS database whereas the unapproved plans were discarded, we trained VPN using the approved plan data via an adversarial training scheme.

VPN was trained on 60 patients with a four-fold cross-validation scheme, and subsequently evaluated on 8 independent test cases. It was able to differentiate approved and unapproved plans with Area under the curve 0.97 for testing data. For unapproved plans, after applying VPN's suggested dose improvement, the improved dose agreed with ground truth with relative differences $2.03 \pm 2.17\%$ for $D_{PTV}[98\%]$, $0.49 \pm 0.29\%$ for $V_{PTV}[95\%]$, $3.08 \pm 2.24\%$ for $D_{PenileBulb}[mean]$, $3.73 \pm 2.20\%$ for $V_{Rectum}[50\%]$, and $2.06 \pm 1.73\%$ for $V_{Bladder}[50\%]$. For more detailed information, readers can refer to the full paper [1].

2.2 Virtual Treatment Planner Network

For VTPN, we constructed it with three sub-networks in a hierarchical form [2], Structure-Net, Parameter-Net, and Action-Net, to mimic the decision-making process of a human planner operating a clinical TPS, Eclipse (Varian Medical System Palo Alto, CA), in terms adjusting treatment planning parameters (TPPs, e.g. priority, dose limit, or volume constraint) in the inverse optimization interface to steer plan towards high quality. Specifically, the VTPN took Dose-Volume Histograms (DHVs) of planning structures as input. By observing the DVHs, the three networks sequentially determine the structure for which TPP should be adjusted, which parameter to adjust for the selected structure, and how to adjust the TPP (increase or decrease.)

Training VTPN was achieved via an end-to-end DRL scheme with a hierarchical Q-learning framework. Let us denote the state of observed DVHs d and the action of TPP adjustment C. The goal of training VTPN is to determine the action-value function $Q^*(d,C)$ represented by the hierarchical networks.

For this purpose, we trained the VTPN to operate an in-house developed GPU-based Eclipse-like TPS with the same inverse planning optimization algorithm implemented and realistic dose engine. The purpose of using this in-house TPS rather than Eclipse was to overcome the low efficiency challenge when using Eclipse repeatedly in the long DRL training process. VPTN repeatedly interacts with the plan optimization process with certain actions and observes plan changes. Each interaction step l generates a data (d^l, C^l, d^{l+1}, r^l), where d^{l+1} is the DVHs for the plan after applying the action C^l to the TPS's optimization engine to generate a new plan, and r^l is the reward by going from d^l to d^{l+1}, which will be presented in the next subsection.

Through the data generated in this process, the network training was achieved by the update:

$$Q^*(d^l,C^l) \leftarrow Q^*(d^l,C^l) + \alpha \cdot [r^l + \gamma \cdot \max_C Q^*(d^{l+1},C^{l+1}) - Q^*(d^l,C^l)], \tag{1}$$

where α is the learning rate to control the weight given to the new information obtained. $\gamma \in [0,1]$ is a discount factor that determines the importance of future rewards in the learning process.

2.3 Physician Preference-guided VTPN

In clinical practice, physicians determine the treatment plan's acceptability based on quantitative metrics on plan quality and other considerations such as patient conditions. To realize PgVTPN, we incorporated the previously developed VPN in the VTPN training process from two angles.

First, PgVTPN employs a reward function combining the ProKnow system and the probability of plan approval from the VPN to represent the requirements both from quantitative plan evaluation and the physician's preference. The ProKnow system comprises 15 metrics specific to prostate SBRT, resulting in a full plan score denoted as $P(d)$ with a maximum of 150. We further denote the plan approval probability of the trained VPN as $VPN(A,d)$, where A is the patient anatomy represented by structure masks. We define the reward when changing a plan with DVHs d to d' as $r(d',d,A) = [P(d') + \kappa VPN(A,d')] - [P(d) + \kappa VPN(A,d)]$, where $\kappa = 150$ is a scaling factor to equally weight the ProKnow score with the plan acceptance probability. The reward r measured the difference in plan quality considering both quantitative metrics and the likelihood of physician approval. A positive reward represented the plan quality was improved, and a negative value otherwise.

Second, PgVTPN incorporates in the training process suggested dose improvements by the VPN. An important aspect in training the VTPN is to have it interact with the TPS, and observe the consequences represented by the reward for making certain actions, namely changes in TPPs. Two aspects have to be balanced in this training process. First, this process has to allow exploration of various actions to allow the VTPN to discover effective strategies in TPP adjustment. We achieved this via a ε-greedy scheme. At each step, PgVTPN

randomly selects an action of TPP adjustment with probability ε to the solution space. Second, meanwhile, it is equally important to have the process to adequately observe effective TPP adjustments to learn the strategy to generate them. Yet it is conceivable that effective TPP adjustments are much more scarce than ineffective ones. It is hence imperative that the training process can effectively generate effective TPP adjustment actions to improve overall training efficiency. As such, we used the dose improvement suggested by the VPN to identify which planning structure should be prioritized to optimize and which direction of TPP adjustment should be implemented in the next planning iteration.

2.4 Evaluation

We retrospectively identified 25 prostate cancer patients who were treated with 45 Gy in 5 fractions SBRT using VMAT between 2017 and 2020 at our institution [3]. We conducted a thorough assessment to evaluate the performance of the IATP framework. For testing, we coupled the trained PgVPTN with the real clinical Eclipse TPS, as opposed to the in-house one in the development, using ESAPI to allow PgVTPN to directly control the plan optimization process. Like a human planner, the trained PgVPTN was able to autonomously operate Eclipse TPS to generate a plan for each patient. Standard clinical metrics, emphasizing target coverage and critical organ sparing in line with clinical guidelines, were employed to assess all RT plans. Additionally, a blinded assessment was carried out by three clinical reviewers (a radiation oncologist and two experienced physicists), who compared the plan generated by PgVPTN with the clinical plan used in treatment for each patient case. To assess the clinical deliverability of the IATP system, we randomly selected three plans and performed patient-specific quality assurance (QA).

3 Results

3.1 Planning Process of a Representative Case

Fig.2 shows improvements in plan quality for a representative case during the planning process. The ProKnow score $P(d)$, VPN plan approval score $\kappa VPN(A,d)$, and total plan scores $P(d)+\kappa VPN(A,d)$ exhibited an overall increasing trend, rising from 138.98 to 142.10, 43.33 to 122.37, and 182.31 to 264.47, respectively. The clinical plan used in treatment for this case scored 140.06, 109.15, and 249.21. Fig.2 (b) display cross-sectional dose distributions at planning steps 0, 4, and 10 for PgVPTN to operate Eclipse TPS, and (c) compare DVHs at these steps, demonstrating the autonomous behavior of PgVPTN in making decisions to perform treatment planning like a human.

3.2 Comprehensive Assessments on 25 Patients

Fig. 3 (a) displays the distribution of plan scores in terms of likelihood of plan approval $\kappa VPN(A,d)$ and quantitative ProKnow scores $P(d)$. The IATP plans generated by PgVPTN are slightly more clustered in the upper right corner, indicating the higher plan quality in both two aspects. Fig. 3 (b-d) plots the plan scores $P(d)+\kappa VPN(A,d)$, ProKnow scores $P(d)$, and the likelihood of plan approval $\kappa VPN(A,d)$ for each patient, with the two axes being the clinical plan and IATP plan scores. The distribution below the diagonal line revealed that IATP plans exhibit superior quality with higher scores.

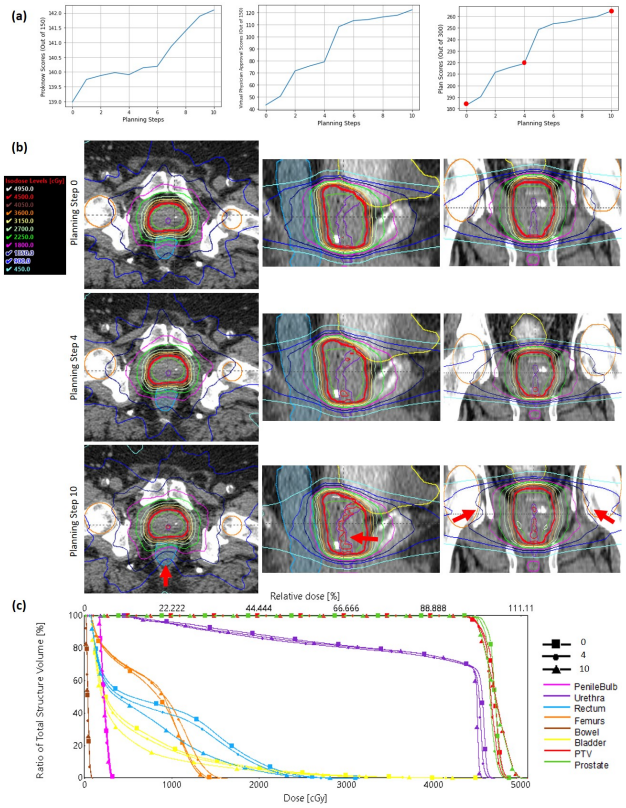


Figure 2: (a) Evoluation of plan scores during the planning process. (b) and (c) are cross-sectional dose distributions and DVHs at steps 0, 4, and 10 of PgVPTN operating Eclipse.

Fig. 4 provides a detailed comparison of clinically relevant plan quality metrics, monitor units (MUs), VPN plan approval scores, and total plan scores between clinical plans and their IATP counterparts. The IATP-generated plans achieved an average total score of 252.99 ± 15.57 and an average VPN plan approval score of 121.63 ± 13.44 compared to the clinical plans scored at 237.29 ± 18.72 ($p = 0.003$) and 110.41 ± 17.63 ($p = 0.017$), respectively. Overall, the IATP-generated plans outperformed the clinical plans, achieving higher scores in both total plan score and VPN plan approval score, with statistically significant differences. Furthermore, IATP-generated plans also demonstrated statistically significant improvements over human-planned ones in several specific metrics such as the Conformality Index of PTV ($p = 0.0$), $V_{Rectum}[36Gy]$ ($p = 0.013$), $D_{Rectum}[40\%]$ ($p = 0.0$), $D_{Urethra}[20\%]$ ($p = 0.0$), $D_{Bowel}[1cc]$ ($p = 0.049$), and $D_{RFemoralHead}[max]$ ($p = 0.018$). While the average monitor units for IATP-generated plans were slightly higher

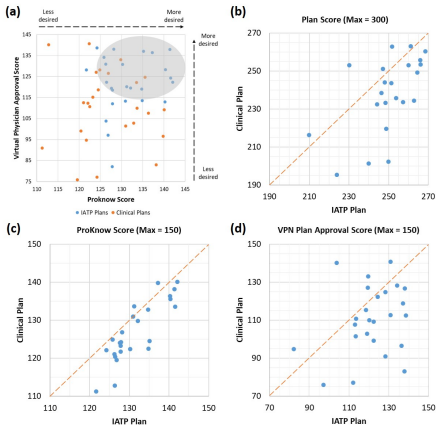


Figure 3: (a) Distribution of plan scores for the clinical plan and the PgVTPN generated plans. The shaded area indicates a cluster of high-quality plans. (b-d) Comparison of total plan scores, ProKnow score, and VPN plan approval score between clinical plans developed by human planners and those generated by IATP.

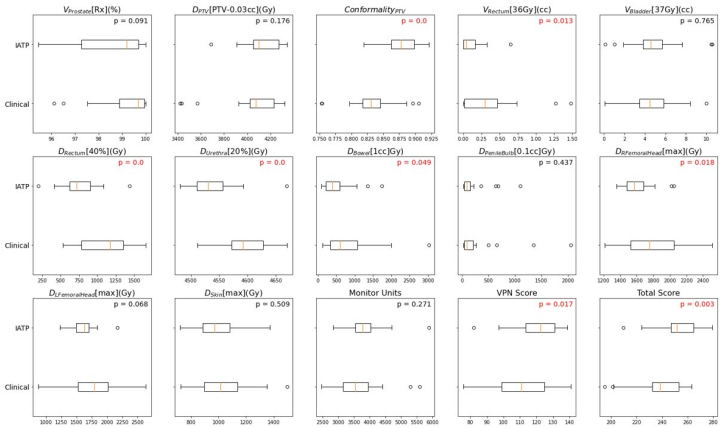


Figure 4: Boxplots of clinically relevant plan quality metrics, MUs, VPN plan approval score, and the total plan score, comparing clinical plans and their corresponding plans generated by IATP. The left and right quartiles define the edges of the boxes, and any point exceeding 1.5 times the interquartile range is considered an outlier. Vertical lines inside the boxes represent median values. The p-values below 0.05 are highlighted to indicate statistical significance.

(3836.62 ± 552.75) than clinical plans (3627.56 ± 733.54), the difference was not statistically significant $p = 0.271$. The IATP workflow spent an average of 42.05±10.71 minutes to generate a high-quality prostate SBRT plan, with the majority of time spent on plan optimization and dose calculation in Eclipse. The decision made by PgVPTN to adjust TPP was almost instantaneous, highlighting the efficiency. All three randomly selected PgVPTN-generated plans for QA successfully passed the gamma index analysis (3%/3 mm), with a passing rate of an average of 96.3%, confirming its clinical deliverability.

3.3 Blinded Evaluation by Three Reviewers

Figure 5 displays the results of a blinded assessment conducted by three reviewers including one physician and two

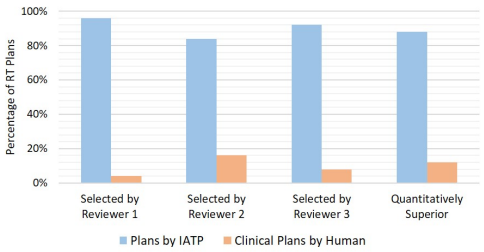


Figure 5: Three clinical reviewers evaluated 25 patients, each having one plan generated by IATP and one clinical plan, in a blinded assessment.

medical physicists. Reviewers favored IATP-generated plans over clinical plans in 96% of patient cases (24/25), 84% of cases (21/25), and 92% of cases (23/25), respectively. The one IATP-generated plan not chosen by all reviewers involved a patient with a left femoral head hip replacement, where the human planner utilized beam angle avoidance, which the IATP system failed to consider.

4 Conclusion

This study presented the IATP framework for prostate cancer SBRT treatment planning. Through end-to-end DRL training, a PgVTPN was developed to autonomously operate a clinical TPS, like a human planner, to generate plans meeting dosimetric requirements and aligned with physicians' clinical preferences. The integration of IATP into the clinical workflow was achieved seamlessly with a commercial TPS. The IATP-generated plans exhibited a statistically superior total score of 252.99 ± 15.57 compared to 237.29 ± 18.72 of the clinical plans. IATP outperformed in dosimetric metrics such as PTV conformality, $V_{Rectum}[36Gy]$, $D_{Rectum}[40\%]$, $D_{Urethra}[20\%]$, $D_{Bowel}[1cc]$, and $D_{RFemoralHead}[max]$ with statistical significance. In a blind evaluation conducted by three experienced reviewers, PgVPTN-generated plans were preferred over the corresponding clinical plans in 90% of cases on average, demonstrating the effectiveness of our development.

References

[1] Y. Gao, C. Shen, Y. Gonzalez, et al. "Modeling physician's preference in treatment plan approval of stereotactic body radiation therapy of prostate cancer". *Physics in Medicine & Biology* 67.11 (2022), p. 115012. DOI: 10.1088/1361-6560/ac6d9e.

[2] C. Shen, L. Chen, and X. Jia. "A hierarchical deep reinforcement learning framework for intelligent automatic treatment planning of prostate cancer intensity modulated radiation therapy". *Physics in Medicine & Biology* (2021). DOI: 10.1088/1361-6560/ac09a2.

[3] L. Chen, B. S. Gannavarapu, N. B. Desai, et al. "Dose-Intensified Stereotactic Ablative Radiation for Localized Prostate Cancer". *Frontiers in Oncology* 12 (2022), p. 779182. DOI: 10.3389/fonc.2022.779182.

An exploration of the complexities of CTV mapping

Marcel van Herk, Matt Craddock, Mairead Daly, Eliana Vasquez Osorio

Division of Cancer Sciences, University of Manchester, UK
The Christie NHS Foundation Trust, Manchester, UK

Abstract. A naïve CTV description could be a population-averaged cell density of microscopic disease. We show, by simulation, that there is a huge impact of the *mode* of cell density variation on margin requirements. Our model consists of a spherical GTV, surrounded by small nodules at given distance. Geometrical errors were simulated and TCP calculated for many scenarios with varying margin. The three modes of CTV nodules variation were: 1: varying cell density; 2: varying nodule probability, with all nodules present/absent at once, 3: varying nodule probability with each nodule independent. We report the CTV+PTV margin at 2% TCP loss from large margins. Per 0.1 times density reduction (mode 1) CTV margin reduces 1mm. For mode 2, CTV margin reduced by 4mm at p=0.1. For mode 3, the margin is in-between. Since all these modes share the same population averaged tumour cell density, we demonstrated that this representation is insufficient.

1 Introduction

There is currently a renewed interest in providing a more detailed description of the CTV, rather than using fixed or anatomical margins based on historical data. According to ICRU, "*The Clinical Target Volume (CTV) is a volume of tissue that contains a demonstrable GTV and/or subclinical malignant disease at a certain probability considered relevant for therapy".* In the lung setting, there have been clinical studies omitting a CTV margin in higher stage patients with apparently no loss of local/marginal control [1]. Also, the latest pathology findings show that current 5 mm CTV margins may be generous in small lung tumours [2]. However, to qualify such results a more quantitative definition of the CTV would be required.

A naïve assumption would be to describe the CTV by a population averaged tumour cell density [3]. In clinical practice quantitative information about the CTV is very hard to collect, and significant efforts are required [4]. Better TCP modeling and statistical methods are required to make a more scientific definition of the CTV possible. But even if such information would be available, is it enough? Here we simulate a very simple model of GTV and CTV to demonstrate the large impact of the details of *how* the tumour density varies on the required safety margins, given that the actual CTV would be unknown. We hope that these results contribute to a better understanding of the requirements of CTV 'maps' that form a core element of the current discussions and upcoming ICRU recommendations.

2 Materials and Methods

Our (very simple) tumour model consists of a spherical GTV with radius of 2.5 cm, surrounded by 26 small nodules with 1/1000 of the GTV volume at 8 mm distance (**Figure 1**). This distance is the median nodule distance observed previously in lung cancer [5]. We chose 26 nodules to easily cover all directions around the sphere. Random and systematic geometrical errors were simulated (Σ=3.9mm, σ=2.6mm), requiring about 11 mm PTV margin according to the van Herk recipe [6]. We assumed a dose prescription of 66Gy (in 2Gy fractions). Density variation of the nodules was simulated in three modes (**Figure 2**): 1: by only varying cell density, all nodules at the same time, 2: by varying the probability of nodule presence, with all nodules present our absent at the same time, 3: by varying the probability of nodule presence with each nodule independently absent or present. Using an existing simulation framework [7] (**Figure 3**), TCPs were calculated for a range of isotropic margins. Our parameters where chosen to give a more or less realistic TCP of 62% for stage 3 lung cancer, $\alpha = 0.35\text{Gy}^{-1}$, $\sigma_\alpha = 0.08\text{Gy}^{-1}$ and tumour cell density of $1.8*10^7/\text{cm}^3$. We arbitrarily define the optimal CTV margin as that margin where the population-based TCP loss is 2% compared to very large margins. Such a choice represents a reasonable balance of the risk of geometrical miss versus organ at risk damage at larger margins.

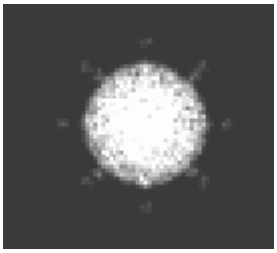


Figure 1: Visualisation of the very simple simulated GTV and CTV. The GTV is surrounded by 26 nodules at 8 mm distance and equally distributed in all directions in 3D. Depending on the simulation, nodules may be turned off such that the average number of present nodules can be much smaller.

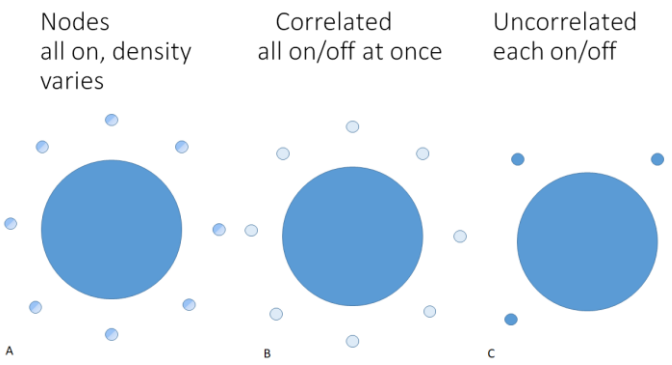


Figure 2: Visualisation of the difference of CTV simulation. The 26 nodules are A) all varied in cell density, B) all switched on or off at the same time at given probability, or C) independently switched on or off at a given probability. Note that these three simulations are fully identical in terms of population-averaged cell density.

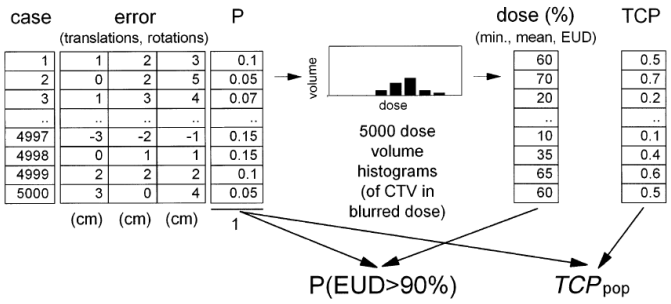


Figure 3: Simulation framework (taken from [7]).The simulation applies 5000 error scenarios and for each scenario a random CTV is drawn according to the modes shown in **figure 2**. An idealised dose distribution with varying margins is used as input. The population TCP is reported here.

To summarise our simulations, we plot contour lines of TCP as function of (PTV+CTV) margin and density/ probability, the latter simulated over a logarithmic range from 1 to 0.0001 in half decade steps.

3 Results

Figure 4 shows the summary plots for our experiments, where we observed the highest TCP to be 0.62.

The required margin to achieve 60% TCP as function of density or probability is quite different for the three types of simulations. Cell density reduction allowed a linear reduction of margin at about 1% per decade lower density (**Figure 4a**).

Nodule probability reduction behaves quite different, because at low probability the population TCP is no longer affected, since e.g. at p=0.001 only one in thousand patients will fail. There is a large difference between correlated (**Figure 4b**) and uncorrelated probability (**Figure 4c**). With uncorrelated probability, the required margin increases by 2mm for at a 0.1 probability level.

4 Discussion

This work simulates three possible modes of variation for tumour density and nodule presence probability in the same geometry that have identical cell density distributions on a population level. However, the margin required at 2% TCP loss depends strongly on the mode of variation. As expected, we find that all margins are the same at p=1, about 16mm. This is about 3 mm less than calculated by the van Herk recipe (11 mm) plus the distance from the GTV to the simulated nodes (8 mm). A somewhat lower margin based on TCP considerations is expected because TCP based margins are more resilient to

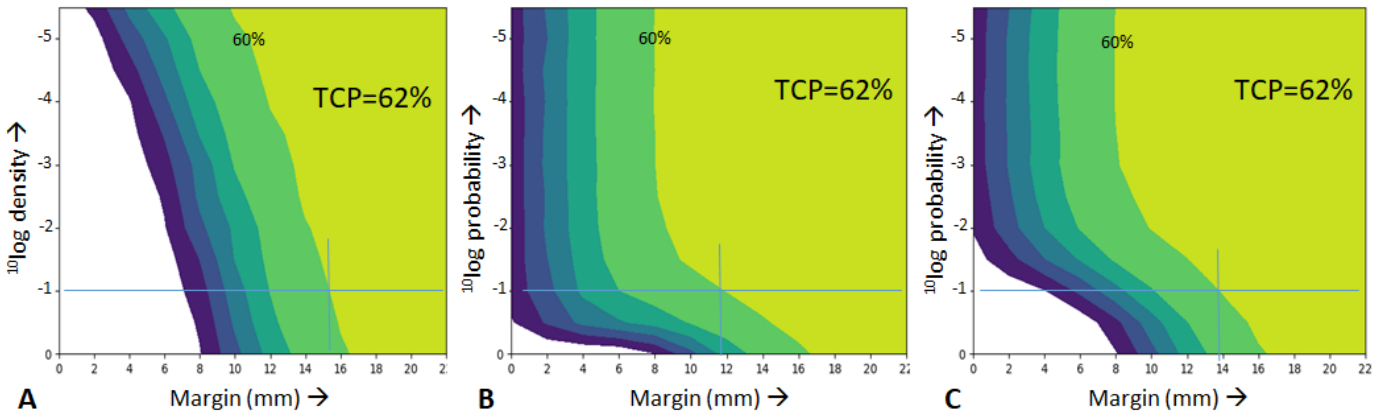


Figure 4. Dependence of TCP (colour) on margin (horizontal axis) and average tumour density/probability (vertical axis) for different modes of tumour density variations. In all three graphs, the population-based cell density distribution is the same. Colour steps are per 5% TCP. The green-yellow transition is at a TCP of 60%, while the highest TCP is 62%. A) Cell density reduction, the same for all patients and nodules. B) Nodule probability reduction, where all nodules are either present or absent at the same time for each individual patient. C) Nodule probability reduction where each nodule is treated independently. The lines indicate the required PTV+CTV margin at a probability or density of 0.1 to achieve a TCP of 60%.

edge under-dose than dose-probability based margins [7].

Most surprising is the difference between correlated and uncorrelated probability of node presence. We believe this finding demonstrates a combination of two effects. First, there is an effect of CTV distribution on margins for *geometrical* uncertainties. If nodules are present all around the GTV, geometrical errors in any direction will cause a nodule miss, i.e. the CTV is at risk in 3 dimensions. At lower probability, there will be fewer nodules. This means that only errors exactly in the direction of a nodule will lead to miss. In other words, the CTV will be at risk in less than 3 dimensions and therefore a reduced margin can be accepted [6]. However, here we find an increased margin requirement. This is related to the degrees of freedom in the variation. In mode 2 there is one degree of freedom, if all nodes are on at once they be missed everywhere at once. In mode 3 there are up to 26 degrees of freedom, and the likelihood of missing one or more of them is bigger for $p>0.04$. The combined effect is a margin increase of 2mm at $p=0.1$.

We chose to use 26 nodules and of course mode 3 will be enormously affected by this number as they translate to degrees of freedom. In modes 1 and 2, the total nodule volume is nodules is more likely to determine the behaviour. However, it is very likely that the results for mode 3 will scale, e.g. taking 10 times more nodules at 10 times lower probability would give the same results, The degrees of freedom for geometrical miss relates to a spatial correlation in nodule presence that *must* be taken into account for proper margin calculation. This is a similar situation as described for contouring, where Mckenzie et al [8] described ‘dimples and pimples’, and Vasquez Osorio et al used spatial filtering on a level set to create realistic contours [9]. Biologically, such a behaviour may be plausible due to development of different clonal regions in a tumour where only those with certain mutations start disseminating. Correlation descriptions are required to correctly add CTV and PTV margins and for robust treatment planning for the CTV.

By necessity our simulation is very simplistic and it would be of interest to add more variation. However, our results do clearly show the insufficiency of a population-averaged cell density model.

5 Conclusion

This is a first simulation of the mode of CTV density variation and its effect on margin requirements. The average density does not and cannot reflect the mode of cell density variation and is insufficient for defining a CTV margin, It would not give the expected average TCP in a real patient population with unknown mode of density variations, and should not be used for CTV mapping. Correlation must be considered to correctly add CTV and PTV margins. Given the many unknowns it is important to perform clinical trials to determine the optimal CTV margin for each clinical site.

Acknowledgements

We would like to acknowledge Peter Remeijer, Marnix Witte and Joos Lebesque at the Netherlands Cancer Institute who in the past greatly contributed to the simulation platform utilized here.

References

[1] Cui T, Zhang A, Cui J, Chen L, Chen G, Dai H, Qin X, Li G, Sun J. Feasibility of omitting the clinical target volume under PET-CT guidance in unresectable stage III non-small-cell lung cancer: A phase II clinical trial. Radiother Oncol. 2023 Apr;181:109505. doi: 10.1016/j.radonc.2023.109505.

[2] Schmitt M, Aussenac L, Seitlinger J, Lindner V, Noël G, Antoni D. Evaluation of Microscopic Tumour Extension in Localized Stage Non-Small-Cell Lung Cancer for Stereotactic Radiotherapy Planning. Cancers (Basel). 2022 Mar 2;14(5):1282. doi: 10.3390/cancers14051282.

[3] Brahme A, Agren AK. Optimal dose distribution for eradication of heterogeneous tumours. Acta Oncol. 1987;26(5):377-85. doi: 10.3109/02841868709104364.

[4] Apolle R, Rehm M, Bortfeld T, Baumann M, Troost EGC. The clinical target volume in lung, head-and-neck, and esophageal cancer: Lessons from pathological measurement and recurrence analysis. Clin Transl Radiat Oncol. 2017 Mar 21;3:1-8. doi: 10.1016/j.ctro.2017.01.006.

[5] van Loon J, Siedschlag C, Stroom J, Blauwgeers H, van Suylen RJ, Knegjens J, Rossi M, van Baardwijk A, Boersma L, Klomp H, Vogel W, Burgers S, Gilhuijs K. Microscopic disease extension in three dimensions for non-small-cell lung cancer: development of a prediction model using pathology-validated positron emission tomography and computed tomography features. Int J Radiat Oncol Biol Phys. 2012 Jan 1;82(1):448-56. doi: 10.1016/j.ijrobp.2010.09.001.

[6] van Herk M, Remeijer P, Rasch C, Lebesque JV. The probability of correct target dosage: dose-population histograms for deriving treatment margins in radiotherapy. Int J Radiat Oncol Biol Phys. 2000 Jul 1;47(4):1121-35. doi: 10.1016/s0360-3016(00)00518-6.

[7] van Herk M, Remeijer P, Lebesque JV. Inclusion of geometric uncertainties in treatment plan evaluation. Int J Radiat Oncol Biol Phys. 2002 Apr 1;52(5):1407-22. doi: 10.1016/s0360-3016(01)02805-x.

[8] McKenzie A. A novel way to allow for uncertainties in delineation and changes in shape of target volumes in radiotherapy. Proceedings of the 11th Annual Scientific Meeting. Institute of Physics in Engineering and Medicine, 2004.

[9] Vásquez Osorio E, Shortall J, Robbins J, van Herk M. Contour Generation with Realistic Inter-observer Variation. 19th International Conference on the use of Computers in Radiation Therapy, Montreal, 2019.

Biologically-guided multi-criteria optimization and exploration to generate clinical plans for functionality-sparing radiation therapy trials

Dimitre Hristov[1], Matthew Jinsoo Kim[1], Jen-Yeu Kim[1], and Elizabeth Kidd[1]

[1]Department of Radiation Oncology, Stanford University, Stanford, CA 94305

Abstract

Purpose: Functional imaging of heterogeneous organ function offers the opportunity to explore and evaluate novel treatments that maximize preserved organ functionality while satisfying other constraints. Designing such treatments with clinical planning systems is a challenge since these systems cannot drive an optimization according to an image-based functionality surrogate. The purpose of our work is to demonstrate an optimization approach that overcomes this challenge entirely with tools available in a commercial treatment planning system (TPS).

Methods: We implemented the optimization approach with multicriteria optimization (MCO) and scripting tools in Eclipse TPS v15.6 (Varian, Palo Alto). We start with a standard plan for which an organ with heterogeneous imaging appearance meets clinical dose-volume constraints. We segment this organ in non-overlapping sub-regions of equal volumes and increasing functional image intensity. For these sub-volumes and planning targets, starting with the clinical plan, we generate a database of plans with (MCO). We then optimize functionality sparing with simulated annealing exploration of the plan database. In this exploration we select a plan by setting sub-volume objectives to track closely their randomly assigned maximum limits. We gauge the utility of a selected plan by a user-defined cost function that calculates a surrogate metric of preserved functionality from image and dose values at voxels within the organ.

Results: We used this approach with interactive exploration to support a Phase I/II study," Bone Marrow Sparing Image-Guided Radiation Therapy Incorporating Novel Use of GCSF and FDG-PET Imaging", NCT04514692. The surrogate metric of preserved functionality was the cumulative FDG activity receiving doses below 20 Gy in pelvic bone. For 70% of the enrolled patients treatment proceeded with an MCO plan as functional sparing was deemed significant.

Conclusion: We have demonstrated a clinically deployable planning approach to support single- and multi-institutional clinical trials of novel biologically-guided function-preserving radiotherapy treatments.

1 Introduction

Functional and molecular imaging of heterogeneous organ function offers the opportunity to design and conduct clinical trials that evaluate the benefit of novel treatments that incorporate such imaging in plan optimization to maximize preserved organ functionality while still satisfying other treatment planning constraints. While designing such treatments with clinical treatment planning systems (TPSs) is highly desirable to assure all treatment workflow and quality aspects are met, it poses several challenges. First, once structures (organs) are segmented from anatomical, molecular, or functional images clinical TPSs treat these structures as homogeneous in function, completely ignoring the functional imaging context. Second, clinical TPSs optimize treatment plans with hard-coded pre-defined (typically) dose-volume histogram (DVH) metrics of plan quality (cost functions) which cannot drive an optimization according to user-defined image-based functionality metrics that will vary depending on which organ and function are considered for preservations and which imaging modality is used to capture the heterogeneity of this function.

One potential approach to overcome these challenges is to segment a structure with heterogeneous function in non-overlapping sub-regions (sub-structures) with equal volumes and increasing functional intensity and then include these structures with the hard-coded TPS objective (cost) functions (also termed objectives) and user-assigned objective weights in the standard TPS optimization to generate a plan. Using TPS scripting tools, the image-based functionality metric can be then calculated and in principle used to guide a new selection of the objective weights automatically or interactively.

A conceptual "viscous circle" type of problem with this approach is that in selecting an optimal plan a clinician will have to gauge the relative merit of gains and losses of standard DVH metrics against gains and losses of a newly investigated functionality metric, when the goal of the clinical study itself is to establish the validity and the merit of using such functionality metric. Hence, in what follows we propose and demonstrate a novel "First do no harm" optimization approach to functionally-guided radiotherapy that is entirely implemented with tools available in a commercial TPS, thus making it immediately deployable clinically.

2 Materials and Methods

Description of Optimization Strategy

Our optimization approach is based on the use of multicriteria optimization (MCO) available in commercial TPSs and is described below with reference to a simple scenario shown in Figure 1. We start with a clinical plan that is optimized as per physician and institutional guidelines and accepted for treatment. For this plan the structure with heterogeneous function attains a clinical DVH objective such as V_D (absolute or relative volume below dose D) equal to $V_{attained}$ (see Figure 1). We segment this organ in N (N = 2 in Figure 1) non-overlapping sub-structures of equal volumes and increasing functional image intensity (substructure **R**educed **I**mage **I**ntensity or ***RII***, and **E**levated **I**mage **I**ntensity or ***EII*** in Figure 1). Starting with the clinical plan, for these sub-structures and for planning targets (the TPS requires targets to be included), we generate a database of plans with multicriteria optimization (MCO) after removing the initial structure from the optimization. The initial target value of the DVH objectives for the substructures ($^{RII}V_D$ and $^{EII}V_D$ in Figure 1) are set to match the values attained

by the clinical plan. To enforce the "First do no harm" principle other objectives are excluded from the MCO and thus effectively locked at their previously achieved and clinically acceptable values with the clinical plan.

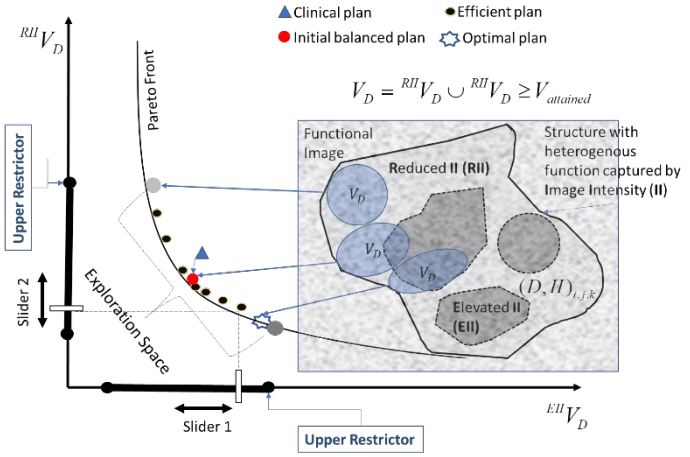


Figure 1 A simple example to illustrate the MCO strategy for maximizing spared functionality. The initial balanced plan (red marker) generated from a clinical plan (blue triangle) spares a given region V_D (in blue) from the structure with heterogeneous function as reflected by substructures with elevated (EII) and reduced image intensity (RII). MCO optimization with trade-off objectives ${}^{RII}V_D$ and ${}^{EII}V_D$ creates exploration space of Pareto efficient plans (black dots) with dose distributions that trade between target objectives, ${}^{RII}V_D$, and ${}^{EII}V_D$ resulting in different dose distributions, illustrated by different V_D regions. In exploration the optimal plan is selected as the one that maximizes a user-defined cost function that calculates a surrogate metric of preserved functionality from paired dose and image intensity values $(D,II)_{i,j,k}$ while still maintaining the initial V_D.

Given a set of competing (trade-off) objectives and starting with an initial balanced plan generated from the clinical plan, the MCO tool in the commercial TPS generates a database of Pareto efficient plans which belong to a certain exploration space of the Pareto Front [1, 2]. Any plan that belongs to the Pareto front balances the trade-off objectives in efficient manner that is a trade-off objective cannot be further improved without worsening at least one or more other trade-off objective. This database of efficient plans is then made available for exploration in a "Trade-Off Exploration Context". In interactive mode a user manipulates sliders linked to each trade-off objective to express preferences towards different trade-off objectives and evaluates the generated treatment plan that corresponds to these preferences.

When the goal of such exploration is to maximize preserved organ functionality while conforming to other plan constraints, the user's preference should additionally be guided by an update of a cost function that calculates a surrogate metric of preserved functionality from paired dose and image intensity values for the currently selected plan.

While interactive exploration may be intuitive and feasible for small number of substructures, it becomes non-intuitive and impractical when an increased number of substructures is used to capture significant imaging heterogeneity within a structure. For this reason, we maximize functionality sparing with simulated annealing [3] exploration of the plan database ("Trade-Off Exploration Context"). In this exploration, at each step, we select a new plan candidate by setting new upper restrictors (Figure 1) for the substructures with random step from the upper restrictors current values. We then set the trade-off objective (Sliders in Figure 1) to track the upper restrictors as closely as possible while satisfying all other constraints. The utility of a plan candidate selected by the values of the trade-off objective is then gauged by a cost function of preserved functionality and acceptance or rejection of the candidate is performed according to the simulated annealing algorithm. Further detail on the simulated annealing exploration are given in the pseudocode in Figure 2.

```
INIT Temperature, upperRestrictorDelta
CALL FunctionalCostFunction (TradeoffExplorationContext, FunctionalImage,
Structure) RETURN currentBestCost /* Get the cost for the balanced plan */

WHILE stoppingCriterion NOT met
    CALL GetTradeoffExplorationContext
    RETURN currentBestUpperRestrictors, currentBestTradeOffObjectives

    /*Generate a random feasible neighbour of the current plan by setting
    in random order and directon feasible UpperRestrictors and
    tradeoffObjectives within restrictrorDelta of their current values */
    CALL SetNewUpperRestictorsAndObjectives (TradeoffExplorationContext,
    restrictrorDelta) RETURN newUpperRestrictors, new tradeoffObjectives

    /* Select a new plan candidate */
    CALL SetTradeoffExplorationContext(TradeoffExplorationContext,
    newUpperRestrictors, newtradeoffObjectives)

    /* Get the cost for the new plan candidate */
    CALL FunctionalCostFunction (TradeoffExplorationContext,
    FunctionalImage, Structure)  RETURN newCost

    COMPUTE AcceptanceProbability (newCost, currentCost, Temperature)

    IF (AcceptanceProbability > aRandomNumberBetweenZeroAndOne)
        SET currentBestCost to newCost
    ELSE
        /* Revert to the current best plan */
        CALL SetTradeoffExplorationContext(TradeoffExplorationContext,
        currentBesUpperRestrictors, currentBestTradeOffObjectives)
    DECREASE Temperature /* according to annealing schedule */
    DECREASE upperRestrictorDelta /* according to rate of decrease */
ENDWHILE
```

Figure 2 Pseudocode for the simulated annealing exploration of the MCO database.

Implementation

We have implemented the optimization strategy discussed above to enable functionality optimized planning for a pilot Phase I/II study," Bone Marrow Sparing Image-Guided Radiation Therapy Incorporating Novel Use of GCSF and FDG-PET Imaging", NCT04514692. We have developed the simulated annealing exploration using the Eclipse Scripting Application Programing Interface, Varian Medical Systems, Palo Alto, CA (ESAPI) v15.5. The surrogate

metric of preserved functionality is the cumulative FDG-PET activity $CA_{20\ \mathrm{Gy}}^{ABM}$ that receives doses lower that 20 Gy in an **A**ctive **B**one **M**arrow structure (ABM) segmented from the FDG-PET image. The corresponding functionality cost function is the relative Functional Loss

$$Functional\ Loss = 100 \times \left(1 - \frac{CA_{20\ \mathrm{Gy}}^{ABM}}{CA_{0\ \mathrm{Gy}}^{ABM}}\right)$$

whereby $CA_{0\ \mathrm{Gy}}^{ABM}$ is the total ABM activity (that would be spared if zero dose is delivered). To set the initial temperature we sample extreme (but unlikely to be clinically acceptable because of other planning constraints) plans from the plan database with maximum functionality sparing, calculate the differences in cost between these extremes and the balanced plan, and set the initial temperature T_0 high enough to allow high initial acceptance (0.8) of uphill moves for such differences. We use a standard acceptance probability function with Boltzmann distribution, and annealing schedule that reduces the initial temperature as $T_0 \times (0.95)^k$ where k is the iteration number of the annealing schedule. The initial step size for the upper restrictors is set to 0.5 (a half of the 0 to 1 exploration range) and at each iteration the step size is reduced by a factor of 0.98. Since setting a trade-off upper Restrictor and Objective automatically impacts all other tradeoff-objectives and available ranges for exploration, at each step we set the order of manipulating the different upper restrictors and objective randomly and we update the "Trade-Off Exploration Context" after manipulating a pair of an upper Restrictor and corresponding Objective. Furthermore, the currently manipulated Objective is scaled back if necessary by automatically evaluating clinical constraint and measures for the current plan candidate and subsequently updating the "Trade-Off Exploration Context". Thus, the optimal objectives for the final functionally sparing plan are not set solely by the values of the upper restrictors but also by the enforcement of clinical DVH constraints at each iteration step and the competition between all MCO objectives.

3 Results

Figure 3 shows an example of simulated annealing exploration of an MCO generated database for a clinical case. For this patient an active bone marrow structure was segmented at the mean FDG uptake in the pelvis bone and further sub-segmented in 2 structures HA (High Activity) and LA (Low Activity). Figure 3 shows that our approach led to ~ 8% reduction in functionality loss (from 89% to 81%). Figure 4 compares the optimized plan with the initial balanced plan to demonstrate the changes in dose distributions at 20 Gy that spare uptake activity in the pelvic bone. Figure 4 also illustrates that with regard to other clinical objectives and constraints the plans are equivalent.

We have also used our approach with interactive exploration to generate clinical plans for the pilot Phase I/II study," Bone Marrow Sparing Image-Guided Radiation Therapy Incorporating Novel Use of GCSF and FDG-PET Imaging", NCT04514692. In 70% (5/7) of the enrolled patients activity sparing with MCO was deemed significant to proceed with the MCO plan.

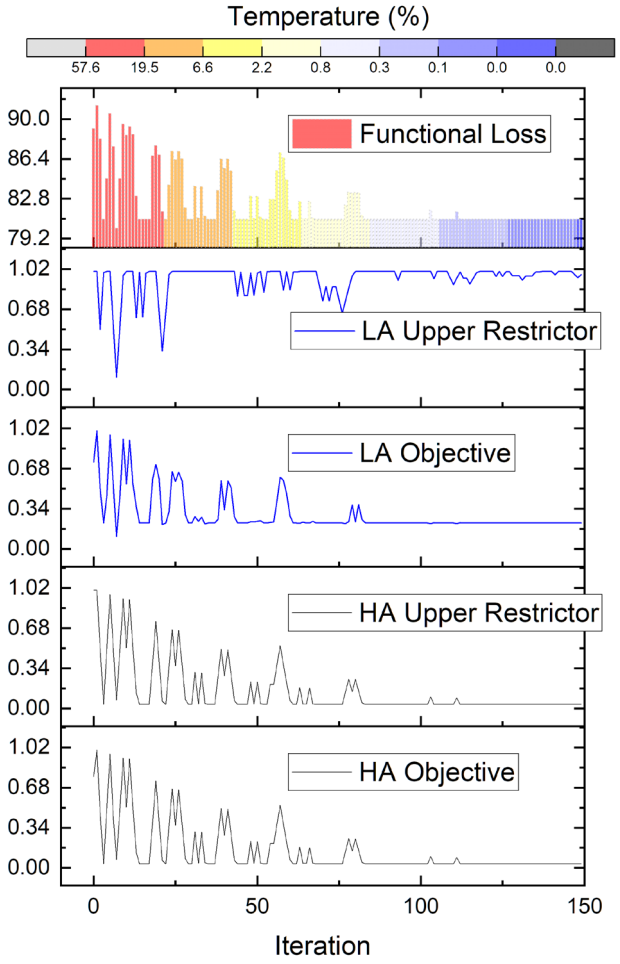


Figure 3 Example of simulated annealing exploration of an MCO generated database for a clinical case. HA and LA correspond to a High Activity Substructure and Low Activity Substructure. The accepted values of the Functional Loss are shown as a function of the iteration number colored with the temperature. Also shown are accepted values of restrictors and objectives during the optimization. Note that while the values for the HA Objective track closely these of the set values for the corresponding HA Upper Restrictor, at the solution the accepted LA Objective values are not limited by the corresponding LA Upper Restrictor but other clinical constraints.

4 Discussion

Our optimization approach offers several benefits.
First, it allows functional and molecular imaging information to guide the optimization of clinically deliverable plans to spare heterogeneous function for a

given organ while maintaining clinical DVH constraints for other structures by excluding these structures from the MCO.



Figure 4 ***Top:*** *Color wash dose display for a plan optimized with our approach.* ***Middle:*** *Color wash dose display for an initial balanced plan. Note that dose at 20 Gy tis redistributed in the optimized plan to spare activity in the pelvic bone. At the same time other constraints are met by both plans, as indicated by the green indicators below the images.* ***Bottom:*** *DVHs showing equivalent target coverage (in red) but decreased dose at 20 Gy for the Low Activity and High Activity substructures. MCO plan: solid lines, balanced plan: dashed lines.*

Second, our approach offers a way to navigate the MCO ("Trade-Off Exploration Context") in case where large heterogeneity in functional images requires many sub-structures to be segmented and interactive navigation is no longer intuitive. Third, it generalizes immediately to any functionally defined cost function, as no specific knowledge about the structure of the cost function is necessary. Furthermore, by normalizing the functionality loss from 0 to 100% and automatically selecting the initial temperature as discussed, it is likely that the hyper-parameters of the algorithm are applicable to other functionallity cost functions.

5 Conclusion

We have presented and implemented an optimization framework based on simulated annealing and multi-criteria optimization to support planning for clinical studies of functionality sparing radiotherapy based on heterogeneous functional imaging.

References

[1] K. H. Küfer *et al.*, "Intensity-modulated radiotherapy a large scale multi-criteria programming problem," (in English), *Or Spectrum,* vol. 25, no. 2, pp. 223-249, May 2003. [Online]. Available: <Go to ISI>://WOS:000182752600006.

[2] T. Bortfeld, C. Thieke, K. H. Küfer, and H. Trinkaus, "New approaches in intensity-modulated radiotherapy -: A new optimization paradigm," (in English), *Progress in Radio-Oncology Vii, Proceedings,* pp. 251-258, 2001. [Online]. Available: <Go to ISI>://WOS:000176997100035.

[3] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, "Optimization by Simulated Annealing," *Science,* vol. 220, no. 4598, pp. 671-680, 1983, doi: doi:10.1126/science.220.4598.671.

Deep Learning Guided Radiotherapy Modality Selection via Multi-Task Dose Prediction

Austen Maniscalco[1,2], Ezek Mathew[1,2], David Parsons[2], Justin Visak[1,2], Mona Arbab[2], Prasanna Alluri[1,2], Xingzhe Li[1,2], Narine Wandrey[2], Mu-Han Lin[1,2], Asal Rahimi[1,2], Steve Jiang[1,2], and Dan Nguyen[1,2]

[1]Medical Artificial Intelligence and Automation Laboratory, University of Texas Southwestern Medical Center, Dallas, Texas, USA
[2]Department of Radiation Oncology, University of Texas Southwestern Medical Center, Dallas, Texas, USA

Abstract In radiation therapy (RT), choosing an appropriate treatment modality is a crucial yet complex task due to the unique features of each modality that impact dose distribution. This complexity is compounded by the lack of a quantitative, patient-specific method for modality selection. To address this gap, we developed an efficient multi-task convolutional neural network. This model utilizes patient-specific data, such as computed tomography image and dosimetric goals, to predict dose distributions for various modalities simultaneously. To evaluate our model's performance, we used a dataset consisting of 92 plans from 28 patients undergoing accelerated partial breast irradiation and divided it into training, validation, and test subsets. Across the test subset, the multi-task model dose predictions achieved Mean Absolute Percent Error of 1.1033 ± 0.3627%. Our approach shows promise as an efficient, quantitative tool for guiding clinicians in RT modality selection. In the future, we aim to expand upon this work with a site-agnostic model to improve its applicability and generalizability in RT.

1 Introduction

Selection of the most suitable treatment modality in radiation therapy (RT) is a critical and nuanced task. The landscape of External Beam Radiation Therapy (EBRT) presents a variety of modality designs, including Cobalt-60-based systems and linear accelerators (LINACs) with C-arm, O-ring, or robotic arm design. Aside from design, modalities may differ in other aspects such as beam energy, beam collimation, mechanical degrees of freedom, and shielding. A modality's unique attributes, advantages and limitations crucially influence a patient's corresponding RT treatment plan and dose distribution.

Adding to the complexity, dose distributions may be further impacted by the choice of treatment technique, treatment planning system (TPS), and dose calculation algorithm. Advanced treatment techniques like intensity-modulated radiation therapy (IMRT), volumetric modulated arc therapy (VMAT), and helical tomotherapy (HT) can offer precise shaping of dose distributions but may increase treatment plan complexity. Additionally, the accessibility of these techniques varies across machines. This complexity emphasizes the need for a comprehensive, quantitative, and patient-centric assessment to aid in modality selection.

Addressing this need is challenging given the myriad of planning options and the time-intensive nature of plan generation. In current clinical practice, it is impractical to create multiple treatment plans for each patient as guidance for final modality selection. For example, Cilla *et al.* [1] utilized auto-planning to reduce time spent on creation of a single treatment plan but it still required 60-80 minutes. This limitation in clinical efficiency and flexibility poses a significant hurdle in providing quantitative, patient-specific data for modality selection.

In consideration of this hurdle, we believe a deep-learning based approach is desirable for its ability to rapidly predict quantitative, patient-specific data. By utilizing patient-specific data, a trained model could predict dose distributions to provide clinicians with essential quantitative information to aid in the decision-making process. Furthermore, we desired an efficient approach that could predict multiple unique dose distributions simultaneously while leveraging the shared knowledge across the broader domain.

In response to this need, we developed a deep learning-based multi-task convolutional neural network (CNN) that is capable of predicting dose distributions for various RT modalities simultaneously. Our network, inspired by the U-Net architecture, consists of a shared encoder, shared decoder, and task-specific convolutional layers to output predicted modality-specific dose distributions. We hypothesize that this model will exhibit statistically significant differences in dose prediction performance as compared to traditionally trained single-task dose prediction networks

2 Materials and Methods

Figure 1 shows a high-level schematic of our deep learning-based multi-task CNN that was developed using the PyTorch framework. The network processes input data as a 4-D NumPy array to represent a channel dimension and the three spatial dimensions. The input data has four distinct channels: computed tomography (CT) image, planning target volume (PTV) contour, a PTV distance array, and a goal array. For model training, the input data is formatted into patches of size 112x112x64 and grouped into batches of 4. Detailed pre-processing steps will be explained later in this section.

The input data is sent to the network's encoder, which is composed of four encoding layers. Each encoding layer includes two convolutional layers (kernel size 3, stride 1) and a downsampling layer. The downsampling layer concatenates the outputs from 3D max and 3D average pooling operations (both with kernel size 2) before passing them through a convolutional layer (kernel size 1, stride 2).

The convolutional layers in this network generally sequence through convolution, activation (using Convolved Concatenated ReLU or CCReLU), normalization (group normalization across five groups), and dropout (3D DropBlock with a rate of 0.05 and a maximum block size of 5x5x5). Our CCReLU activation function concatenates outputs from a positive ReLU function and a negative ReLU function, followed by convolution (kernel size 1, stride 2).

Following the final encoding layer, the network integrates two additional convolutional layers. Subsequently, the decoder contains four decoding layers. Each decoder layer consists of an upsampling layer and a

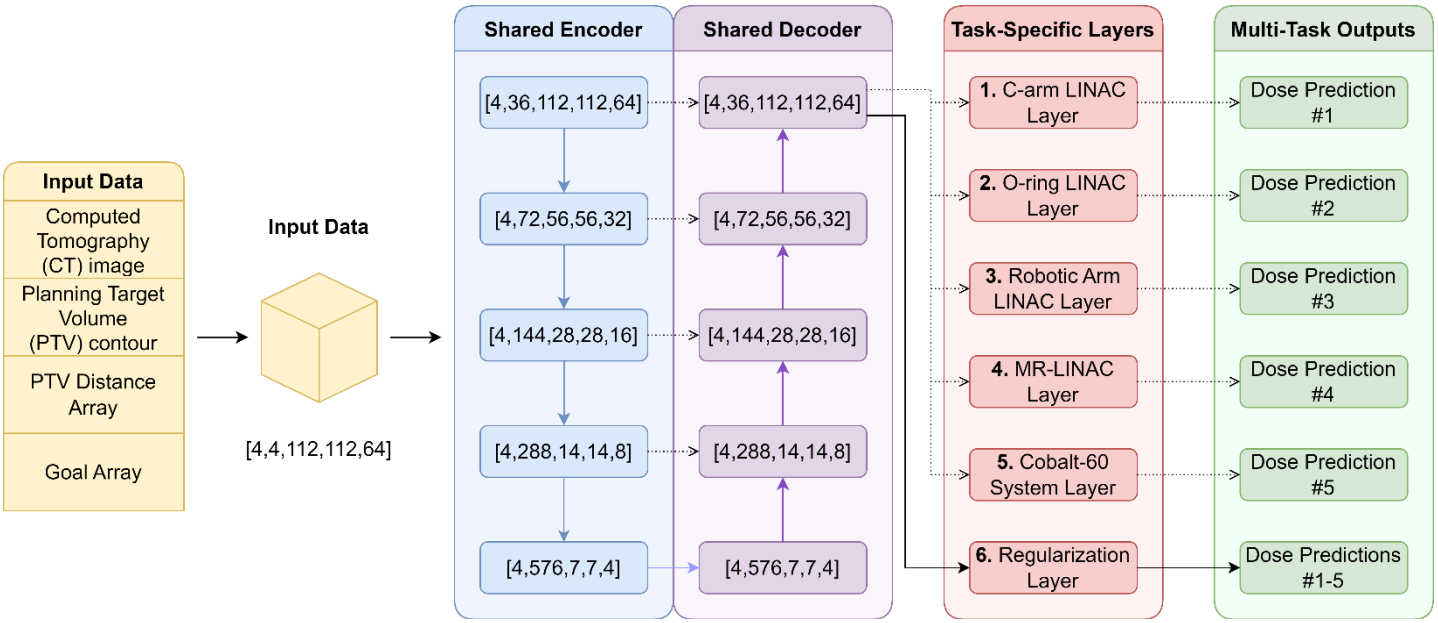


Figure 1: Our convolutional neural network architecture. Dotted lines represent the path that used skip connections.

two simultaneously utilized paths. In the upsampling layer: First, the input data undergoes nearest-neighbor interpolation, trilinear interpolation, and transposed convolution. Then, the operations' outputs are concatenated and finally passed through a convolutional layer (kernel size 1, stride 1). The output of each upsampling layer is sent to two unique paths: The first path utilizes skip connections, concatenating the upsampling layer's output with the corresponding encoder layer output, whereas the second path directly proceeds with the upsampling layer's output. The output from both of these paths is passed through two convolutional layers (kernel size 3, and stride 1).

The bifurcation of feature maps in the decoder, which results in two separate outputs, is a distinctive feature of our CNN. The output derived from the skip connection path is shared with each modality-specific convolutional layer, whereas the output that bypasses skip connections is fed into a regularization task convolutional layer. The modality-specific and regularization task convolutional layers have independent weights and all perform a single convolution and a linear activation to produce the final network outputs. The modality-specific convolutional layers, which receive input from the skip connection path, can utilize high-resolution features for dose prediction. In contrast, the regularization convolutional task layer only employs the decoder's upsampled, high-level data representation to predict doses for all trained modalities.

In overall consideration of our network, we sought to leverage shared knowledge within the broader domain by sharing feature maps across the modality-specific convolutional layers. Additionally, our regularization task aims to encourage a robust, generalized representation of the input data, and to prevent the model from overfitting to modality-specific tasks. To briefly summarize the flow of data in this network, the encoder receives an input tensor of shape [4, 4, 112, 112, 64], based on the batch size, number of channels, and extent of each spatial dimension. The output from the first encoding layer is [4, 36, 112, 112, 64], with 36 representing the initial number of feature maps. Immediately prior to the decoder, the tensor shape is [4, 576, 7, 7, 4]. The final outputs include the modality-specific task predictions, shaped as [4, 1, 112, 112, 64], and the regularization task predictions, shaped as [4, m, 112, 112, 64] where 'm' represents the number of modality-specific tasks.

To test the performance of our model, we selected a group of 28 patients that underwent Accelerated Partial Breast Irradiation (APBI). This patient cohort comprised two groups of 16 supine patients and 12 prone patients, based on their CT position. Each of the 16 supine patients had adaptive treatment plans for 4 distinct modalities: C-arm LINAC, O-ring LINAC, robotic arm LINAC, and Magnetic Resonance LINAC (MR-LINAC). Additionally, each of these supine patients had a conventional treatment plan for the C-arm LINAC. Lastly, each of the 12 prone patients had an adaptive treatment plan for a Cobalt-60-based system. In total, there were 92 treatment plans across the 28 patients. All patients were prescribed 30 Gy over 5 fractions to the PTV.

Gross Tumor Volumes (GTVs) were contoured by radiation oncologists. Two types of Clinical Target Volumes (CTVs) and PTVs were created for each patient uniformly based on our institutional guidelines: an adaptive treatment CTV and PTV, and a conventional treatment CTV and PTV. The following relevant OARs were delineated for all patients: body, left & right breasts, heart, left & right lungs, ribs, skin, and spinal canal. Additionally, dosimetric goals were identical for all patients in this study.

The first step in our data pre-processing was conversion of DICOM data to NumPy arrays. The CT array values were clipped to range from -1000 Hounsfield units (HU) to 5000 HU, and then normalized to an absolute range of 0 to 1. Each segmented structure was represented with a binary array in the same shape of the CT array, and assigned a

value of 1 for voxels where the structure is present. Dose arrays were divided by prescription dose so that their values ranged from 0 to approximately 1, with the upper limit determined by the maximum dose in the treatment plan.

A novel aspect of our data pre-processing is the creation of a PTV distance array, which has same shape as the CT array and contains values ranging from 0 to 1. For this array, voxels within the PTV contour are assigned a value of 1.0 and values for voxels outside the PTV gradually decrease towards 0 as the distance from the PTV increases.

An additional novel step in our data pre-processing is the development of a goal array. Conventionally, each OAR would be assigned to its own channel of the input data. This has two major consequences: First, a model trained with such input is inflexible, as it only has capacity for specific pre-trained OARs. Therefore, this limits the generalizability and clinical practicality of the model. Secondly, there is significant redundancy in this type of input data because many of the OAR arrays will only hold non-zero values in a relatively small percentage of its total size. In consideration of these limitations, we desired a single array that conveys all relevant OAR information so that we could minimize the number of overall input channels without sacrificing the quality of the input data. Furthermore, we desired a method that could represent a physician's dosimetric goals in the same shape of the CT so that it could be used as input data for a CNN.

Our proposed solution to these issues, which we refer to as a goal array, serves as an efficient representation of the most pertinent OARs and dosimetric goals. We established rules to fill the goal array based on each OAR's dosimetric goals, so that the goal array conveys maximum preferred dose on a voxel-by-voxel basis. For instance, one dosimetric goal requests that the contralateral breast max dose is less than 0.9 Gy, so voxels within it were assigned values equal to 0.9 Gy. For volumetric-based goals, we solved for a sub-volume of the organ that: 1) satisfies the sparing criteria, and 2) is most distal to the PTV. For the ipsilateral lung, a goal requests that the volume of it receiving 9 Gy or more is less than 10%. Our function would then solve for 90% of ipsilateral lung that is most distal to the PTV and assigning those voxels to 9 Gy. Any voxel of an OAR that was unspecified by dosimetric goals was assigned to an arbitrary value equal to 105% of prescription dose, conveying a desire to minimize high dose. Overlapping objectives were consolidated by taking the minimum value for each voxel. Next, voxels within the body that remained unassigned by any of these rules were assigned to arbitrary values equal to 95% of prescription dose, conveying that prescription dose is undesirable outside of the target. Lastly, voxels outside the body were assigned values of 0. The goal array was then divided by the prescription dose (as the dose array was).

The input data arrays were assembled by stacking the PTV array, PTV distance array, dose array, and goal array, yielding a 4-dimensional array. Next, the arrays were resized to represent $5mm^3$ voxel spacing. Then, the final input data array was created with a randomized patch selection function and a data augmentation function. The patch center was selected by sampling from a Gaussian distribution with the GTV center of mass as the mean, and the patch extent was increased in random directions until it met our specified shape of [112, 112, 64] in the y-, x-, and z-dimensions respectively. We also performed random spatial translation with a shift tolerance of 5 elements per dimension. The remaining data augmentation techniques included a 50% chance of random rotation (with an additional 25% chance of rotating either 0, 90, 180 or 270 degrees) and a 50% chance of random flip.

Our dataset was split prior to model training as follows: The data from 8 patients was withheld, totaling 24 test plans (4 plans per modality). The data from the remaining 20 patients were used during model training, which includes 56 training plans (10 plans per supine modality, 6 plans for the prone modality) and 12 validation plans (2 plans per modality).

Finally, we commenced supervised model training by using the original, human-planned dose distribution as the label. All model training in this study used the AdamW optimizer, an identical fixed random seed, and began with randomly initialized weights. Loss was calculated by comparing the label and the predicted dose. The overall loss was a sum of task-specific loss functions, and each task's loss function was a sum of mean squared error and mean squared logarithmic error. Each model trained for 5,000 epochs with a learning rate ranging from 1e-6 to 1e-4 over one cycle. The regularization task's weights were frozen for all of model training. We also froze task-specific weights if their validation loss failed to improve across 500 epochs.

In addition to the multi-task model, five single-task models were trained. Single-task models were trained with identical network architecture as our proposed multi-task CNN, except that the network only had one task that predicted a dose distribution for a single modality. Dataset splits were identical for the multi-task and single-task trainings, though single-task trainings omitted data from untrained modalities.

3 Results

Mean Absolute Percent Error	Robotic Arm LINAC	MR-LINAC	O-ring LINAC	C-arm LINAC	Cobalt-60 System
Multi-Task Model Average	1.3262 ± 0.1993%	1.4685 ± 0.2808%	1.1135 ± 0.3168%	0.8053 ± 0.3302%	0.8029 ± 0.1490%
Single-Task Model Average	1.5298 ± 0.2534%	1.5106 ± 0.2802%	1.2308 ± 0.2861%	0.8187 ± 0.3567%	1.1029 ± 0.3552%

Table 1: Model performance across the test dataset.

The trained multi-task model and single-task models were used to predict dose distributions across the test dataset. We calculated the Mean Absolute Percent Error (MAPE) for each dose prediction based on its difference

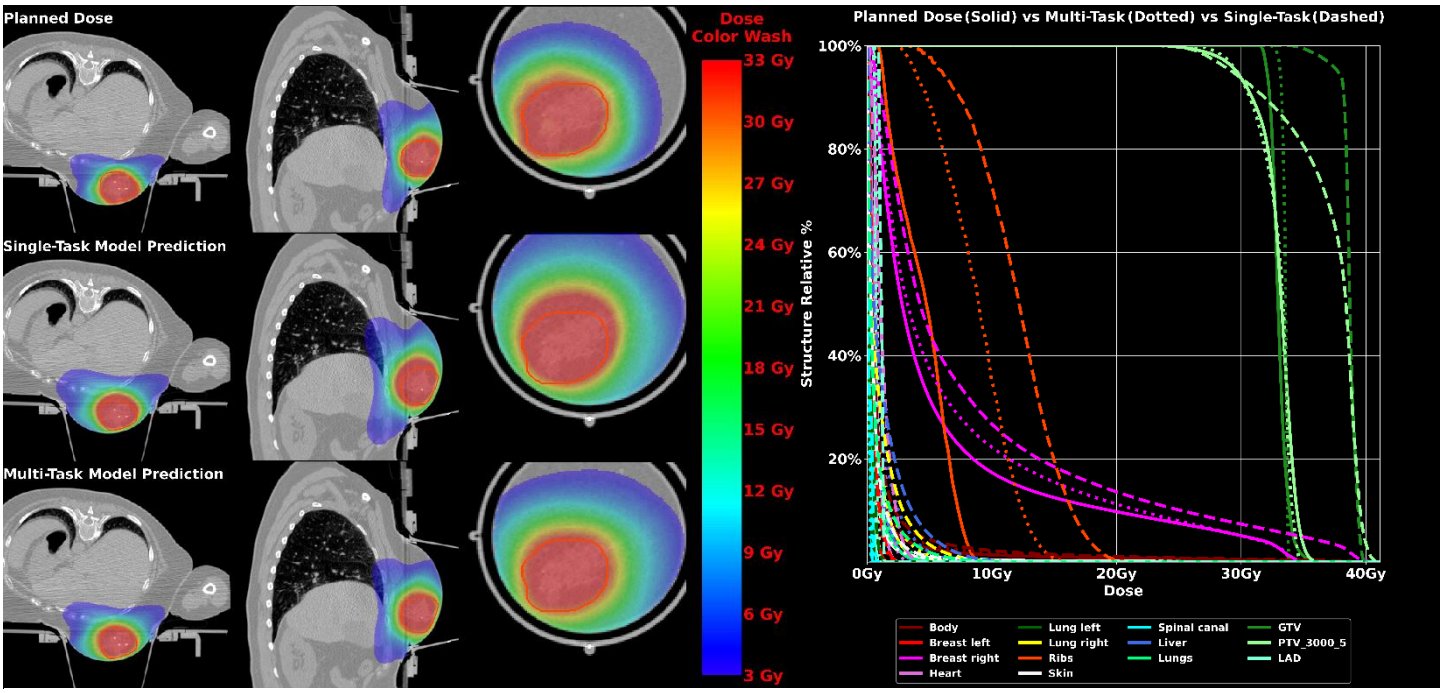


Figure 2: On the left, a patient's dose distributions for the Cobalt-60-based system are overlaid on their computed tomography image, with their planning target volume contoured in red. On the right, dose volume histogram (DVH) values are plotted for all three dose distributions. We aim for the predicted dose to align closely with the planned dose.

with the original human-planned dose distribution. Then, we calculated the average MAPE across the test dataset for the multi-task model and the single-task models as seen in Table 1. The multi-task model exhibited a lower average MAPE across each modality compared to the single-task models. Across the entire test dataset, the MAPE was 1.1033 ± 0.3627% for the multi-task model and 1.2386 ± 0.3872% for the single-task models. We performed a two-tailed Wilcoxon signed-rank test with these 24 paired MAPE values, yielding p=0.0003 and 95% confidence interval = [-0.0865, -0.0712].

In figure 2, we display an example of a prone patient's dose predictions and their corresponding dose volume histogram (DVH) for plans using the Cobalt-60-based system. The DVH demonstrates the multi-task model prediction's improved accuracy versus the single-task model in structures such as the heart, left breast, and PTV. This dosimetric difference can also be observed qualitatively as the single-task model prediction demonstrates excessive high dose outside of the PTV.

4 Discussion

Our study demonstrated superior performance from the multi-task model across the dataset as compared to a traditional single-task model in modality-specific dose prediction. We observed that the modality with the largest number of training plans had the smallest performance gap (C-arm LINAC with 20), and the largest performance gap was in the modality with the least number of training plans (Cobalt-60 system with 6). Based on this, we suspect that the multi-task CNN's shared, modality-agnostic backbone may allow for more rapid adaptation to emerging modalities than a single-task network because it trains shared layers with a larger effective sample size. Furthermore, this design may only require fine-tuning a single, modality-specific layer to add a new modality, or removal of a modality-specific layer for a modality that is phased out of clinical practice. We also suspect that this network can train a model for technique-specific dose prediction.

A trained multi-task model can facilitate quantitative decision-making for physicians and patients. If deployed as a clinical tool, it could be used to compare dosimetric outcomes across multiple modalities. Prior to the treatment planning process, involved parties could weigh the model's predicted dosimetric information alongside the treatment location, appointment availability and/or financial cost on a per-modality basis. For example, patients may consider a specific modality based on geographical proximity to their home if the dosimetric outcome is acceptable compared to a geographically distant but cutting-edge modality.

5 Conclusion

Our study developed a multi-task deep learning model for predicting dose distributions across multiple radiation therapy modalities. This flexible, modality-agnostic tool can aid clinical decision-making and resource allocation. In future works, we aim to improve the model with a larger dataset and investigate the addition of different related tasks like beam angle selection. Longer term, we aim to test site-agnostic dose prediction.

References

1. Cilla S, Ianiro A, Romano C, et al. Template-based automation of treatment planning in advanced radiotherapy: a comprehensive dosimetric and clinical evaluation. *Scientific reports.* 2020;10(1):423. DOI: 10.1038/s41598-019-56966-y.

Producing Deliverable Treatment Plans with Deep Learning: A Fully Automated Pipeline Outside the Treatment Planning System

Gerd Heilemann[1], Tufve Nyholm[2], Gregor Goldner[1], Joachim Widder[1], Dietmar Georg[1], and Peter Kuess[1]

[1]Department of Radiation Oncology, Medical University of Vienna, Vienna, Austria
[2]Department of Diagnostics and Interventions, Umeå University, Umeå, Sweden

Abstract **Background**: This study introduces a novel, fully automated one-click planning pipeline for radiation oncology, capable of operating independently of a traditional treatment planning system (TPS) without human intervention.
Methods: An integrated pipeline combining a dose prediction model and a model for generating treatment plans was developed. In this pipeline, we can produce a DICOM RT plan file directly from the segmented CT images of a prostate patient. The plans were analyzed with respect to speed and clinical acceptability based on the recalculated dose distributions.
Results: The pipeline efficiently produced 40 deliverable treatment plans with an average time of (38 ± 6) seconds. All plans met minimum PTV coverage requirements, but a few exceeded maximum dose limits for the bladder and rectum.
Conclusion: This innovative, fully automated pipeline demonstrates potential for rapid and efficient treatment planning in radiation oncology, independent of any TPS.

1 Introduction

Treatment planning in radiotherapy is inherently complex, with plan quality largely depending on the planner's skill, experience, and time investment. This complexity poses significant challenges in an era where demand for radiotherapy is soaring, and advancements in personalized radiotherapy, particularly in online treatment adaptation, are set back by traditional planning workflows.
Deep learning for treatment planning could be a game-changer, offering superior speed, accuracy, and consistency. Numerous models capable of dose prediction [1–3] or predicting the fluence [4] of Volumetric Modulated Arc Therapy (VMAT) plans have been developed. However, a common limitation of these models is their reliance on traditional treatment planning systems (TPS) to convert predicted doses or fluences into DICOM RT plans, that are executable by a LINAC.
Our group has made a significant stride in this area by developing a model that can predict the machine parameters for a VMAT plan directly [5]. Building on this, our current study introduces a pioneering pipeline that combines this model with an existing dose prediction model [3]. This integration culminates in the first fully automated pipeline capable of generating deliverable treatment plans independently of a TPS, achieving this at unprecedented speeds.
The implications of this development are profound. By circumventing traditional TPS constraints, our pipeline not only accelerates the treatment planning process but also brings us closer to the goal of personalized, adaptive radiotherapy. This advancement could potentially streamline workflows, democratize access to high quality radiotherapy, and pave the way for more dynamic treatment adaptation strategies in clinical settings.

2 Materials and Methods

2.1 Patient data

The objective was to develop a model capable of generating treatment plans that meet the latest clinical standards of our department. For this purpose, only treatment plans created using RayStation TPS were initially considered. However, to enrich the dataset, we also incorporated treatment plans developed with the previous TPS, Monaco (ELEKTA, Sweden), for data augmentation and external validation purposes. This study included patients diagnosed with localized prostate cancer, with or without seminal vesicle involvement, between January 2016 and September 2023. A comprehensive overview of the patient cohort and their respective contributions to the study is detailed in Table 1.
All patients underwent computed tomography (CT) imaging with a defined protocol of 2 mm slice thickness. The planning CT mandates a full bladder and a rectal balloon (if not contraindicated). Magnetic Resonance (MR) images were rigidly registered to the CT scans to allow a more precise delineation of the prostate, capsule penetration assessment, and seminal vesicle infiltration. The clinical target volumes (CTV) were determined based on the risk category: the prostate alone for low risk, and the prostate along with seminal vesicles for medium risk. These targets were delineated by a radiation oncologist with over 20 years of experience. Planning target volumes (PTV) were derived by adding a 7 mm margin to the CTV, reduced to 5 mm in cases where gold markers were used. For the purpose of protecting organs-at-risk (OAR) during primary localized prostate radiotherapy, structures such as the rectum, bladder, and both femoral heads were delineated. Before Febuary 2023, experienced radiotherapy technologists (RTT) or physicians performed the contouring of OARs manually. Starting from February 2023, all OAR delineations were conducted using ArtPlan (Therapanacea, France) [6].
Treatment planning was performed using RayStation (Ray-

Description	Number	Used for	Which models
Clinical RayStation plans (2022-2023)	91	Training	Dose prediction, Plan generation
Script-based RayStation plans (same)	61	Training	Plan prediction
Mimicked Monaco plans (2020-2021)	124	Training	Plan prediction
Prostate patients (random from 2016-2020)	40	External validation	Dose prediction + Plan generation

Table 1: Summary of Data Used for Model Training and Validation

Search, Stockholm, Sweden), with a prescribed dose of 60 Gy delivered in 20 fractions (3 Gy per fraction). Detailed clinical goals for planning are outlined in Table 2.

2.2 Models and Training

Architecture: The models were previously published [3, 5]. For this study the models were modified and enhanced. The dose prediction model was extended to produce dose distributions for prostate cancer cases and to incorporate a dose gradient loss to better reproduce gradients observed in the dose distribution (e.g., posterior to spare the rectum). Moreover it was modified to be compatible with the second model, the plan generation. The latter was modified to be compatible with RayStation treatment plans, which is used today for all treatment planning at our department. More detailed information on the models' architectures have been described earlier [3, 5].

Data Augmentation: In order to enhance our dataset, we implemented two distinct augmentation strategies: dose mimicking and script-based plan generation.

1) Dose Mimicking: This method expanded our pool of treatment plans by utilizing clinically approved dose distributions from Monaco plans. Specifically, we used the dose mimicking module in RayStation to generate new plans. This process involved replicating the dose distributions of existing Monaco plans in RayStation, thus producing a set of augmented treatment plans.

2) Script-Based Plan Generation: Using the scripting interface of RayStation, we developed an "offline" auto-planning pipeline within our clinic, which relies on a compiled library of prior plans. This knowledge-based method initializes the optimization process using parameters derived from anatomically similar patients. Through an iterative optimization cycle, the script aims to match the Dose-Volume Histogram (DVH) metrics with those of the anatomically similar cases. While the results of this innovative script-based model have not yet been published, the quality of the generated plans was satisfactory.

These strategies effectively almost tripled our collection of RayStation plans.

Training: The dose prediction model was trained on the 91 clinical RayStation treatment plans. The plan generation model was trained on the same clinical RayStation plans, 124 mimicked plans and an additional 61 script-based plans (see Table 1).

2.3 Fully-automated Pipeline

In our fully automatic pipeline (see Figure 1) for radiotherapy treatment planning, the process begins with preprocessing the DICOM CT image and structure files into matrices. Specifically, the CT images were stored as matrices containing the Hounsfield units in a configuration of 128x128x128 voxels. The DICOM structure data was processed to generate separate masks for the PTV, rectum, bladder, and left/right femoral heads, resulting in a composite shape of 5x128x128x128.

The **dose prediction and enhancement** phase was a critical part of the pipeline. Initially, a dose prediction model [3] was used to generate a preliminary dose distribution from the CT and OAR data in a 128x128x128 format. This dose output was then interpolated to a size of 320x128x128 to ensure compatibility with the subsequent model. The predicted dose undergoes enhancement to improve dose conformality and homogeneity within the targets, while concurrently reducing the doses to the OARs. This was achieved through the application of a series of filters designed to adjust the maximum dose in the PTV and lower the doses in sensitive areas such as the rectum and bladder.

In the **plan generation** phase, the enhanced, predicted dose, along with the CT and structure data, was fed into a second model. This plan generation model, as described in [5], was designed to predict the necessary machine parameters for VMAT plan execution. It generated leaf and jaw positions, as well as monitor units, for all 180 control points of the VMAT arc. Following this, a post-processing step was implemented where the predicted leaf and jaw positions, along with the monitor units, were checked for any potential machine violations, such as minimum leaf gaps and travel speeds between control points. Once validated, this data was stored into a DICOM plan template, which includes the meta-information from the patient's CT and structure DICOM files, ensuring compatibility.

The final stage involved dose calculation and plan review. For this, the generated plans were sent to the TPS, to provide dose calculation with a certified dose engine. The calculated doses were normalized to 57.5 Gy to 95% of the PTV to guarantee minimum coverage. To assist in the review process, the clinical goals were automatically checked, and confor-

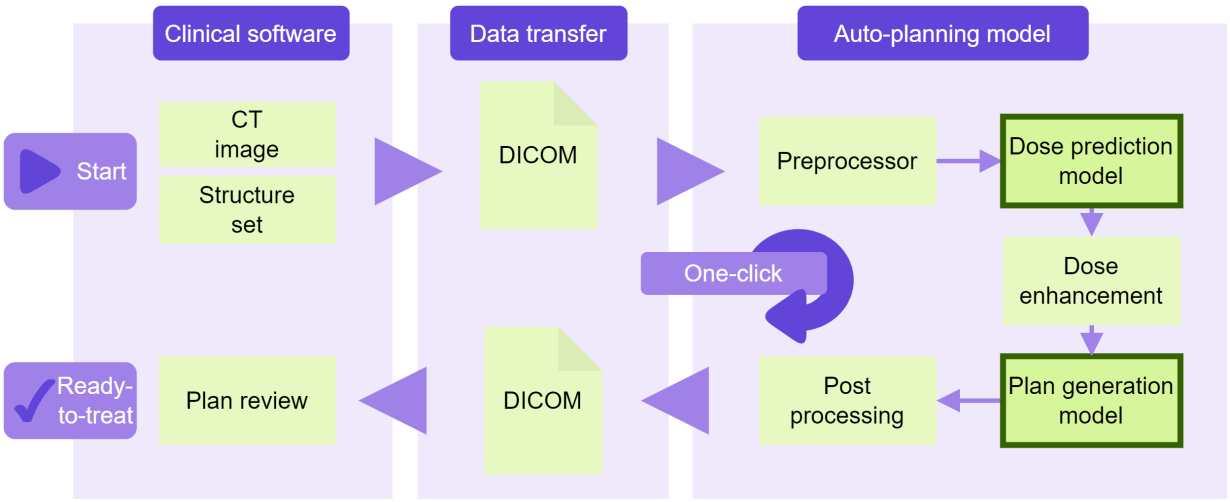


Figure 1: Flowchart of the fully-automatic treatment planning pipeline.

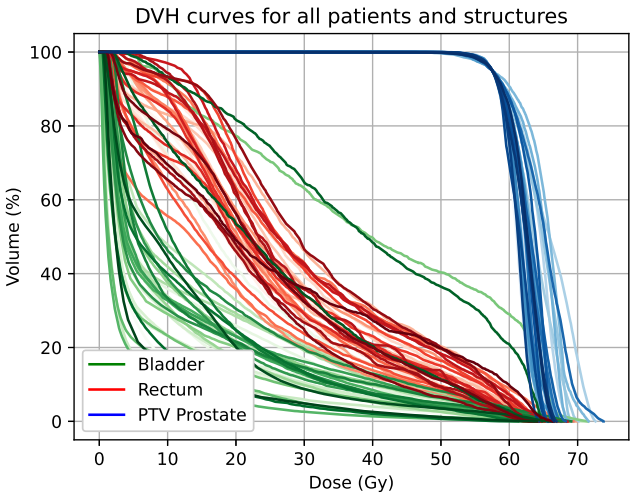


Figure 2: Dose-volume histograms for all external validation patients.

mance was indicated with positive check marks, facilitating a thorough and efficient evaluation of the plan's suitability. This pipeline was tested with an external validation set of 40 prostate patients (see Table 1).

3 Results

Figure 2 shows the DVH curves for all 40 external validation patients. The graph distinctly demonstrates the normalization of treatment plans to meet our minimum coverage criteria. Most treatment plans follow a steep dose falloff with the more than half of the plans dropping to 0% of the target volume below the permitted maximum dose of 66 Gy (21/40 treatment plans).

In the graph one can clearly see how the treatment plans were normalized with respect to satisfy the minimum coverage that we allow.

To evaluate the clinical acceptability, we report how the external validation patients performed with respect to the clinical goals (see Table 2). By design of our normalization each plan complies with the minimum required coverage of the PTV. The bladder maximum dose was exceeded in 7 out of 40 plans (17%), the low dose goals in the bladder were typically achieved. Similarly, no problems were registered in the low dose clinical goals for the rectum. Still, here the V_{20Gy} and the maximum dose were exceeded in 40% and 25% of all evaluated plans, respectively.

The average end-to-end time to produce a treatment plan from a segmented CT with no interaction necessary was 38 ± 6 seconds. The inference was done on the CPU of a regular workstation (AMD Ryzen 7 3700-X 8 Core 3.6 GHz).

ROI	Criteria	Goal	Achieved
Bladder	V_{54Gy}	20%	95%
	V_{48Gy}	40%	98%
	V_{44Gy}	50%	100%
	V_{30Gy}	80%	100%
	$D_{0.01\%}$	66Gy	83%
FemoralHead_L	V_{46Gy}	5%	100%
FemoralHead_R	V_{46Gy}	5%	100%
PTV Prostate	$D_{95\%}$	57%	100%
	$D_{0.01\%}$	66Gy	53%
Rectum	V_{48Gy}	40%	100%
	V_{40Gy}	45%	100%
	V_{30Gy}	60%	100%
	V_{20Gy}	54%	60%
	$D_{0.01\%}$	66Gy	75%

Table 2: Performance of 40 prostate plans w.r.t. clinical goals.

4 Discussion

The pipeline was successful in producing VMAT treatment plans from segmented CT images without the need for any manual interference. The pipeline produced a DICOM RT plan file that is ready to treat without the need for any further

interaction with a TPS. In this study, the TPS was utilized solely for dosimetric verification and review of the predicted plans. However, these functions could potentially be replaced by any secondary dose check and plan review software.
The demonstrated technique of augmenting the training dataset of the plan generation model (model 2) with script based replanning was successfully employed to more than double the amount of training data available. In contrast to standard augmentation techniques like flips, rotations and the addition of noise, this method produced a more diverse set of training data on the same segmented CT images. The key benefit of this technique is that it alters the dose distribution, requiring the optimizer to create a completely new MLC sequence, as opposed to minor modifications to an existing MLC sequence from the training set. We believe such techniques to be very valuable in training the plan generation model which proved to be particularly data demanding. Large improvements could be observed when increasing the dataset from only about 100 available clinical plans to almost 300 plans.
Our original plan generation model was trained on more than 500 clinical treatment plans, originally planned in Monaco TPS. As we pointed out in the study [5] the training dataset was very heterogeneous, with a melange of single fraction doses and primary and post-operative patients mixed together. It is promising, that even for roughly half the number of plans used for training we achieved a comparable level predicting RayStation-like treatment plans.
The dose prediction model (model 1) was trained only on clinically accepted cases. Dose prediction models were shown to work even for smaller sets of up to 100 training data. The augmented treatment plans which originated from the script based replanning were not used because we wanted to guarantee clinical acceptable dose predictions as the basis for the plan generation model (model 2).
The majority of treatment plans produced this way achieved the clinical goals that were used at our department. Some plans failed to comply with the high dose constraints for the rectum. However, these have been shown to be particularly demanding and most often pose the most difficult challenge in manual treatment planning too. Morevover, while the majority of the plans did not exceed the allowed maximum target dose, a notable amount of plans put some hot spots inside the PTV. These were typically within a few percent higher than the acceptable limits, but in a minority of cases significant hot spots were observed (see also Figure 2).
There are some limitations in this study. The treatment plan quality was not always on a clinically acceptable level. This will be addressed as more training data of prostate treatment plans planned in RayStation become available. The pipeline is currently only designed for local primary prostate patients. Dedicated models to produce plans for post-operative patients and for patients with lymph-node involvements are underway. Early tests to train a a generalized model that can cover all these cases at the same time was not successful.
In the current pipeline, more than 2/3 of time is spent on the preprocessing of the segmented CT images. Particularly, creating the mask matrices from the DICOM structure files is very time consuming. The actual dose and machine parameter prediction happened within a few seconds. We believe that this process can be sped up significantly to push the pipeline to produce a plan in under 20 seconds.
The development of this fully automated treatment planning pipeline represents a significant advancement in the field of radiation oncology, particularly in its potential to democratize access to high-quality radiotherapy, e.g., in low-income countries. By simplifying the planning process and reducing the dependency on specialized training and resources, this technology can potentially make advanced radiotherapy treatments more accessible and affordable.

5 Conclusion

Our team has developed what we believe to be the fastest fully automated treatment planning pipeline to date. This pipeline generates deliverable treatment plans independently of traditional TPS. In this study, we successfully integrated our two dose prediction and plan generation models, creating a seamless, fully automated workflow. This one-click solution requires no human interaction. While there were instances where the pipeline did not fully meet the clinical criteria for an acceptable treatment plan, this pioneering methodology showcases promising potential.

References

[1] D. Nguyen, T. Long, X. Jia, et al. “A feasibility study for predicting optimal radiation therapy dose distributions of prostate cancer patients from patient anatomy using deep learning”. *Scientific Reports* 9 (1 2019), pp. 1–10. DOI: 10.1038/s41598-018-37741-x.

[2] A. Babier, B. Zhang, R. Mahmood, et al. “OpenKBP: The open-access knowledge-based planning grand challenge and dataset”. *Medical Physics* 48 (2021). DOI: 10.1002/mp.14845.

[3] L. Zimmermann, E. Faustmann, C. Ramsl, et al. “Technical Note: Dose prediction for radiation therapy using feature-based losses and One Cycle Learning”. *Medical Physics* 48 (9 Sept. 2021), pp. 5562–5566. DOI: 10.1002/mp.14774.

[4] X. Li, J. Zhang, Y. Sheng, et al. “Automatic IMRT planning via static field fluence prediction (AIP-SFFP): A deep learning algorithm for real-time prostate treatment planning”. *Physics in Medicine and Biology* 65 (17 2020). DOI: 10.1088/1361-6560/aba5eb.

[5] G. Heilemann, L. Zimmermann, R. Schotola, et al. “Generating deliverable DICOM RT treatment plans for prostate VMAT by predicting MLC motion sequences with an encoder-decoder network”. *Med Phys* 50 (8 Aug. 2023), pp. 5088–5094. DOI: 10.1002/mp.16545.

[6] G. Heilemann, M. Buschmann, W. Lechner, et al. “Clinical Implementation and Evaluation of Auto-Segmentation Tools for Multi-Site Contouring in Radiotherapy”. *Physics and Imaging in Radiation Oncology* 50 (Nov. 2023), p. 100515. DOI: 10.1016/j.phro.2023.100515.

Enhancing Domain-Specific Knowledge of LLMs through Retrieval Augmented Multi-Process Prompting (RAMP)

Jace Grandinetti[1] and Rafe McBeth[1]

[1]Department of Radiation Oncology, University of Pennsylvania, Philadelphia, USA

Abstract Despite advancements in Large Language Models (LLMs), their application in specialized fields, such as medical physics, remains challenged by the need for domain-specific knowledge and avoidance of hallucinations. This study introduces a novel framework called RAMP that incorporates Retrieval Augmented Multi-process Prompting techniques to enhance the domain-specific accuracy of LLMs without the need for extensive retraining. By benchmarking on a medical physics multiple-choice exam, we found that our proposed model significantly outperforms standard LLMs and humans in performance, demonstrating improvements by as much as 68% and achieving a high score of 90% on the exam. This method reduces hallucinations and increases domain-specific performance, underscoring the potential for clinical use where accurate information is essential. This generalized technique is easily adaptable to different domains and is independent of LLM models, demonstrating its significant potential.

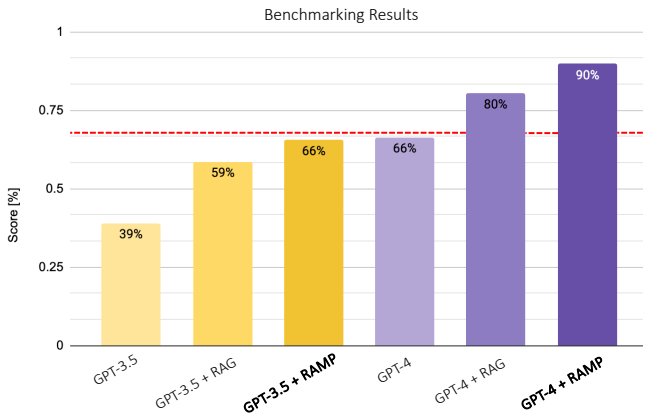


Figure 1: Bar graph illustrating the scores of different model configurations on a medical physics benchmark exam, highlighting the enhanced performance of our proposed RAMP model that achieved a score of up to 90%. The dashed red line represents the average score of 68% achieved by a cohort of medical physicists on a similar exam, serving as a comparative human benchmark [6].

1 Introduction

Natural Language Processing (NLP) models have been experiencing a rapid rise in adoption across many domains due to the Transformer-based Large Language Models (LLMs) [1]. Many leading commercial LLMs have been trained across a vast dataset of knowledge to serve as a "generalist" artificial intelligence (AI) language model, allowing for use in a broad range of specialties. With a generalist approach, these models have performed remarkably well on a range of topics, spanning from English literature to coding, to high-school STEM examinations in physics and biology. For instance, the leading state-of-the-art LLM model, GPT-4 (OpenAI, San Francisco, CA), scored $\sim 90^{th}$ percentile on the Bar Exam, $\sim 99^{th}$ percentile on the GRE Verbal, and a 75% on the Medical Knowledge Self-Assessment Program [2].
However, despite these great advancements, a disparity in knowledge for highly specialized domains (e.g., medical physics, radiation oncology) has been observed due to limited training datasets [3–5]. A study evaluating the performance of GPT-3.5 and GPT-4 on a medical physics exam found that the base models scored only 35% and 67% respectively [6], while GPT-4 similarly scored 68% on a surgical knowledge test [3]. Solutions to create or fine-tune existing models in these domains remain an open challenge. It's been demonstrated that fine-tuning and retraining methods can lead to undesired outputs, loss of prior knowledge, and overall degraded performance [7, 8]. Moreover, LLMs lack access to updated information due to a knowledge cutoff date, struggle with sourcing citations, and often experience hallucinations [9–11]. Falsified or misleading outputs of LLM models can also be especially dangerous in clinical use where medical professionals may use the outputs to guide their decisions, ultimately impacting patient care.

To address this disparity in specialized domain knowledge and to minimize hallucinations, Retrieval Augmented Generation (RAG) has rapidly emerged as a robust method for enhancing LLM models in domain-specific areas without the need for additional training [12]. RAG employs a retriever to locate documents and information corresponding to the user's inquiry, supplying this context to the transformer model during the inference process. This has been shown to significantly reduce hallucinations and provide the necessary contextual information to augment the model's knowledge base without the need for fine-tuning or retraining new models from scratch [13, 14].

Although RAG has demonstrated promising improvements, its current limitations, specifically in handling context length (i.e., number of tokens) and effectively accessing relevant information from retrieved contexts, have posed significant challenges. It has been observed that LLMs struggle with managing information when dealing with extended input contexts, which restricts their ability to handle large databases or extensive documents effectively [15]. In this study, we propose a novel framework that utilizes **R**etrieval **A**ugmented and **M**ulti-process **P**rompting (**RAMP**) techniques to improve the performance of LLM models in the domain of med-

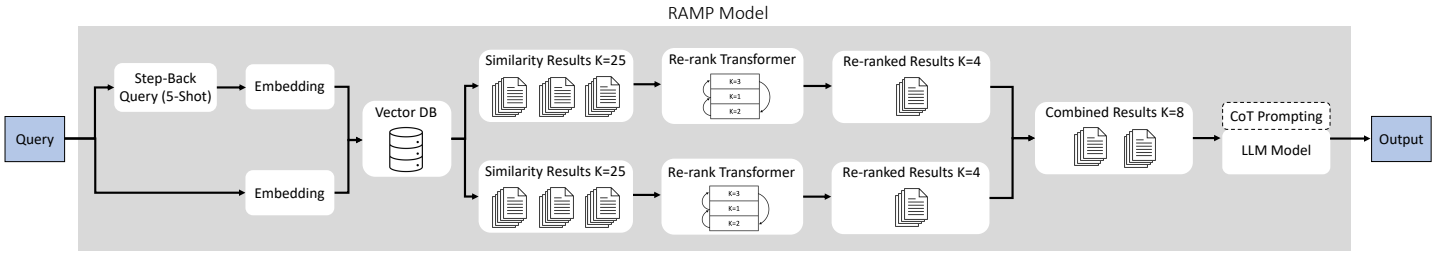


Figure 2: Figure illustrating the architecture of the proposed RAMP model which incorporates Step-Back Prompting, vector embeddings, similarity search in the vector database, a re-ranking transformer, and CoT prompting on an LLM Model for refined output generation. This workflow facilitates the accurate retrieval of pertinent documents, ensuring the model's performance is maximized across all domains in the benchmarking exam.

ical physics. We benchmarked our model using a multiple-choice medical physics exam to compare performance against standard LLM models and a reported average score of a human cohort of medical physicists who participated in a similar exam [6]. This technique is easily adaptable to different domains using various LLM models, highlighting the robust potential of this approach.

2 Methods

The proposed model uniquely integrates RAG and multi-process prompting to enhance performance, starting with Step-Back (SB) prompting to maximize the relevance of retrieved documents. These documents are then refined through a re-ranking transformer to prioritize the selection of those with the highest relevance to the input query for contextual understanding. Finally, Chain-of-Thought (CoT) prompting method is added to further improve reasoning during inferences. Detailed explanations of these individual methods are provided in Sections 2.1 - 2.4.

2.1 Retrieval Augmented Generation

We created a retrieval database by aggregating open-source reports (e.g., AAPM Task Groups and MPPGs), scholarly publications, and textbooks (e.g., IAEA Radiation Oncology Physics) pertinent to medical physics. Our objective was to develop a robust dataset that extensively covers a wide array of fundamental topics and their associated practices within medical physics, with a focus on areas like quality assurance, treatment planning, radiation safety, and medical imaging. A collection of around 50 documents was compiled and converted into text files by employing an open-source Python library [16], contributing to a vector database with a count size of about 10,000. Each document was chunked into vector sizes of ~ 500 and semantically embedded using OpenAI's embedding model. The embeddings were stored in a vector database (Pinecone, San Francisco, CA) where cosine similarity was used to retrieve vector embeddings with the highest similarity indexes.

2.2 Step-Back Prompting

Prompt engineering has proven to be a valuable technique for improving the outcomes of models by adding a layer of instructions that can help guide the LLMs as they make inferences. This has been notably demonstrated in Chain-of-Thought (CoT) prompting where multi-step reasoning is employed to solve intermediate steps from complex queries [17, 18]. A unique application of this CoT prompting is a method proposed by Google's DeepMind known as Step-Back (SB) prompting [19] which creates a secondary "step-back" query and attempts to break down and simplify a question that has multiple components. Expanding on this work, we found that we were able to apply this prompting technique to extract important keywords and rephrase queries in specialized domains, such as medical physics, to improve vector similarity searches. 5-shot prompting examples were also added to improve in-context learning to assist the model with expected outputs.

2.3 Re-Ranking Transformer

Due to the model token limits and degraded performance of LLM models with long contextual inputs, we aimed to minimize the number of retrieved documents. We achieved this by utilizing a re-ranking transformer endpoint (Cohere, Toronto, CA) specifically designed for contextually compressing search results [20]. With this approach, we were able to refine an initial set of 50 retrieved vectors to just 8 highly relevant results, significantly enhancing the speed and accuracy of our inferences. Our findings indicated diminishing returns in the value of providing more than 8 results, as pursuing additional results often led to missed information, notably when dealing with results sized into chunks of approximately 500 tokens. Furthermore, the re-ranking transformer had the added benefit of reducing token usage for API-dependent models like OpenAI's GPT, which can be cost-effective.

2.4 Proposed Model

While SB prompting enhances the retrieval of relevant information, we observed that its application in specialized

	Physics Fundamentals	Therapy Fundamentals	Treatment Planning	Radiation Safety	Imaging	Advanced Treatments	Calculation-Based	All
GPT 3.5	85.71%	36.67%	20.00%	30.00%	41.67%	38.71%	6.67%	39.06%
GPT 3.5 + RAG	85.71%	66.67%	40.00%	54.55%	50.00%	61.29%	40.00%	58.59%
GPT 3.5 + RAMP	85.71%	66.67%	53.33%	63.64%	75.00%	64.52%	46.67%	65.63%
GPT 4	**100.00%**	70.00%	46.67%	36.36%	75.00%	74.19%	53.33%	66.41%
GPT 4 + RAG	**100.00%**	**83.33%**	56.67%	81.82%	**83.33%**	90.32%	60.00%	80.47%
GPT 4 + RAMP	**100.00%**	**83.33%**	**80.00%**	**90.91%**	**83.33%**	**96.77%**	**86.67%**	**89.84%**

Table 1: Table detailing the percentage of correct responses of GPT-3.5 and GPT-4 across seven categories of the 2023 RAPHEX Therapy multiple-choice exam. The models enhanced with our proposed RAMP architecture outperformed the baseline in four out of seven sections, with notable advancements in Treatment Planning and Calculation-Based questions.

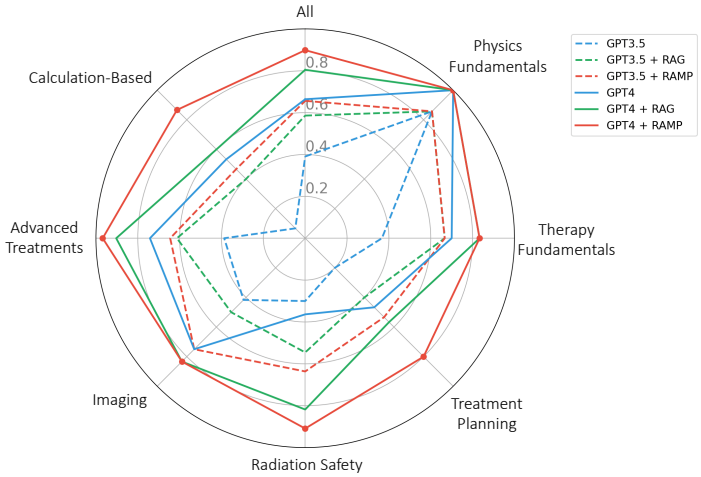


Figure 3: Radar chart depicting the benchmarking results of each model type (GPT-3.5 shown dashed, GPT-4 shown solid) by exam section. Performance percentages are charted where each vertex represents a section category and culminates to 100% at the periphery of the chart.

domains like ours occasionally led to the omission of crucial concepts and keywords necessary for accurate query responses. To mitigate this, we incorporated the original query into the vector database to yield similarity results. This integration of the outcomes from both the SB-prompted and original queries significantly improved the retrieval of pertinent documents. We employed a final CoT prompting technique at the last step to further enhance the model's performance by asking to think carefully step-by-step. It's important to note that we prompted the model to output its reasoning as it solved each problem to minimize guessing and to evaluate where the model may be making mistakes (i.e. retrieval step, incorrect calculation, hallucination, etc.). Fig. 2 illustrates the proposed model where the input query is split for embedding and SB prompting to create a second query. The top 25 similarity results are retrieved for each route which are then re-ranked through the end-point transformer before combining and then passing through the CoT prompted LLM mode. A unique advantage of this model is that it is not LLM model-dependent, so it can be easily implemented with different models.

2.5 Benchmarking Exam

To benchmark LLMs in a way that reduces subjective bias, multiple-choice questions are often used for evaluations. Since there are no existing benchmarking exams in medical physics, we compiled a set of 128 multiple-choice questions, each with 4-5 possible answers, from the RAPHEX 2023 Therapy exam. This was done to guarantee a fair and broad representation of the subject matter, recognizing that crafting a new exam from scratch could unintentionally introduce bias. We omitted questions that couldn't be answered without reference to tables or images, as our model cannot process visual information. A previous study reported that a group of medical physicists at Mayo Clinic Arizona achieved an average score of 68% on a set of RAPHEX questions, some of which were identical and others similar to those in our study. We present these findings to establish a baseline for expected human performance [6].

One drawback to benchmarking with multiple-choice questions is that the method does not inherently penalize correct guesses. Given the potential for using LLMs in clinical settings, it's imperative that the model demonstrates a reliable understanding of the correct answers and is not hallucinating or generating them by chance. To mitigate this issue, we subjected each question to the model five separate times. The model had to answer each question correctly in all five runs, otherwise no points were awarded. This strictly penalizes any incorrect answers and reduces the probability of scoring points through random guessing to less than 0.1%.

3 Results

A total of 6 model variations were benchmarked on the exam: GPT-3.5 and GPT-4, each with and without RAG to serve as a baseline comparison against our proposed RAMP model. The exam's questions were sorted into six categories, with calculation-based questions being doubly classified under their respective topic and the calculation category. The data from Table 1 and Figures 1 (summarized) and 3 (categorized) demonstrate that our proposed model ranked highest in all sections and outperformed the others in 4 out of 7 sections, with the largest improvement in treatment planning, radiation

safety, and calculation-based questions. Overall, improvements increased by approximately 68% using GPT-3.5 and 35% using GPT-4, with GPT-3.5 enhancements nearly matching the performance of the baseline GPT-4 model. Our proposed model also reliably cited the source of information for almost all correct answers, providing additional confidence in its accuracy.

4 Discussion

Significant gains were realized with GPT-3.5, likely due to its narrower training scope compared to GPT-4. Calculation improvements were also notable largely in part due to the CoT prompting and enhanced recall of relevant equations. Furthermore, we observed that a more advanced model like GPT-4 could retrieve pertinent information and process in-context learning with greater efficiency, even though GPT-3.5 showed larger improvements overall. This suggests that integrating RAG and multi-process prompting with more capable models has a synergistic impact, as the foundational knowledge is utilized effectively during the information retrieval and inference processes. Some limitations of this study include a non-comprehensive database due to restrictions on open-source content and information. We expect that we could achieve even better results by extending our dataset to encompass a broader range of knowledge. Moreover, data pre-processing steps to clean the dataset could potentially lead to enhanced precision and the improved sourcing of citations. Finally, this study benchmarked performance exclusively with a single commercial LLM (GPT). Future research would benefit from comparative analysis using models from other developers to provide a more comprehensive understanding of performance across different LLM platforms.

5 Conclusion

In conclusion, we have proposed RAMP: a novel Retrieval Augmented Multi-process Prompting technique that can be easily implemented and deployed on existing LLM models to improve performance in highly specialized domains. From a limited database of open-source materials, we were able to improve the benchmarking performance of the leading GPT-4 model from a score of 66% to 90% in the field of medical physics without the need for fine-tuning or building a new LLM model from scratch. This underscores the technique's robust potential for easy adaptation to various domains, while remaining independent of specific LLM models, demonstrating its applicability beyond medical physics.

References

[1] A. Vaswani, N. Shazeer, N. Parmar, et al. "Attention is all you need". *Advances in neural information processing systems* 30 (2017).

[2] J. Achiam, S. Adler, S. Agarwal, et al. "Gpt-4 technical report". *arXiv preprint arXiv:2303.08774* (2023).

[3] B. R. Beaulieu-Jones, M. T. Berrigan, S. Shah, et al. "Evaluating capabilities of large language models: Performance of GPT-4 on surgical knowledge assessments". *Surgery* (2024).

[4] Z. Liu, A. Zhong, Y. Li, et al. "Tailoring large language models to radiology: A preliminary approach to llm adaptation for a highly specialized domain". *International Workshop on Machine Learning in Medical Imaging*. Springer. 2023, pp. 464–473.

[5] B. van Aken. "Exploration and adaptation of large language models for specialized domains" (2023).

[6] J. Holmes, Z. Liu, L. Zhang, et al. "Evaluating large language models on a highly-specialized topic, radiation oncology physics". *arXiv preprint arXiv:2304.01938* (2023).

[7] Y. Luo, Z. Yang, F. Meng, et al. "An empirical study of catastrophic forgetting in large language models during continual fine-tuning". *arXiv preprint arXiv:2308.08747* (2023).

[8] I. J. Goodfellow, M. Mirza, D. Xiao, et al. "An empirical investigation of catastrophic forgetting in gradient-based neural networks". *arXiv preprint arXiv:1312.6211* (2013).

[9] K. Lee, O. Firat, A. Agarwal, et al. "Hallucinations in neural machine translation" (2018).

[10] Y. Bang, S. Cahyawijaya, N. Lee, et al. "A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity". *arXiv preprint arXiv:2302.04023* (2023).

[11] V. Rawte, A. Sheth, and A. Das. "A survey of hallucination in large foundation models". *arXiv preprint arXiv:2309.05922* (2023).

[12] *Retrieval Augmented Generation: Streamlining the creation of intelligent natural language processing models*.

[13] Y. Wu, J. Zhu, S. Xu, et al. "RAGTruth: A Hallucination Corpus for Developing Trustworthy Retrieval-Augmented Language Models". *arXiv preprint arXiv:2401.00396* (2023).

[14] S. Tonmoy, S. Zaman, V. Jain, et al. "A Comprehensive Survey of Hallucination Mitigation Techniques in Large Language Models". *arXiv preprint arXiv:2401.01313* (2024).

[15] N. F. Liu, K. Lin, J. Hewitt, et al. "Lost in the middle: How language models use long contexts". *arXiv preprint arXiv:2307.03172* (2023).

[16] Unstructured-IO. *Unstructured-IO/unstructured: Open source libraries and apis to build custom preprocessing pipelines for labeling, training, or production machine learning pipelines.*

[17] J. Wei, X. Wang, D. Schuurmans, et al. "Chain-of-thought prompting elicits reasoning in large language models". *Advances in Neural Information Processing Systems* 35 (2022), pp. 24824–24837.

[18] Z. Zhang, A. Zhang, M. Li, et al. "Multimodal chain-of-thought reasoning in language models". *arXiv preprint arXiv:2302.00923* (2023).

[19] H. S. Zheng, S. Mishra, X. Chen, et al. "Take a step back: evoking reasoning via abstraction in large language models". *arXiv preprint arXiv:2310.06117* (2023).

[20] H. Gilbert, M. Sandborn, D. C. Schmidt, et al. "Semantic Compression With Large Language Models". *arXiv preprint arXiv:2304.12512* (2023).

An open source PyMedPhys harness for interfacing MOSAIQ® with Claude®

Stuart J Swerdloff, PhD[1] and Simon Biggs[2]

[1]ELEKTA Pty, Wanaka, New Zealand
[2]Anthropic PBC, Wagga-Wagga, Australia

Abstract We introduce tools added to the PyMedPhys project for connecting an anonymized Oncology Information System database, MOSAIQ® , to a Large Language Model, Claude2® for natural language querying. The approach to anonymizing utilizes a custom strategy file provided to the 'pynonymizer' package, and the interface to Claude2® leverages existing capabilities in PyMedPhys and adds a Streamlit based GUI and application logic for querying via Claude2 using Retrieval Augmented Generation of the schema and data dictionary and prompting.

1 Introduction

Large Language Models have achieved significant capabilities and notoriety in the last year. Maintaining the privacy of health information is paramount, and leakage of Private Health Information to a system not within the complete control of the responsible institution must be avoided.
Oncology Information Systems are repositories for both structured and unstructured data that represent many patients' diagnosis, staging, prescription, treatment, and outcomes. Many of the clinical professionals involved in Cancer Care are pressed for the time to provide all the aspects of patient care. Use of Large Language Models to assist in extracting relevant information, correlating data, and generating first draft reports is expected to be of value to the clinicians.
Claude2® is a Large Language Model based on 'Constitutional AI', with a focus on safety, developed by Anthropic.
MOSAIQ® is an Oncology Information System that is used globally at thousands of facilities, developed by ELEKTA.
The current effort aims to place the tools for exploring the possibilities of using LLM as an AI Assistant in Oncology in the hands of cancer clinicians and clinical researchers and leave as much control over the data and prompting in their hands.

2 Materials and Methods

The software and configuration files are available in PyMedPhys [1].
The preliminary step is to anonymise the database that will later be interrogated. Anonymisation is performed using `pynonymizer` (https://pypi.org/project/pynonymizer/ or https://github.com/rwnx/pynonymizer), driven by a strategy file specific to the MOSAIQ® database. The strategy file lists the columns in tables that contain Private Health Information and the strategy for replacing the values (typically using a Faker update). There are approximately 50 columns that contain identifying information.
The database management system is Microsoft SQL Server (2016 or later).
An extract of the strategy file (in YAML format):

```
tables:
  Patient:
    columns:
      Last_Name:
        type: fake_update
        fake_type: last_name
      First_Name:
        type: fake_update
        fake_type: first_name
```

Steps taken to perform the anonymisation of the database:
The starting point was a sample database typically used for engineering development and testing that doesn't contain PHI (all patient and staff identifiers are fictitious), referred to here as "EDB". The "EDB" was backed up to file and named `pynom_msq.bak`, and then restored as `pynom_msq` in to a Microsoft SQL Server instance.
Pynominizer was run, restricted to the `ANONYMIZE_DB` step (given that the editable copy of the database was already present in SQL Server).
The simplest version of this MOSAIQ AI agent has three key components:

- Chat agent for user interaction
- SQL query creation agent which returns queries based on a schema and the conversation history
- A standard database read tool which accepts the queries provided by the SQL query creation agent executes them in read only mode, and then sends the results to the chat agent in order to converse with the user

A version of the above that makes significant improvements upon this simple version is provided within the discussion.
Given the chat agent and SQL query execution components are relatively standard, we will focus on the SQL query generation method here.
In order to direct Claude in MOSAIQ SQL query generation we provide it with both the conversation transcript and the full MOSAIQ SQL schema. In order to provide the schema to Claude in a format that it would expect we asked Claude

how it would design such a schema format. The format we used was:

```
<?xml version="1.0" encoding="UTF-8"?>
<database>
  <table name="ExampleTable">
    <column name="AnID" type="integer" />
    <column name="AnotherColumn"
        type="varchar(50)" />
  </table>
</database>
```

We then used the following SQL query to extract the desired database schema directly from the database itself, and formatted it according to Claude's preference for inclusion within the prompt:

```
SELECT TABLE_NAME, COLUMN_NAME, DATA_TYPE
FROM INFORMATION_SCHEMA.COLUMNS
```

This schema was about 5% larger than the 200k context window for Claude, so a few tables were excluded from the schema for implementation purposes.
In order to direct the SQL agent to produce appropriate queries the following key components of the prompt were utilised:

- System prompt
- Database schema
- Conversation transcript
- Reminder of key requirements
- Forced initial completion

Although there was more to it, the key parts of the system prompt were the following (shortened for brevity):

> You are an MSSQL AI agent. You respond only with valid Microsoft SQL Queries encompassed within <query> tags. You always provide 10 unique and diverse queries.
>
> Some queries that you request may not return a result, and some tables within the schema may not be relevant to your needs. So make sure that each of your queries targets different tables within the database so that you get good coverage over possible solutions and are able to return something meaningful.
>
> ...
>
> All queries assume the following database schema: schema

After the schema has been provided, the conversation transcript is then also included. After all of this a reiteration of the task is given as so:

> You respond only with valid Microsoft SQL Queries encompassed within <query> tags. All queries assume that the database is designed according to the provided schema within the <database> tags that was provided within your system prompt. You are to provide exactly 10 unique and diverse queries. Make sure that each of your queries targets different tables within the database.

And finally we put "words in the assistant's mouth" by starting their response for them, ensuring adherence to our output schema that we subsequently parse to actually undergo the queries ourselves:

> Assistant: <query> SELECT DISTINCT

The assistant will then undergo its textual completion directly after the DISTINCT keyword in order to create its first query and then subsequent queries.
As an evaluation, for this agent, the following questions were asked and the result was then recorded:

- Please find the patient ID that has had the most radiotherapy treatments
- Please find the most common cancer diagnosis that patients have had
- What is the most common number of fractions used to treat patients? Do not give a range, instead, find exactly the most common fraction number.

The agent was only given one shot in each attempt and was not given the opportunity to refine or further verify its answer. For all testing we used the publicly available version of Claude, version 2.1 via the Anthropic Python PyPI package version 0.14.0. The Mosaiq DB utilised was a testing instance that corresponds to the same schema as that used by MOSAIQ version 3.1.3.

3 Results

See tables 1, 2, and 3 for a summary of the results provided by the agent for each of the provided questions. Of note, the only question for which the majority of the answers were the correct one was for the question "Please find the most common cancer diagnosis that patients have had". An example SQL query that the agent created to answer this question was:

```
SELECT TOP 10 tpg.Description AS CancerType,
    COUNT(m.MED_ID) AS NumDiagnosed
FROM Medical m JOIN Topog tpg ON tpg.TPG_ID =
    m.TPG_ID GROUP BY tpg.Description ORDER BY
    NumDiagnosed DESC
```

Case	Count
Agent said it couldn't find a result	18
31587	2
10003	3
31215	1

Table 1: Results for "Please find the patient ID that has had the most radiotherapy treatments"

For the last question regarding fractionation, in some example responses, it becomes clear that in those particular cases Claude misunderstood the question:

> Based on the queries and analysis, the most common number of fractions used to treat patients is 1 fraction per day. This was determined by:
>
> - Calculating the average number of fractions per day across all patient treatments, which came out to be 1 fraction per day on average.
> - Analyzing the number of treatments (fractions delivered) grouped by number of fractions per day. This showed 1 fraction per day was the most common configuration, used for 54 different sequences of treatment fractions.

Another which was more successful both understood the question and was able to explain its processes:

> Based on the queries I tried, 10 fractions is the most common number used to treat patients in the database. This was determined by:
>
> - Grouping the Fractions field from the Site table and counting the number of patients for each fractions value.
> - Ordering the result by the number of patients in descending order.
> - Taking the top result which showed 10 as the most common number of fractions with 109 patients.

4 Discussion

As can be seen from the results, the agent as built above is inconsistent and while it can find answers to some classes of questions, others it fails dismally.
A second, more complicated version of the above, which aims to improve both speed and quality of the SQL query creation agent splits the SQL creation agent up into the following steps:

- Table selection agent, by first only providing the table name schema and asking it to select just a few tables to work within, the context usage is significantly reduced

Case	Count
Agent said it couldn't find a result	21
Malignant neoplasm of prostate	26
Provided diagnosis codes	1
The "NULL" diagnosis (no diagnosis entered)	3
Prostate cancer	1
Bladder cancer	1
Malignant neoplasm	1
Answered wrong question	1

Table 2: Results for "Please find the most common cancer diagnosis that patients have had"

Case	Count
Agent said it couldn't find a result	20
0 fractions	1
1 fraction	4
10 fractions	3
14 fractions	1
23 fractions	1
152 fractions	1
Answered wrong question	1

Table 3: Results for "What is the most common number of fractions used to treat patients? Do not give a range, instead, find exactly the most common fraction number"

- Query creation agent that is only given the schema for the tables selected by the table selection agent

Then, due to the significantly reduced token context utilised we would be able to run this two step query creation agent about 10 times in parallel with a voter agent selecting the best and worst queries based on the results, and then feed back the previously successful queries into subsequent multiple reruns of the SQL query creation agent in a reflexion like approach [2]. The result of which being both a reduced total time to execute (due to not having the entire schema passed) and a hypothesised improvement on completion quality. This hypothesis could be further tested by repeating the above evaluations and iterating for an optimum configuration.

5 Conclusion

We present here an open source harness for interfacing MOSAIQ® with large language models. As well as providing a means to undergo anonymisation of the underlying database. We also provide an example approach for the evaluation of the quality of the underlying model and its interfacing agent. In this case although valid and relevant SQL queries were able to be produced, in its current implementation the agent set up was only able to answer some classes of questions, and in those cases it still wasn't able to do so reliably.
Nevertheless the open source approach, the significant room

for improvement in the agent infrastructure, as well as improvements in the underlying language model being predicted through the scaling laws [3, 4] shows potential for large language model interfaces to radiation oncology information systems.

References

[1] S. Biggs, M. Jennings, S. Swerdloff, et al. "PyMedPhys: A community effort to develop an open, Python-based standard library for medical physics applications". *Journal of Open Source Software* 7.78 (2022), p. 4555.

[2] N. Shinn, F. Cassano, A. Gopinath, et al. "Reflexion: Language agents with verbal reinforcement learning". *Thirty-seventh Conference on Neural Information Processing Systems*. 2023.

[3] J. Kaplan, S. McCandlish, T. Henighan, et al. "Scaling laws for neural language models". *arXiv preprint arXiv:2001.08361* (2020).

[4] J. Hoffmann, S. Borgeaud, A. Mensch, et al. "Training compute-optimal large language models". *arXiv preprint arXiv:2203.15556* (2022).

Iterative Prompt Refinement for Radiation Oncology Symptom Extraction Using a Teacher-Student Model

Reza Khanmohammadi[1], Ahmed Ghanem[2], Kyle Verdecchia[2], Ryan Hall[2], Mohamed Elshaikh[2], Benjamin Movsas[2], Hassan Bagher-Ebadian[2,4,5], Indrin Chetty[3,5], Mohammad M. Ghassemi*[1], and Kundan Thind*[2,6]

[1]Department of Computer Science and Engineering, Michigan State University, East Lansing, USA
[2]Department of Radiation Oncology, Henry Ford Cancer Institute, Detroit, USA
[3]Department of Radiation Oncology, Cedars Sinai Medical Center, Los Angeles, USA
[4]Departments of Radiology and Osteopathic Medicine, Michigan State University, East Lansing, USA
[5]Department of Physics, Oakland University, Rochester, USA
[6]Department of Medicine, Michigan State University, East Lansing, USA

Abstract This study introduces a novel teacher-student architecture utilizing Large Language Models (LLMs) to improve prostate cancer radiotherapy symptom extraction from clinical notes. Mixtral, the student model, initially extracts symptoms, followed by GPT-4, the teacher model, which refines prompts based on Mixtral's performance. This iterative process involved 294 single symptom clinical notes across 12 symptoms, with up to 16 rounds of refinement per epoch. Results showed significant improvements in extracting symptoms from both single and multi-symptom notes. For 59 single symptom notes, accuracy increased from 0.51 to 0.71, precision from 0.52 to 0.82, recall from 0.52 to 0.72, and F1 score from 0.49 to 0.73. In 375 multi-symptom notes, accuracy rose from 0.24 to 0.43, precision from 0.6 to 0.76, recall from 0.24 to 0.43, and F1 score from 0.20 to 0.44. These results demonstrate the effectiveness of advanced prompt engineering in LLMs for radiation oncology use.

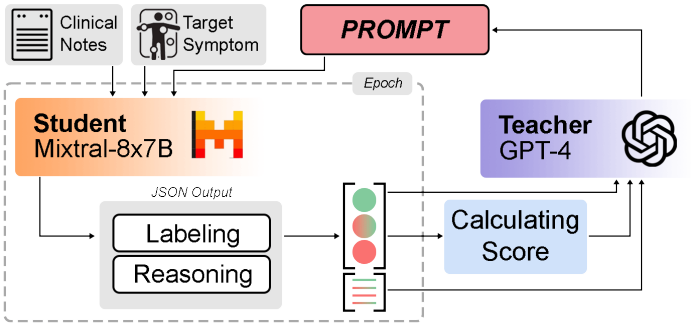


Figure 1: Schematic representation of our proposed method.

1 Introduction

The development of Large Language Models (LLMs) has marked a transformative era in the field of Artificial Intelligence (AI). These advanced models, built on the foundation of extensive data and sophisticated algorithms, have demonstrated remarkable capabilities in understanding and generating human-like text. In the medical domain, specialized versions of these models such as Med-PaLM 2 [1] and HuatuoGPT [2] have been tailored to understand and process medical terminology, patient data, and clinical notes. This focus on healthcare has enabled LLMs to assist in diagnostic processes, treatment recommendation, and patient care, thereby revolutionizing the intersection of AI and medicine [3]. The ability of these models to handle vast and complex datasets has been pivotal in uncovering insights and augmenting the decision-making process in healthcare [4]. However, the effectiveness of LLMs in specialized areas like cancer text analysis is often limited by the availability of relevant training data. Since these models learn and evolve based on the data they are exposed to [5], their performance in domains with scarce publicly available data, such as specific cancer-related texts, can be suboptimal [6, 7]. Further, a major challenge for LLMs in specialized domains is the lack of effective prompt engineering, which is essential for enhancing their performance. Prompt engineering is a process that involves iterative refinement of the input and output formats, as well as the incorporation of domain knowledge and terminology. This process can facilitate LLMs to supplement embedded information in specific domains, and thus increase their applicability and effectiveness in niche fields such as oncology [8]. Recent studies have demonstrated the efficacy of prompt engineering in improving domain-specific inference. Wang et al. [9] showed that combining prompt engineering with paraphrases of rare biomedical words can significantly enhance the performance of pre-trained models in the biomedical domain. Similarly, Alsentzer et al. [10] utilized a large language model for zero-shot interpretable phenotyping of postpartum hemorrhage from clinical notes, highlighting the model's strong performance in extracting granular concepts associated with the condition. These examples underscore the potential of prompt engineering to refine the capabilities of LLMs in specialized medical tasks, offering a promising avenue for the systematic abstraction of critical concepts in fields such as radiation oncology.

Traditional methodologies in prompt refinement have varied, with some relying on manual tuning by experts, while others explore automated techniques. However, these methods often require a balance between human input, computational resources, and specific task requirements [11]. This balance is especially crucial in medical text analysis, where precision and context-specific accuracy are paramount. Recent advancements have been observed in the domain of radi-

*Shared senior author.

ation oncology, particularly in the extraction of key concepts from clinical notes, including the identification of toxicity [12, 13]. Consequently, the development of prompt engineering methodologies for the extraction of key concepts in radiation oncology offers the potential to facilitate large-scale, systematic abstraction of critical concepts in this field.

We introduce a novel, automatic framework for prompt refinement, using a teacher-student dynamic between two LLMs: GPT-4 as the teacher and Mixtral as the student. This novel architecture is shown in Figure 1. This approach is especially pertinent in the clinical domain for key concept extraction, which is reliant on domain specific knowledge. We demonstrate the application of this architecture with prostate cancer radiotherapy toxicity symptom extraction from clinical notes.

2 Materials and Methods

The architecture employs a novel teacher-student framework designed to iteratively refine the prompts used by LLMs for prostate cancer radiotherapy toxicity symptom extraction from clinical notes. The student model, Mixtral, is tasked with the initial extraction of symptoms based on provided prompts. The teacher model, GPT-4, subsequently refines these prompts in a series of rounds and epochs based on the student's performance and reasoning.

The student model, Mixtral-8x7B, operates as the primary extractor of symptoms from the clinical notes. Given a prompt and a target symptom, it classifies the symptom as present, negated, or unknown within the notes. The model's output for each note includes the symptom classification and a reasoning for its decision.

The teacher model GPT-4, receives inputs from the student model's performance across all training notes for a given symptom. These inputs include:

- *Optimal Prompt:* The prompt that has led to the highest level of accuracy in the student's symptom classification.
- *Peak Performance:* The student model's best classification accuracy acquired on the optimal prompt.
- *Decision-making Reasoning:* Insights into the reasoning behind the student model's classification tied to the optimal prompt.
- *Failed Prompt History:* A list of prompts previously tested within the current epoch, which have yielded inferior results compared to the optimal prompt.

Using these inputs, GPT-4 generates refined prompts intended to improve the student's performance in symptom extraction. This refinement process occurs through a sequence of rounds and epochs.

Full optimization encompasses a series of epochs, each consisting of 16 rounds. In every epoch, the teacher model iteratively refines prompts to improve the student model's symptom extraction performance. At each epoch's start, the teacher is updated with the optimal prompt and peak performance metrics from the previous epoch where it has 16 opportunities (one per round) to refine the prompt and exceed the student's current best performance, measured by accuracy. If, in any of the 16 rounds, the teacher model enhances the student's performance beyond its prior peak, the epoch concludes, and the process advances to the next epoch where the teacher is granted a reset of 16 refinement opportunities. The refinement process terminates if the teacher model fails to improve the student's performance in all 16 rounds of an epoch, signaling the achievement of the student's maximum capability within the framework's parameters.

Algorithm 1 Iterative Prompt Refinement for Symptom Extraction on a target symptom

Require: Target symptom, Initial prompt, MaxEpochs, MaxRounds
Initialize Mixtral-8x7B (student model) with initial prompts
Initialize GPT-4 (teacher model)
Set best performance metric to 0
Set optimal prompt to initial prompt for the symptom
for epoch = 1 to MaxEpochs **do**
  for round = 1 to MaxRounds **do**
    Apply current optimal prompt to the student model across all clinical notes
    Gather labels and reasonings
    Calculate performance score
    if performance improves over best performance **then**
      Update best performance and optimal prompt
      Break and proceed to the next epoch
    end if
    Use the teacher model to refine the optimal prompt based on:
    - Current optimal prompt
    - Best performance metrics
    - Reasoning of the student model
    - History of failed prompts
  end for
  if no improvement in any of the 16 rounds **then**
    Terminate refinement for the symptom
  end if
end for

The data consisted of clinical notes from a cohort of prostate cancer patients treated with a radiotherapy (RT) dose of 78 Gy between 2013 and 2020. Clinical notes 6 months post-RT were extracted to focus on long term toxicities. 294 notes with presence of single symptoms were selected. Twelve common post-RT toxicity symptoms: Urgency, Cystitis, Urinary Obstruction, Dysuria, Erectile Dysfunction, Rectal Bleeding, Stricture, Nocturia, Proctitis, Hematuria, Urothelial Carcinoma, and Incontinence were selected. For the training set, we utilized 20 notes per symptom, except for Urothelial Carcinoma, where only 15 notes were available. Correspondingly, the test set comprised 5 notes for each

Symptom	*Single-symptom Notes*								*Multi-symptom Notes*							
	Accuracy		**Precision**		**Recall**		**F1**		**Accuracy**		**Precision**		**Recall**		**F1**	
	Init.	Post.	Init.	Post.	Init.	Post.	Init.	Post.	Init.	Post.	Init.	Post.	Init.	Post.	Init.	Post.
Urgency	0.80	**1.00**	0.87	**1.00**	0.80	**1.00**	0.80	**1.00**	0.37	**0.52**	0.72	**0.69**	0.37	**0.52**	0.27	**0.49**
Cystitis	0.40	**0.60**	0.80	**0.90**	0.40	**0.60**	0.53	**0.67**	0.04	**0.38**	0.27	**0.78**	0.04	**0.38**	0.06	**0.49**
Urinary Obs.	0.40	**0.80**	0.16	**0.67**	0.40	**0.80**	0.23	**0.72**	0.01	**0.12**	**0.98**	0.97	0.01	**0.12**	0.01	**0.19**
Dysuria	0.80	**0.80**	0.65	**0.70**	0.80	**0.80**	0.71	**0.73**	0.56	**0.62**	0.43	**0.67**	0.56	**0.62**	0.44	**0.61**
Erectile Dys.	**1.00**	0.80	1.00	**1.00**	**1.00**	0.80	**1.00**	0.89	**0.26**	0.21	0.38	**0.62**	**0.26**	0.21	**0.19**	0.12
Rectal Bleed.	0.00	**0.20**	0.00	**0.20**	0.00	**0.20**	0.00	**0.20**	0.16	**0.57**	0.16	**0.76**	0.17	**0.57**	0.10	**0.58**
Stricture	0.40	**1.00**	0.16	**1.00**	0.40	**1.00**	0.23	**1.00**	0.04	**0.15**	**0.96**	0.83	0.04	**0.15**	0.02	**0.20**
Nocturia	**1.00**	0.80	**1.00**	0.64	**1.00**	0.80	**1.00**	0.71	0.30	**0.34**	0.28	**0.75**	0.30	**0.34**	0.25	**0.30**
Proctitis	0.00	**0.20**	0.00	**0.80**	0.00	**0.20**	0.00	**0.32**	0.01	**0.20**	0.74	**0.77**	0.01	**0.20**	0.02	**0.32**
Hematuria	0.60	**0.60**	0.80	**1.00**	0.60	**0.60**	0.57	**0.70**	0.72	**0.76**	0.63	**0.75**	0.72	**0.76**	0.66	**0.71**
Urothelial Carc.	0.00	**1.00**	0.00	**1.00**	0.00	**1.00**	0.00	**1.00**	0.05	**0.70**	0.99	**0.99**	0.05	**0.70**	0.08	**0.81**
Incontinence	0.80	**0.80**	0.87	**0.87**	0.80	**0.80**	0.80	**0.80**	0.39	**0.53**	0.64	0.58	0.39	**0.53**	0.30	**0.51**
Average Score	0.51	**0.71**	0.52	**0.82**	0.51	**0.71**	0.49	**0.73**	0.24	**0.43**	0.60	**0.76**	0.24	**0.43**	0.20	**0.44**

Table 1: Symptom extraction test results: Initial (Init.) vs Post-refinement (Post)

symptom, with the exception of Urothelial Carcinoma, which had 4. These notes were labelled with symptom assignment of -1, 0, or 1, signifying its absence, unknown status, or presence, respectively. A separate cohort of 375 clinical notes with multiple symptom within each note were also labelled and used exclusively for validation. Multi-symptom notes ranged from 1 to 10 for the number of distinct symptom mention in each notes, with an average of 5 symptoms mentions per mutli-symptom note. This comprehensive dataset structure and labeling approach provided a robust foundation for evaluating the architecture performance, particularly in demonstrating its scalability and effectiveness in handling complex, multi-symptom clinical notes. Use of clinical notes, and the overall study was reviewed and approved by the institutional review board.

The evaluation methodology utilized accuracy as the primary metric for assessing the student model's performance during the prompt refinement process. To measure the effectiveness of these refined prompts compared to the initial ones, we employed accuracy along with weighted precision, recall, and F1 scores on the test set. These metrics provided a comprehensive view of the refined prompts' performance enhancements over the initial versions. You can find the algorithmic implementation of our method in Algorithm 1.

3 Results

Table 1, summarizes accuracy, precision, recall and F1 scores on both single symptom notes, and multi-symptom notes. Initial and post-refinement refinement results are presented for both categories, with initial performance metrics based on the the initial prompt, and post-refinement performance based on optimized prompt. Improvements across all evaluation metrics is observed. Highlighted entries show performance improvement from initial to post-refined extraction. For single symptom notes, a strong performance increase is observed across all metrics. The average improvement in accuracy from 0.51 to 0.71, precision from 0.52 to 0.82, recall from 0.52 to 0.72, and F1 from 0.49 to 0.73 is observed in single symptom notes. Noteworthy improvements include the symptom "Urgency," where accuracy soared from 0.80 to a perfect 1.00, and precision mirrored this perfect score. Similarly, "Stricture" and "Urothelial Carcinoma" witnessed accuracy enhancements from 0.40 and 0.00 to 1.00, respectively. Performance improvement across all metrics was observed for multi-symptom notes, with average improvement in accuracy from 0.24 to 0.43, precision from 0.6 to 0.76, recall from 0.24 to 0.43 and F1 from 0.20 to 0.44. The improvement was was not as strong as the single-symptom notes. Notable example of "Rectal Bleeding," showcased an increase in accuracy from 0.16 to 0.57 and an F1 score increase from 0.10 to 0.58.

Figure 2 depicts the progression of Mixtral's symptom extraction accuracy across different symptoms over successive epochs, as well as the average and standard deviation of symptom accuracy per epoch. A clear trend of improvement is observed in nearly all symptoms as the number of epochs increases, illustrating the efficacy of our prompt refinement process. Notably, symptoms such as "Urothelial Carcinoma" and "Proctitis" show a remarkable upward trajectory, indicating significant gains from the iterative approach. This demonstrates adaptive learning capabilities of this novel architecture.

4 Discussion

This novel student LLM-teacher LLM architecture demonstrates performance improvement in accuracy, precision, recall and F1 scores for prostate cancer RT symptoms. This novel architecture makes contribution in the following key areas:

- *Automatic Prompt Engineering* This automated method for refining prompts using a teacher-student combination of two LLMs, streamlines the traditionally manual process of prompt optimization.
- *Zero-Shot Learning* This framework enhances LLM performance without additional model training, showcasing the efficiency of our zero-shot learning approach.

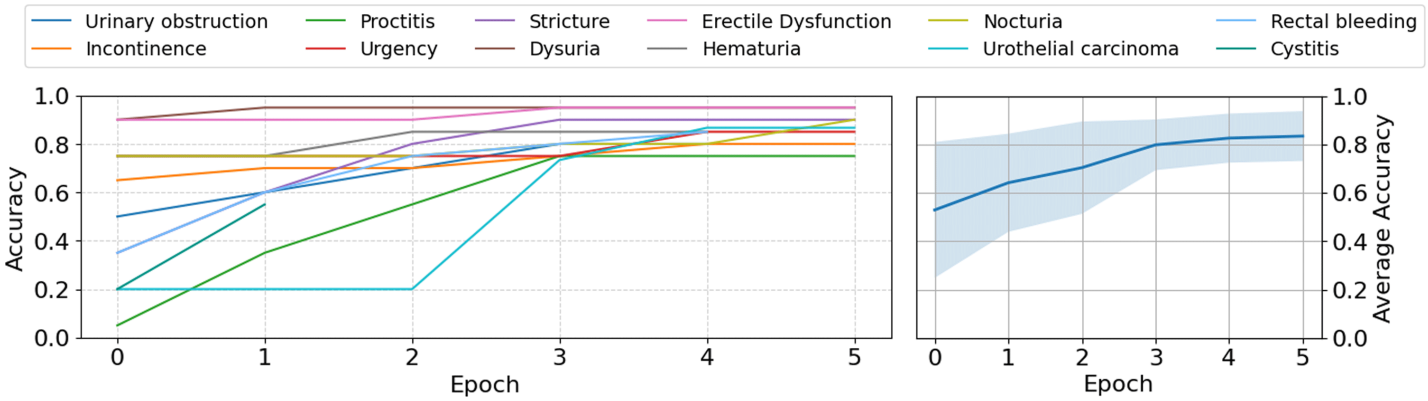


Figure 2: The evolution of Mixtral's symptom extraction accuracy per different symptoms at different epochs.

- *Local Model Improvement for Privacy* The improvement of the student model is conducted locally, ensuring sensitive clinical data, like Protected health information (PHI), is processed securely, addressing major data privacy concerns in healthcare.

The performance improvement across metrics of accuracy, precision, recall and F1 was observed for both single and multi-symptom notes. The prompt-optimization was performed on single symptom notes, and applied across test set of single symptom notes, and multi-symptom notes. This could potentially have contributed to inferior performance improvement in multi-symptom notes. Limited sample size of single symptom notes utilized for optimization may potentially have resulted in over fitting. Higher sample size for optimization, both single and multi-symptom could potentially further improve robustness, and the performance. Hyperparameter selection for both student and teacher models can also have a significant impact on the overall optimization process. Repetition studies were performed to tune hyperparameters of student model to ensure absolute consistency in response over three repetitions. As such, low `temperature` (0.2) and `top_p` (0.1) were utilized for the student model, and high `temperature` (2) and `top_p` (0.9) were used for the teacher model in this architecture. This may have limitations and further experimentation may lead to better results.

5 Conclusion

In conclusion, this novel architecture presents a pathway for self-optimized local LLM agents that can extract key concepts in medical notes with a zero shot learning approach. The complete local inference by the student model is suitable for protecting privacy of healthcare data. This system provides proof-of-concept in the research setting, and is promising for the potential of natural language processing methods to support clinical care.

References

[1] K. Singhal, T. Tu, J. Gottweis, et al. *Towards Expert-Level Medical Question Answering with Large Language Models*. 2023.

[2] H. Zhang, J. Chen, F. Jiang, et al. *HuatuoGPT, towards Taming Language Model to Be a Doctor*. 2023.

[3] M. Karabacak and K. Margetis. "Embracing Large Language Models for Medical Applications: Opportunities and Challenges". *Cureus* 15 (5 2023), e39305. DOI: 10.7759/cureus.39305.

[4] N. Eghbali, D. Siegal, C. Klochko, et al. "Automation of Protocoling Advanced MSK Examinations Using Natural Language Processing Techniques". *AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science* 2023 (June 2023), pp. 118–127.

[5] M. Mardikoraem, Z. Wang, N. Pascual, et al. "Generative models for protein sequence modeling: recent advances and future directions". *Briefings in Bioinformatics* 24.6 (Oct. 2023), bbad358. DOI: 10.1093/bib/bbad358.

[6] R. Khanmohammadi, M. M. Ghassemi, K. Verdecchia, et al. *An Introduction to Natural Language Processing Techniques and Framework for Clinical Implementation in Radiation Oncology*. 2023.

[7] D. Kamali, J. Romain, H. Liu, et al. *Using Persuasive Writing Strategies to Explain and Detect Health Misinformation*. 2023.

[8] B. Meskó. "Prompt Engineering as an Important Emerging Skill for Medical Professionals: Tutorial". *Journal of Medical Internet Research* 25 (2023), e50638. DOI: 10.2196/50638.

[9] H. Wang, C. Liu, N. Xi, et al. "Prompt Combines Paraphrase: Teaching Pre-trained Models to Understand Rare Biomedical Words". *Proceedings of the 29th International Conference on Computational Linguistics*. Ed. by N. Calzolari, C.-R. Huang, H. Kim, et al. Gyeongju, Republic of Korea: International Committee on Computational Linguistics, Oct. 2022, pp. 1422–1431.

[10] E. Alsentzer, M. J. Rasmussen, R. Fontoura, et al. "Zero-shot Interpretable Phenotyping of Postpartum Hemorrhage Using Large Language Models". *medRxiv* (2023). DOI: 10.1101/2023.05.31.23290753.

[11] B. Chen, Z. Zhang, N. Langrené, et al. *Unleashing the potential of prompt engineering in Large Language Models: a comprehensive review*. 2023.

[12] M. T. S. G. Bitterman DS Goldner E Finan S Harris D Durbin EB Hochheiser H Warner JL Mak RH. "An End-to-End Natural Language Processing System for Automatically Extracting Radiation Therapy Events From Clinical Texts". *Int J Radiat Oncol Biol Phys.* 117.3 (2023), pp. 262–273. DOI: 10.1016/j.ijrobp.2023.03.055.

[13] J. C. Hong, A. T. Fairchild, J. P. Tanksley, et al. "Natural language processing for abstraction of cancer treatment toxicities: accuracy versus human experts". *JAMIA Open* 3.4 (Dec. 2020), pp. 513–517. DOI: 10.1093/jamiaopen/ooaa064.

A BFO-compliant Ontology for Radiation Therapy (RTO)

Mark H Phillips[1], Andre Dekker[2], and Jonathan Bona[3]

[1]Department of Radiation Oncology, University of Washington, Seattle, WA, USA
[2]Department of Radiation Oncology, University of Maastricht, MAASTRO, Maastricht, NL
[3]Department of Informatics, University of Arkansas, Little Rock, AR, USA

Abstract This work describes the development of an ontology for the medical domain of radiation therapy. This effort was undertaken to provide a solid foundation upon which other standards within the field can be developed and applied. Given the uncertainties in vocabularies and concepts, the decision was made to harness the W3C ontology tools which provides numerous advantages. These include interoperability between domains, a serialization that is widely acceptable and computable, and an implementation that allows the use of Descriptive Logic. The ontology follows the guidelines of the OBO Foundry and is based on the Basic Formal Ontology. It was written in the OWL language using the software package, Protege. Examples of the utility of this ontology in the clinical environment of radiation oncology are provided.

1 Introduction

The Open Biological and Biomedical Ontology Foundry (OBO Foundry) has emerged as the home of a large number of ontologies. To be included in this repository, the creators of an ontology must follow certain principles and best practices for creating and maintaining ontologies [1]. The framework established by the OBO Foundry serves a number of useful goals. Perhaps most importantly, it guarantees interoperability between ontologies. In addition, the use of a common format reduces the difficulties in using parts or all of an ontology. Finally, the use of unique IRI's and the commitment to avoiding identical labels of ontology entities insures that ambiguities are minimized.

The topics of the OBO Foundry ontologies span a very wide range, including three "base" ontologies: Basic Formal Ontology (BFO), Core Ontology for Biology and Biomedicine (COB), and Relation Ontology (RO). These greatly help in making sure that ontologies as separate as the Gene Ontology (GO) and Ontology for General Medical Science (OGMS) can use classes and relations from each other. Many of the ontologies are basic biology-oriented, while others serve as repositories of general domain knowledge.

In this work, we described the construction of an OBO-compliant ontology for the domain of radiation therapy (RTO). The motivation for this work is several-fold. First, several technologies are emerging that require, or at least benefit from, a comprehensive knowledge representation of this domain. These include initiatives such as mCODE [2] and IHE-RO [3]. Second, the field of radiation oncology suffers from a plethora of database schemas that include several different Oncology Information Systems (OIS), electronic health record systems (EHR), insurance companies, and national and international research organizations. In the absence of a standard representation of the field, these computer systems either cannot communicate or require extensive, and ever-evolving, software to map variables and concepts. Even that environment is fraught with potential errors since it is difficult to resolve ambiguities and/or redundancies. Finally, the rise of machine learning has highlighted the need for comprehensive structured data that is intelligible across medical and biological domains. For example, the potential to find associations between genetic patterns and oncology treatment outcomes requires a common knowledge representation. Even between medical subfields, such as medical, surgical and radiation oncology, unambiguous and explicit definitions are important.

It should be noted that the RTO takes as its domain "radiation therapy" which is not identical with the larger domain of "radiation oncology". The latter includes a much more comprehensive area of knowledge that is more biology-oriented than radiation therapy, which is more oriented towards the physics/engineering and clinical side of radiation oncology. If an ontology of radiation oncology and/or oncology in general is constructed, it is our expectation that the RTO will be easily integrated, thereby reducing the effort and possible duplication and confusion. Of course, it is not possible to cleanly separate radiation oncology from radiation therapy, and RTO does attempt to make bridges when it is necessary.

2 Methods and Materials

This project was implemented by the American Association of Physicists in Medicine (AAPM) as a task group charged with building an OBO-compliant ontology. (Simultaneously, the AAPM Big Data Subcommittee also began working on a non-compliant ontology in conjunction with the mCODE project; this ontology serves as a foundation for future work harnessing the Semantic Web [4].) Task group membership consisted of AAPM volunteers. In addition, an informatician (J Bona) with experience with OBO Foundry ontologies was recruited. Work on the project was carried out using a github repository and with biweekly meetings carried out over the internet. In addition, discussions with other OBO ontology groups were conducted to resolve issues.

The ontology construction was carried out using the standards promulgated by the OBO Foundry organization. The

ontology was also built using the standards and tools of the World Wide Web (W3C) consortium [5]. This was done to insure the maximum flexibility and applicability of the ontology, as well as being able to leverage the software and tools developed under these standards.
The Ontology Development Kit (ODK) was used to initialize the ontology repository. The software package Protege [6] was used to develop the ontology using the W3C language OWL: "OWL is a computational logic-based language such that knowledge expressed in OWL can be exploited by computer programs, e.g., to verify the consistency of that knowledge or to make implicit knowledge explicit." [7]
Our ontology was built upon the Basic Formal Ontology [8, 9] and the Relation Ontology [10]. These ontologies define the taxonomy of the ontology and the relations between individuals, respectively. They are extremely general, defining entities such as "process" and relations such has "has-part", and do not refer to any concepts specific to biomedical domains. In addition, the evolving standard Core Ontology for Biology and Biomedicine [11] was integrated when possible. This arose as it became clear that there were a number of classes that were so general to the domains of biomedicine, e.g. device, that it was useful to include them in a standard ontology which all interested users could import.
The Task Group relied on the tools built into Github to provide a controlled process for the discussion of issues, the addition to and modification of the ontology as it developed, and convenient means of coordinating between members of the task group as well as interested parties.
The knowledge representation of the domain of radiation therapy can be broken up into overlapping subdomains. This is due to the large degree of overlap between disparate fields such as nuclear physics, medical practice, oncology, computer engineering, and biomedical informatics. One of the main challenges in building this ontology is incorporating ontological bridges between these different domains. Given the OBO principles of re-use of terms within the collective ontologies, this led to the need of reviewing a large number of related ontologies, and when necessary, discussing and collaborating with these groups.
Although the subdomains were not completely orthogonal, we were able to sequentially focus on the following areas.

- basic and nuclear physics;
- production and measurement of ionizing radiation;
- radiation therapy workflow;
- clinical devices specific to radiation therapy.

Finally, an important component of the development of the ontology was the recognition that clinical practice evolves, new technologies are developed and terminology is difficult to control. Therefore, this ontology was built to contain as little of current vocabularies–rather, it tries to encapsulate the most basic concepts with well-defined terms and definitions. More specific classes can be built using the Descriptive Logic methods that are an integral part of the OWL standard. It is expected that users can apply this formal ontology to the their specific needs by using axioms to build specific complex classes from the elemental concepts in RTO.

3 Results

The Radiation Therapy Ontology (RTO) was built and published under the auspices of the OBO Foundry. This version of RTO is a backbone upon which further details and refinements can be made. The classes and properties that were implemented were done so with an eye to several practical applications–namely mirroring the variables used in the Operational Ontology of Radiation Oncology [4], and providing support for a multi-institutional research project on reducing errors in radiation oncology [12].
Following the dictum that classes defined in other ontologies should be used, rather than re-defined, classes were imported from a number of different ontologies, including Basic Formal Ontology, Relation Ontology, Core Ontology for Biology and Biomedicine, Information Artifact Ontology, Radiation Biology Ontology, Chemical Entities of Biological Interest, Ontology for Biomedical Investigations, and Ontology for General Medical Science. In addition, the NCI Thesaurus is sometimes linked to definitions.
The goal of avoiding overly specific classes and vocabulary was achieved to some degree by making use of the concept of defined classes. For example, the concept of Intensity modulated radiotherapy (IMRT) means different things to different stakeholders. As a simplified example, a user could define a class `IMRT` as a subclass of `techinque action specification` and then restrict it as follows:
`has Part min 1 inverse_planned_specification`
and `has Part min 1 multileaf_collimator`
and `has Part min 5 control_point`
where the objects of these three triples are classes already defined.
The other goal was to interface with other clinical/research systems in radiation oncology. For this, we relied on Ontop [13]. Ontop is a Virtual Knowledge Graph system. It exposes the content of arbitrary relational databases as knowledge graphs. These graphs are virtual, which means that data remains in the data sources instead of being moved to another database. Ontop translates SPARQL queries (opens new window) expressed over the knowledge graphs into SQL queries executed by the relational data sources and relies on R2RML mappings. Again, as with the defined classes, this approach allows a maximum of flexibility between established data sources and emerging standards. R2RML has a default mapping scheme or it can be individually tailored as needed.

4 Discussion

The development of a radiation therapy ontology has become necessary in order for the field to be able to leverage the ad-

vances made machine learning as well as biological advances in other biomedical fields. The AAPM has been at the forefront of this effort with a two-pronged approach. First, it has worked with other stakeholders, such as American Society of Clinical Oncology and IHE-RO and HL7 FHIR, to construct an ontology based on the commercial and clinical environment [4]. The second is to work to apply the advantages and power of the Semantic Web to data curation by the creation of a formal ontology, RTO.

One of the persistent problems in making the most of the data we are continually acquiring is the set of conflicting standards, particularly in vocabulary. For example, the term "modality" is a term of art that is commonly used but not clearly defined. DICOM-RT Second Generation uses the term in multiple places with varying definitions depending on the context. The term is also defined in the Operational Ontology for Oncology. The standard defines a value set, none of which matches the DICOM standard. This leads to significant difficulties when trying to query relational databases.

There are several different approaches one can take with this ontology. If one is using Ontop, then R2RML can be used to map the RTO class to the corresponding appropriate column variable in the database. Alternatively, one can add a class specific to the application and then use the equivalence axiom to map one to the other.

As discussed earlier, the use of standards promulgated by the W3 Consortium provides a wealth of possibilities for expanded and novel uses of an ontology in radiation oncology. These include the ability to tap into the expertise of related domains, the reduction in the effort needed to broadcast data from a given domain, and the ability to leverage algorithms based on knowledge graphs and descriptive logic [14–16].

The representation of data in this approach differs from traditional relational databases. The core concept is the "triple"– subject, predicate, object. For example, `J Smith has_role patient`.

Figure 1 represents the way in which the overall concepts of the ontology are modelled for RT. In the database, each of the classes shown in Figure 1 would be instantiated by individuals, represented by an Internationalized Resource Identifier (IRI) which is unique. That is, similar to DICOM UID's, a given `radiation therapy plan specification` for a given patient would have a unique IRI which provides an unambiguous link to the resource.

An important aspect of the approach we have taken to building an ontology is the ability to use inference. The use of a defined syntax and semantics (in this case, RDF/XML) allows one to build a "reasoner" which is software that applies Desciption Logic to infer new facts from the stated ones using the axioms of the ontology. Simple examples include inferences such as "If A is a spouse of B, then B is a spouse of A", "If C is the child of D, D is the parent of C", and "If Patient X is treated on Y therapy device, and Y is a gamma-emitting device, then X is treated using gamma rays." In the latter example, the reasoner builds the latter triple using only

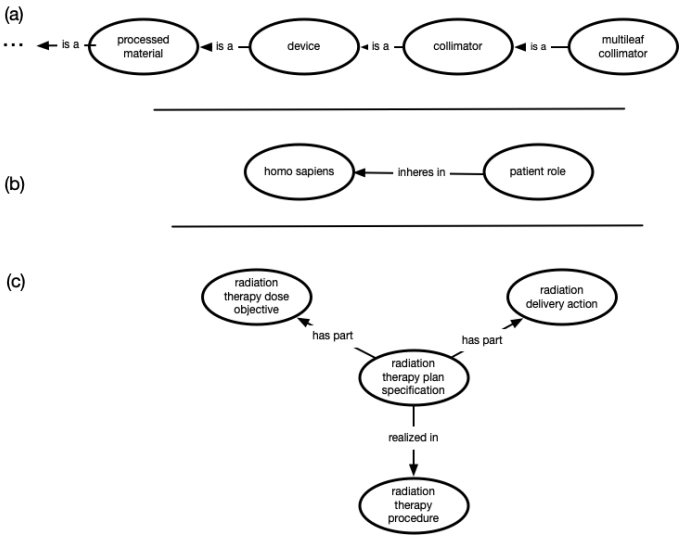


Figure 1: Elements of the ontology expressed as triples. For reasons of space, elements are rendered separately. (a) taxonomy representing a part of a linac; (b) representation of humans as patients; (c) elements that go into defining the use of radiation therapy as a treatment.

the first two without explicit construction.

As discussed earlier, this ability to use inference allows a specific application to define a useful class even though the available data does not explicitly contain that structured data. In this way, the formal RTO ontology can be used to provide building blocks for a wide range of applications and, with relatively minor additions, be tailored for the specific purposes.

The implementation of data using OWL has other advantages as well. For one, the ability of the reasoner to use inference makes it easy to do complex searches that would require difficult SQL joins. For example, searching a database for patients that met a complex set of eligibility criteria for a clinical trial can be simpler, depending on the criteria and size of the database. If one is going to store the data in the triple store approach, it can also be much easier to modify the schema as knowledge evolves and progresses relative to the corresponding effort with relational databases.

Finally, by designing RTO as an OBO-compliant ontology, progress in knowledge in related fields, such as genetics, can more easily be integrated into research and clinical practice of radiation oncology. In this way, the Semantic Web can provide computational tools to explore and understand how to better use radiation to treat cancer.

5 Conclusion

The motivation, methods and advantages of building a formal ontology following OBO Foundry priniciples for the radiation therapy domain has been described. This work was not meant to reduce the effectiveness of other approaches. Rather, it has been carried out in anticipation of its utility in the future development of computational methods utilizing a

wide range of biological and biomedical data. Using OWL and Protege is a reflection that sometimes the model is as important or more important than the data whereas the relational database is useful when data are more important than the model.

References

[1] *Open Biological and Biomedical Ontology Foundry*. https://obofoundry.org/principles/fp-000-summary.html. Accessed 20 Jan 2024.

[2] *minimal Code Oncology Data ELements*. http://hl7.org/fhir/us/mcode/ImplementationGuide/hl7.fhir.us.mcode. Accessed 20 Jan 2024.

[3] *Integrating the Healthcare Enterprise – Radiation Oncology*. https://www.ihe.net/uploadedFiles/Documents/Radiation Oncology/IHE-RO-TF-Vol1.pdf. Accessed 20 Jan 2024.

[4] C. S. Mayo, M. U. Feng, K. K. Brock, et al. "Operational Ontology for Oncology (O3): A professional society-based, multi-stakeholder, consensus-driven informatics standard supporting clinical and research use of real-world data from patients treated for cancer". *International Journal of Radiation Oncology* Biology* Physics* 117.3 (2023), pp. 533–550.

[5] *World Wide Web Consortium*. https: www.w3.org. Accessed 25 Jan 2024.

[6] M. Musen. "The Protege Project: a look back and a look forward". *Association of Computing Machinery Specific Interest Group in Artificial Intelligence* 1 (4 2015). http://protege.stanford.edu.

[7] *"Web Ontology Language (OWL)*. https://www.w3.org/OWL/. Accessed 21 Jan 2024.

[8] *Basic Formal Ontology*. https://basic-formal-ontology.org. Accessed @3 Jan 2024.

[9] *Building Ontologies with Basic Formal Ontology*. MIT Press, 2015.

[10] *OBO Foundry: Relation Ontology*. http://obofoundry.org/ontology/ro.html. Accessed 23 Jan 2024.

[11] *OBO Foundry: Core Ontology for Biology and Biomedicine*. http://obofoundry.org/ontology/cob.html. Accessed 23 Jan 2024.

[12] P. Kalendralis, S. M. Luk, R. Canters, et al. "Automatic quality assurance of radiotherapy treatment plans using Bayesian networks: A multi-institutional study". *Frontiers in Oncology* 13 (2023), p. 1099994.

[13] *Ontop*. https://ontop-vkg.org/guide/#versions. Accessed 26 Jan 2024.

[14] J. Sanjak, J. Binder, A. S. Yadaw, et al. "Clustering rare diseases within an ontology-enriched knowledge graph". *Journal of the American Medical Informatics Association* 31.1 (2024), pp. 154–164.

[15] D. Kim, J.-G. Joung, K.-A. Sohn, et al. "Knowledge boosting: a graph-based integration approach with multi-omics data and genomic knowledge for cancer clinical outcome prediction". *Journal of the American Medical Informatics Association* 22.1 (2015), pp. 109–120.

[16] S. S. Hasan, D. Rivera, X.-C. Wu, et al. "Knowledge graph-enabled cancer data analytics". *IEEE journal of biomedical and health informatics* 24.7 (2020), pp. 1952–1967.

Patient and Public Involvement and Engagement on Natural Language Processing for Real World Cancer Medical Notes

Wuraola Oyewusi[1], Eliana Vasquez Osorio[1,2], Goran Nenadic[3], Issy MacGregor[4], and Gareth Price[1,2]

[1]Division of Cancer Sciences, The University of Manchester, Manchester, UK
[2]Radiotherapy Related Research, The Christie NHS Foundation Trust, Manchester, UK
[3]Department of Computer Science, The University of Manchester, Manchester, United Kingdom
[4]Vocal, NIHR Manchester Biomedical Research Centre, Manchester Academic Health Sciences Centre, Manchester University NHS Foundation Trust,UK

Abstract Beyond clinical trials with a limited number of participants, Real-World Data (RWD) has the potential to improve inclusion in healthcare research. In deploying AI like Natural Language Processing (NLP) for healthcare, careful data and process curation is as important as the algorithms themselves. This ensures patient-focused outcomes, effective modeling, ethical use, and real-world clinical relevance. Therefore, stakeholder involvement through methods like Patient and Public Involvement and Engagement (PPIE) is crucial.
This work explores the process and findings from implementing PPIE for using NLP in cancer related free text electronic health records. The identifiable nature of this data presents unique challenges compared to working with easily anonymizable data. The PPIE event was an in-person focus group discussion with cancer survivors and caregivers. Key questions and feedback focused on data use, acceptable consent processes, research communication, and participation. Regarding consent, two-thirds of the contributors preferred a national opt-out approach, while one-third preferred project-specific consent.

1 Introduction

Real World Data (RWD) refers to all medical data collected during routine patient care rather than during clinical trials. Modern cancer care aims to be driven by evidence-based clinical decisions on treatment strategies. This evidence is often generated using clinical trials. It is well known, however, that many patient populations are under-represented in these trials, therefore, trial findings are not always generalizable. RWD can be leveraged to generate evidence in the real-world settings [1]. Unlike traditional clinical trials with narrow selection criteria, Real-World Data allows us to learn from a wider range of patients.
With the advent and adoption of Electronic Health Records, the availability of real-world data from healthcare systems has exponentially increased. In clinical oncology practice, the details of patients' cancer care including diagnosis and treatment, pathology and radiology reports, ambulatory care, medical and family history are often found in free text narrative and scanned documents [2]. It is estimated that up to 80% of electronic medical records are free text written medical notes, meaning that only 20% of RWD is structured (e.g. in tables, coded data). Structured data are algorithm-friendly, easier to anonymize, and process. The conversion of unstructured data contained within free text into a structured format can increase the scope of useful data, reduce missingness and sparseness, thereby enhancing learning from all patients. In this context, Artificial Intelligence techniques such as Natural Language Processing (NLP) are likely to play an important role.

NLP is the subfield of Artificial Intelligence that focuses on sequences like text and speech data. Intuitively, NLP gives computers the ability to understand text and spoken words in a way similar to humans by combining computational linguistics and rule-based modeling of human language with statistical, machine learning and deep learning models[3]. Working with free text written medical notes is an application of real-world data. In this work, we investigated the patient's perspective about the application of NLP to oncology medical notes. Unstructured data within free text presents unique challenges compared to structured data as even anonymization methods do not fully guarantee complete de-identification of personal identifiable information. Details from free text data can still include residual identifiers that can be cross-referenced with external sources. This means we need to work in more secure ways, and we need to check that the way we are working is acceptable and adequately communicated to patients and the public. These important questions can't be answered without engagement with the community the data is about, hence the need for PPIE.

The key aspect is to understand patients' views on the use of free text data for this type of research, their perspective on the data storage and processing methods proposed, how patients can consent to participate in research involving NLP processing of their (potentially identifiable) written medical notes, the type of information to be communicated about the project and the suitable method of delivery of this communication. To explore these aspects, we performed a Patient and Public Involvement and Engagement (PPIE) event.

Patient and Public Involvement and Engagement is the inclusion of End users such as patients, carers or any member of the community who typically will not be included in a health research design process, but their perspective is crucial because they are stakeholders in the research outcomes. PPIE is about working with ordinary people to shape and produce better research and how results are shared [4]. Public involvement in research as research being carried out 'with' or 'by' members of the public rather than 'to', 'about' or

'for' them. It is an active partnership between patients, carers and members of the public with researchers that influences and shapes research[5]. PPIE usually involves the public or specific patient groups[6] as in our case where we focus on patients living with Cancer.

2 Methods

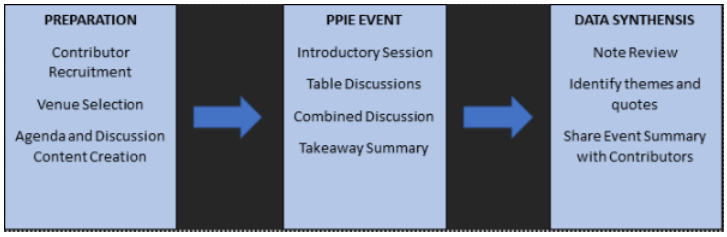


Figure 1: Process flow for a Patient and Public Involvement and Engagement for NLP on medical notes

Here we present an approach to deliver a PPIE event discussing NLP on medical notes mainly focused on three broad headings: 1) Preparation towards the Event, 2) the PPIE Event and 3) Data Synthesis of information gathered from the event. Figure 1 shows the process flow for the PPIE event.

Preparation

In preparation for the PPIE event, we leveraged community support by engaging a local cancer focus network. This community is focused on delivering services and innovative projects that bring patients, researchers, scientists and communities together to enhance health research. The Recruitment of contributors was done by collaborating with the PPIE specialist in the community. The selected Venue was central and accessible. The team then worked with the PPIE specialist to design the event, curate discussion questions and decide on method of event delivery. Table 1 shows the curated key questions

PPIE Event

The PPIE for NLP on medical notes was a 2-hour in-person focus group with 13 participants, nine contributors (Cancer survivors and Carers) and four facilitators (Researchers). Two discussion tables, each with 4-5 Contributors and 2 facilitators, were set up. The key questions were discussed at each table before a combined group discussion to consolidate the individual table discussions. To set context and prepare contributors for table discussions, a team member gave an introductory presentation discussing clinical trial limitations in cancer treatment (underrepresentation, cost, implementation), real-world data advantages (inclusivity, affordability) in addressing these limitations, the roles of structured real-world data, past PPIE findings, the risk of working with written medical notes and key event questions. Printed copies of the presentation were provided per table for reference.

The discussion questions focused on key concepts like data use, research participation, and communication about applying and implementing artificial intelligence techniques such as NLP to cancer medical notes. Contributors documented their insights on post-its stuck on to a large sheet, while dedicated facilitators took notes. The question about how people consent to this type of Research Participation was based on voting. In the UK, National opt-out is patients' choice to block their health data from being used in non-essential research while Study specific opt-out is choice to opt out of specific research project only [7]. The Contributors (9 people) voted YES or NO to National data opt-out versus Study specific opt-out

Concept	Key Questions	Discussion Points/Action
Data Use What is important to consider for our research project?	Contributors' initial views on the use of written medical notes in cancer research	What is important to patients and carers when working with potentially identifiable written medical note data?
Research Participation Is a study-specific opt-out also needed? Vote: Is the national data opt-out or study-specific opt-out most appropriate for our research project?	Contributors' perspective on National data Opt-Out vs. Project Specific Opt-Out for work with potentially identifiable written medical data	Is the national data opt-out appropriate for our research project?
Research Communication How should information be shared?	Contributors' opinions on information that needs to be disseminated about this work	What type of information is important to communicate about our research project?

Table 1: Key questions asked during the PPIE event on NLP for written cancer medical notes

Data Synthesis

After the event, facilitators used thematic analysis on the focus group notes to identify common themes and representative quotes." We then compiled a report summarizing the insights from the PPIE event, which was shared with the contributors via email.

3 Results

During the focus group discussion, contributors' questions and feedback about applying NLP to medical data were documented. Based on our findings, we have synthesized the key questions about the concepts with exemplar quotes from participant notes related to each concept.

Table 2 presents key data use concerns from participants, like patient inclusion, data accuracy/completeness, and anonymization with supporting quotes from contributors. Table 3 displays contributor feedback and comments on how consent for this type of research should be obtained. It in-

Data Use Themes	Exemplar quotes
Inclusion	How are you going to ensure inclusion of people from different ethnic backgrounds? Sometimes minority patients are excluded from medical research. "How are you going to ensure inclusivity, sometimes patients don't get a choice"
Data Accuracy	How accurate is the data? Some contributors inferred that data accuracy is more important than privacy. Lack of confidence about what is written, inferred, and recorded in medical notes, i.e., is the text data correct? Concerns over bias in source data, e.g., when written medical notes don't give the full picture. "Quality of data entered" "My GP has very detailed notes on me electronically. I am not concerned at all" "The data is as good as the recorder"
Data Completeness	How is data about patients treated at multiple centres included – would all the related and historical medical records be included in this work? Why is the focus of the study on only cancer data and not others like mental health? Will handwritten data be used in the study? "Risk of incorrectly transcribing (Handwritten) Data 'GARBAGE IN GARBAGE OUT!'" "Previous cancer history is important would be missed? = partial results/accuracy"
Data Usage	Could the structured data be used to go back to the patient to check accuracy of the data? "If something historical is discovered that may affect patient prognosis, do you get in touch to address it? – I would hope so!"
Masking/Anonymization	What is the balance between anonymization and missing important details in the dataset "What if important information is masked"

Table 2: Questions and exemplar quotes about data use from the contributors during the PPIE event on NLP for written cancer medical notes

Research Participation Themes	Exemplar quotes
Voting Result	National Opt-Out: 6/9 Project Specific Opt-Out: 3/9 One contributor mentioned that hospitals often do audits, and there is no extra need to ask for opt-out in those contexts

Table 3: Contributors' opinion on research participation consent during the PPIE event on NLP for written cancer medical notes

cludes the results of voting on if a National Opt-Out is sufficient or a Project Specific Opt-Out should be implemented and Table 4 shows specific questions and verbatim quotes related to communication about this research. Contributors asked questions about various

4 Discussion

In discussions about the application of NLP, it is far more common to find research related to the algorithms, performance and related technical concepts compared to those related to the real-world users such as those with experience with cancer and radiotherapy as in our work. Beyond the algorithms, hype and techniques, Artificial Intelligence models such as those used in Natural Language Processing are as good as the data used in training. Human experience transcends collected data and to design practical and useful solutions for people, it is valuable to involve them in the

Research Communication Themes	Exemplar quotes
Consent	Does this study include international patients? Will opting out of the research affect a person's treatment progress and standard of care? "Just say "research to improve future patient care" nothing more complex"
Inclusion	Is there provision for non-English speakers? Digital-based solution excludes some patients Why is the study covering only a few years as treatments can change – will data be out of date? Why is this study only at the Christie? "How will the research ensure inclusivity of people with protected characteristics" "2020 - 2014 Old DATA not Current"
Data privacy and security	What is the provision against data leaks in this project? What is the procedure to mitigate data exposure? Commercial use concerns "Is missed personal data screening a big risk because it's partial and less specific / individual i.e., don't worry"
Research info dissemination	Leaflets Support groups QR Codes Online Website On Screen on TVs Multilingual "Communication - Information on what the research is about What is being collected How it is going to be used What will the outcome be or goal Who will be involved. Cohort of patients? What to do if you want to be involved? / How not to be involved? Who to ask if you have questions"

Table 4: Questions and exemplar quotes about research communication from contributors during the PPIE event on NLP for written cancer medical notes

process and this is one of the cases for PPIE in healthcare research.

There is literature documenting both in person and online PPIE Events for different health research such as Vitamin B12 deficiency[1], availability of public database of statutory notification of adverse events from residential care facilities (RCFs)[6], optimal Vitamin D Status [8] and lived experiences of people with long COVID [9].To the best of authors knowledge, there is no literature documenting the PPIE process and findings for NLP in cancer medical notes. Some of the key insights about the research questions from the stakeholder perspective PPIE for NLP on real world cancer medical notes. The findings on the 3 key concepts discussed, their implications and project design adjustment made based on this event are;

Data use: Contributors prioritized balancing information availability with data privacy, not opting for absolute privacy as would have been assumed. Many contributors indicated they were also comfortable with less masking and anonymization of identifiable information if it is authorized researchers working with their data. The initial data use timeline (how old the medical notes must be before use) proposed was at least 2-year-old such that mostly non active patients were included, the contributors did not agree with this, the majority considered it a type of exclusion from beneficial research, the team is adjusting for a more inclusive data use timeline.

Research Participation: Following an introductory session, participants voted on the preferred consent model. Two thirds of the contributors favored National Opt-Out (pre-existing database of non-research objectors), while one third preferred Project-Specific Opt-Out meaning consent should be sought for the specific research. The team aligned with the majority vote on this.

Research Communication: Before proceeding to the Research Ethics Committee, to effectively communicate our work to patients and ensure appropriate and relevant information is accessible. We are going to include some of the contributors in our project information (posters, digital copies) review process. We have planned a subsequent video meeting showing what we have so far and clarifying if it is well communicated.

One limitation of this work is the number of contributors: only nine (9) from the same local area participated in the PPIE event. This suggests an opportunity for future work with a larger and more diverse group. Additionally, while the contributors were active, the short focus group format may not have captured all participant perspectives comprehensively.

5 Conclusion

In this work, we shared our process and findings from a Patient and Public Involvement and Engagement Event on applying Natural Language Processing to real-world cancer medical notes. Data Privacy remains important, but contributors prioritized balancing information availability with protecting privacy. Regarding the preferred consent model, two thirds (66.6%) favored National Opt-Out compared to one third (33.3%) for Study Specific Opt-Out. Contributors also emphasized the importance of clarity and inclusion in the research communication method. These insights have enriched our perspective and strengthened the research design as the project on applying NLP to cancer medical notes proceeds with ethical approval.

References

[1] N. Tyler, S. Giles, G. Daker-White, et al. "A patient and public involvement workshop using visual art and priority setting to provide patients with a voice to describe quality and safety concerns: Vitamin B12 deficiency and pernicious anaemia". *Health Expectations* 24.1 (2021), pp. 87–94. DOI: 10.1111/hex.13152.

[2] J. R. Mitchell, P. Szepietowski, R. Howard, et al. "A Question-and-Answer System to Extract Data From Free-Text Oncological Pathology Reports (CancerBERT Network): Development Study". *Journal of medical Internet research* 24.3 (2022), e27210. DOI: 10.2196/27210.

[3] *IBM: What is Natural Language Processing?* Retrieved January 18, 2024, from https://www.ibm.com/topics/natural-language-processing.

[4] *ReACH: Patient and public involvement and engagement (PPIE)*. Retrieved January 16, 2024, from https://www.manchester.ac.uk/discover/social-responsibility/better-health/patient-and-public-involvement-and-engagement/.

[5] *National Institute for Health Research (NIHR)*. Briefing notes for researchers - Public involvement in NHS, health and social care research. Retrieved January 16, 2024, from https://www.nihr.ac.uk/documents/briefing-notes-for-researchers-public-involvement-in-nhs-health-and-social-care-research/27371.

[6] N. McGrane, P. Dunbar, and L. M. Keyes. "To summarise the approach to and findings of the PPIE undertaken as part of a programme of secondary research with a vulnerable, hard to reach population during the COVID-19 pandemic". *Research involvement and engagement* 9.1 (2023), p. 31. DOI: 10.1186/s40900-023-00416-7.

[7] NHS Digital. *Understanding the National Data Opt-Out*. Retrieved January 31, 2024. 2024.

[8] C. Lee, N. Tanna, M. Blair, et al. "Getting underneath the skin: A community engagement event for optimal vitamin D status in an 'easily overlooked' group". *Health Expectations* 22.6 (2019), pp. 1322–1330. DOI: 10.1111/hex.12978.

[9] R. Owen, R. Ashton, L. Skipper, et al. "Long COVID quality of life and healthcare experiences in the UK: a mixed method online survey". *Quality of Life Research* 33.1 (2024), pp. 133–143. DOI: 10.1007/s11136-023-03513-y.

An AI-Driven Radiation Therapy Workflow Management Platform

Doori Rose[1], Grace Chow[1], Laura I. Cervino[1], Joseph O. Deasy[1], Sharif F. Elguindi[1], Gregory Niyazov[1], Jean M. Moran[1], Darian Pazgan-Lorenzo[1], Ellen Pinto[1], Lakshmi Santanam[1], Niral M. Shah[1], Pengpeng Zhang[1], Anyi Li[1]

[1]Department of Medical Physics, Memorial Sloan Kettering Cancer Center, New York, USA

Abstract Our team has engineered PlanQ, an innovative AI-powered radiotherapy workflow management platform, designed to streamline clinical decision-making and enhance operational efficiency within high-volume radiation therapy settings. PlanQ, with its flexible and modular design, integrates data across diverse systems and formats, offering real-time multimodal data curation, while incorporating practitioner insights as definitive inputs. By leveraging a sophisticated language model and natural language processing, PlanQ adeptly interlinks disparate clinical incidents, facilitates the examination of historical patient treatment information, and gauges re-irradiation hazards. Additionally, the platform centralizes the visualization and oversight of clinical workflows, and an AI-driven Operational Research model adeptly prioritizes planning schedules for a large team based on complexity and urgency. Empirical results indicate PlanQ's efficacy in diminishing coordination workload by 70%, reducing planning tasks by 10%, and saving an hour daily in plan verification efforts. The platform's language model demonstrates a remarkable 95% precision in synthesizing isolated clinical occurrences, ensuring immediate workflow transparency.

1. Introduction

A significant obstacle in radiotherapy emerges from the dispersion of patient data across disparate locations, providers, and systems. This issue is particularly pronounced within larger healthcare institutions, where crafting a digital patient profile often entails manual integration of data from numerous sources, leading to a process that's both time-intensive and susceptible to errors(Lustberg, et al. 2017).

Data on patients is frequently segmented across multiple systems. Simply amassing this data in one spot doesn't inherently synthesize a cohesive digital profile. The true utility of the data manifests when it's synthesized, reflecting the dynamic nature of a patient's health status that continuously evolves throughout their treatment. Adapting patient information in real-time to reflect these changes, and converting the data into a universally accepted format applicable for various radiotherapy scenarios(Field, et al. 2021), presents another layer of complexity.

Effective radiotherapy treatment relies on a well-orchestrated care team, where fluid communication, coordination, and immediate access to an updated digital patient profile are crucial for successful outcomes and error prevention. Real-time updates to the patient's status within their digital profile are vital for the care team to promptly adjust treatment plans. Yet, with frequent reliance on email for communication, tracking patient updates becomes challenging, particularly in high-volume settings, leading to communication overload and potential risks in patient care.

Our project introduces an AI-enabled system(Vandewinckele, et al. 2020) designed to streamline the management of radiotherapy workflows. This system forges connections between diverse clinical events and crafts a continuous digital narrative for each patient, enhancing clinical decision-making. Deployed in high-volume radiation therapy centers, the platform aims to improve workload efficiency and adapt to a broad spectrum of clinical scenarios.

2 Materials and Methods

The cornerstone of PlanQ platform is the harmonization, integration, and capturing of federated patient data to construct comprehensive digital profiles. We have developed a versatile connector layer capable of interfacing with a multitude of established vendor systems. By leveraging advanced distributed big data technologies, our system can interface with over 13 hospital-wide systems and seamlessly integrate common RT vendor systems such as MIM, Aria, and Mosaiq. Our integration process creates a robust foundation for each patient's digital profile by combining data from various modalities, including pathology and radiology reports and diagnostic and simulation DICOM images. PlanQ processes this data through a multi-stage ingestion approach, incorporating both nightly batch updates and real-time event-driven updates, ensuring that all care team members have access to the most current and comprehensive patient profiles from any location and device.

The pivotal step lies in curating the data to map out the interconnected patient journey, which is essential for clinical decision support. Our sophisticated curation engine employs patient matching, data normalization, cutting-edge medical language models, and knowledge graphs to analyze and vectorize multimodal patient data. This process includes extracting structured reports from unstructured simulation orders and prescriptions, using the widely embraced Fast Healthcare Interoperability Resources (FHIR) standards to store and standardize longitudinal patient information.

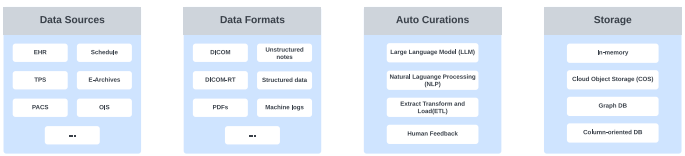


Figure 1. Highlights of PlanQ features: Its heterogeneous architecture adapts to various vendor systems and multimodality environments. The auto curation algorithms and infrastructures enable the near-real-time ingestion and curations from enterprise-wide data sources

PlanQ introduced an assistant feature that relies on expert human guidance to fine-tune the algorithm's curation process, ensuring the identification and inclusion of all relevant

patient data elements, such as historical RT treatment records, to create a high-quality digital profile. The patient events are organized into a 'subway map' visualization, providing a clear and continuous update on each treatment phase, giving the management team a comprehensive overview of clinic performance. This centralized system for documents, records, and team communication streamlines interactions and clarifies the treatment planning process.

PlaQ is built as a scalable and modular entity that accommodates data from an array of vendor systems and environments. Data curation occurs in real time, adhering to FAIR data principles, with human feedback anchoring the veracity of the curation process. Using an expansive language model and natural language processing, it connects disparate clinical events, sifts through and interprets past treatment data, and assesses the risk of re-irradiation. The platform consolidates the visualization and oversight of the clinical workflow, offering a singular, integrated viewpoint. Furthermore, we have integrated an AI-enhanced Operational Research model to refine the scheduling process, adjusting for the varying complexities and priorities within a team of 70 planners.

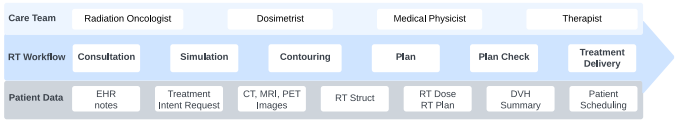


Figure 2. Disconnected patient data and communications between multiple clinical teams make RT workflow challenging. PlanQ integrates and curates data in real-time and makes the data accessible through cloud computing, improving the coordination between care teams and supporting clinical decisions

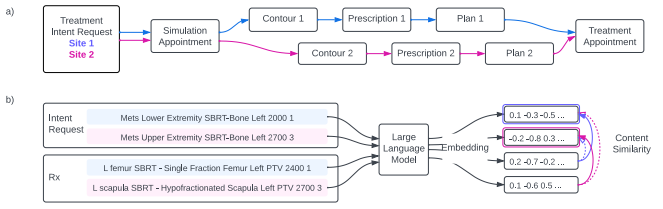


Figure 3. a) The intelligent curation module uses AI to connect the unstructured, non-standard clinical notes at the site level when multiple sites are treated at the same or different timelines.
b) We use an LLM trained in a medical knowledge base to link clinical events described with non-standardized medical terminology. We extract features from treatment intent reports and prescriptions using a sentence transformer, store the vector embeddings in an in-memory database, and calculate the vector similarities to link the evens at the site level

3 Results

Trials and implementation of PlanQ have demonstrated enhanced operational efficiency in the management of radiotherapy workflows. Notably, we have documented a 70% reduction in the effort required for treatment planning coordination, a 10% decrease in planning tasks, and a daily time saving of one hour for planning verification activities. Through PlanQ's effective data curation, retrieving historical radiation records has become more efficient, saving valuable time for physicians, planners, and physicists and reducing the risk of re-irradiation. The sophisticated Large Language Model (LLM) utilized by PlanQ exhibits a 95% accuracy level in synthesizing and associating separate clinical events, which grants immediate, real-time access to the radiotherapy planning process for all clinical teams involved. The platform's vendor-agnostic design ensures its compatibility and flexibility across a diverse array of radiotherapy workflows and clinical environments.

4 Discussion



Figure 4. Screenshots of PlanQ
a) Multiple sites treatment intents
b) Patient journey with Rx (LLM)
c) Prior irradiation records (NLP classifier)
d) DVH summary
e) Patients view from various of sources

PlanQ is an AI-enhanced, ready-for-clinic informatics platform specifically designed to streamline the complexities of managing radiotherapy workflows. PlanQ effortlessly conforms to vendor-neutral systems and multi-modality clinical settings with its versatile architecture. At its core, PlanQ constructs a comprehensive, visual timeline of each patient's treatment, anchored in the FAIR data principles, weaving together disparate data streams to aid instantaneous clinical decision-making. This integration provides a longitudinal view of the clinical workflow and detailed patient information.

PlanQ refines the accuracy of treatment plans and enhances the synergy between therapists, planners, physicists, and physicians. In essence, PlanQ not only streamlines communication but also redefines the landscape of clinical informatics, establishing a new paradigm for patient-focused radiotherapy management.

5 Conclusion

The PlanQ platform is a clinic-ready AI-driven enterprise-wide radiotherapy workflow management and clinical decision support system, showing promising results in improving efficiency and accuracy in treatment planning workflow. Its adaptability makes it a powerful and user-friendly tool to handle the complexities of radiotherapy workflows and support clinical decision-making in a variety of setups.

References

Field, Matthew, et al.
2021 Machine learning applications in radiation oncology. Physics and Imaging in Radiation Oncology 19:13-24.
Lustberg, T., et al.
2017 Big Data in radiation therapy: challenges and opportunities. Br J Radiol 90(1069):20160689.
Vandewinckele, Liesbeth, et al.
2020 Overview of artificial intelligence-based applications in radiotherapy: Recommendations for implementation and quality assurance. Radiotherapy and Oncology 153:55-66.

A mixture of hidden Markov models to predict the lymphatic spread in head and neck cancer depending on primary tumor location

Roman Ludwig[1,2], Julian Brönnimann[1,2], Yoel Pérez Haas[1,2], Esmée Lauren Looman[1,2], Panagiotis Balermpas[2], Sergi Benavente[11], Adrian Schubert[3,4,7], Dorothea Barbatei[8], Laurence Bauwens[8], Jean-Marc Hoffmann[2], Olgun Elicin[3], Matthias Dettmer[6,10], Bertrand Pouymayou[2], Roland Giger[4,5], Vincent Grégoire[8], and Jan Unkelbach[1,2]

[1]Department of Physics, University of Zurich, Zurich, Switzerland
[2]Department of Radiation Oncology, University Hospital Zurich, Zurich, Switzerland
[3]Department of Radiation Oncology, Bern University Hospital, Bern, Switzerland
[4]Department of ENT, Head & Neck Surgery, Bern University Hospital, Bern, Switzerland
[5]Head and Neck Anticancer Center, Bern University Hospital, Bern, Switzerland
[6]Institute of Tissue Medicine and Pathology, Bern University Hospital, Bern, Switzerland
[7]Department of ENT, Head & Neck Surgery, Réseau Hospitalier Neuchâtelois, Neuchâtelois, Switzerland
[8]Department of Radiation Oncology, Centre Léon Bérard, Lyon, France
[9]Department of Head and Neck Surgery, Centre Léon Bérard, Lyon, France
[10]Institute of Pathology, Klinikum Stuttgart, Stuttgart, Germany
[11]Departement of Radiation Oncology, Hospital Vall d'Hebron, Barcelona, Spain

Abstract We previously developed a mechanistic hidden Markov model (HMM) to predict the lymphatic tumor progression in oropharyngeal squamous cell carcinomas. To extend the model to other tumor subsites in the head and neck defined by ICD-10 codes, we develop a mixture model combining multiple HMMs. The mixture coefficients and the model parameters are learned via an EM-like algorithm from a large multi-centric dataset on lymph node involvement. The methodology is demonstrated for tumors in the oropharynx and oral cavity. The mixture model groups anatomically nearly subsites and yields interpretable mixture coefficients consistent with anatomical location. It allows the prediction of differences in lymph node involvement depending on tumor subsite.

1 Introduction

Head and neck squamous cell carcinomas (HNSCC) frequently spread through the lymphatic system [1, 2]. Current diagnostic imaging modalities are unable to detect microscopic lymph node metastases [3, 4]. To avoid nodal recurrences, large volumes in the neck, which are at risk of harbouring occult disease, are irradiated electively. Guidelines about which lymph node levels (LNLs) to irradiate [5] are currently not based on a patient's individual risk, but on the overall prevalence of nodal disease as reported in the literature [1, 2].

To personalize the prediction of the risk for occult disease, given a patient's individual diagnosis, we first published a large, multi-centric dataset that reports per patient which LNLs were clinically/pathologically involved [6, 7].

And subsequently, building on this work, we published an interpretable hidden Markov model (HMM), trained with this data, to predict the risk for occult nodal disease, given an individual patient's diagnosis [8].

Such a personalized risk prediction may allow clinicians to safely reduce the elective clinical target volume (CTV-N) and thus reduce side-effects that degrade the patient's quality of life without compromising treatment efficacy [9].

Here, we extend the previous work by incorporating the primary tumor location (specified as ICD-10 code) into the model of lymphatic tumor progression, focusing on tumors in the oropharynx and the oral cavity. HNSCC patients with primary tumors at different subsites show different patterns of lymphatic spread [1, 2]. So far, this could be handled by training different models for broader categories of tumor locations, e.g. oropharynx and oral cavity tumors. However, this approach does not describe differences in lymphatic spread between different subsites within the oropharynx and oral cavity. To address this issue, we present an approach using mixtures of HMMs. The intuition is that the lymphatic spread of a tumor that lies anatomically at the boarder of oropharynx and oral cavity (e.g. tumors in the palate) may be described by a mixture of different models. Tumor subsites used in this work are sketched in figure 1.

2 Materials and Methods

Each LNL $v \in \{1, 2, \ldots, V\}$ considered in our model is represented by a binary random variable X_v representing the true state of that level (0 for "healthy" and 1 for "involved"). A patient's state of lymph node involvement can be represented in a random vector $\mathbf{X} = (X_1, X_2, \ldots, X_V)$. When a patient is diagnosed with HNSCC, we only observe the clinical lymph node involvement based on imaging, which we denote as another binary random variable Y_v. To compute the personalized risk of occult disease $\mathbf{X}$, given a diagnosis $\mathbf{Y}$, we apply Bayes' law:

$$P(\mathbf{X} \mid \mathbf{Y}) = \frac{P(\mathbf{Y} \mid \mathbf{X})\, P(\mathbf{X})}{\sum_{\mathbf{X}^\star} P(\mathbf{Y} \mid \mathbf{X}^\star)\, P(\mathbf{X}^\star)} \tag{1}$$

In the above equation, the term $P(\mathbf{Y} \mid \mathbf{X})$ is given by the sensitivity and specificity of the diagnostic procedure. The term $P(\mathbf{X})$ represents the prior probability of involvement, which depends on the probability of the tumor to spread

through the lymphatic system. The main task of the HMM is to model $P(\mathbf{X})$ and the main contribution of this paper is to incorporate the primary tumor subsite into the model of $P(\mathbf{X})$.

2.1 Hidden Markov Model for Lymphatic Progression

A patient's state of lymph node involvement $\mathbf{X}[t]$ evolves over discrete time steps t. Let us enumerate all 2^V possible states, representing all combinations of LNLs. In this paper, we consider ipsilateral LNLs I, II, III and IV, which amounts to 16 possible states. The HMM is specified by a *transition matrix*:

$$\mathbf{A} = (A_{ij}) = P\left(\mathbf{X}[t+1] = \boldsymbol{\xi}_j \mid \mathbf{X}[t] = \boldsymbol{\xi}_i\right) \tag{2}$$

whose elements A_{ij} contain the conditional probabilities that a state $\mathbf{X}[t] = \boldsymbol{\xi}_i$ transitions to $\mathbf{X}[t+1] = \boldsymbol{\xi}_j$ over one time step. The transition matrix is specified and parameterised via the graphical model shown in figure 1. The red arcs in the graph of figure 1 (right panel) are associated with the probability that the primary tumor spreads directly to a LNL (parameters b_v). The blue arcs describe the spread from an upstream LNL – given it is already metastatic – to a downstream level (parameters $t_{v\to v+1}$). Now, let $\boldsymbol{\pi}$ be the *starting distribution*

$$\boldsymbol{\pi} = (\pi_i) = P(\mathbf{X}[0] = \boldsymbol{\xi}_i) \tag{3}$$

denoting the probability to start in state $\boldsymbol{\xi}_i$ at time step 0. Assuming that every patient started with all LNLs being healthy, we set π_i to zero for all states except the completly healthy state $\boldsymbol{\xi} = (0,0,0,0)$, which has probability one.
Using the quantities introduced so far, the probability $P(\mathbf{X}[t] = \boldsymbol{\xi}_i)$ to be in state $\boldsymbol{\theta}_i$ in time step t can now be conveniently expressed as a matrix product:

$$P(\mathbf{X}[t] = \boldsymbol{\xi}_i) = (\boldsymbol{\pi} \cdot \mathbf{A}^t)_i \tag{4}$$

This evolution implicitly marginalizes over all possible paths to arrive at state $\boldsymbol{\xi}_i$ after t time-steps. Additionally, we must marginalize over the unknown time of diagnosis using a time-prior $P_T(t)$. This finally defines the probability distribution over all states of lymph node involvement used in equation 1.

$$P(\mathbf{X} = \boldsymbol{\xi}_i \mid \boldsymbol{\theta}) = \sum_{t=0}^{t_{\max}} P_T(t) (\boldsymbol{\pi} \cdot \mathbf{A}^t)_i \tag{5}$$

where $\boldsymbol{\theta} = \{b_v, t_{v\to v+1}\}$ denotes the set of all model parameters (7 in our case). Fortunately, the exact length and shape of this distribution has little impact as previously shown [8]. We set $t_{\max} = 10$ and $P_{\text{early}}(t)$ to a binomial distribution with parameter 0.3. Further details on the HMM can be found in Ludwig, Pouymayou, Balermpas, et al. [8] and Ludwig [10].

2.2 Mixture of HMMs

We now assume that primary tumors at different subsites have different patterns of lymphatic spread, corresponding to different model parameters $\boldsymbol{\theta}$. Training a separate model for every possible subsite (ICD-10 code) would require a sufficiently large dataset for every tumor site. However, anatomically nearby locations are expected to show very similar patterns of LNL involvement. Therefore, we consider a mixture model. Let us assume that we have a dataset $\mathbf{D}$ that is specified via the number of patients N_{is} that were diagnosed in LNL involvement state i and had a primary tutor in subsite s. Let us further assume that we want to describe this dataset using a mixture of M HMMs, each with a different set of model parameters $\boldsymbol{\theta_m}$. As the generative model of the data, we assume that a patient with subsite s is generated with probability π_{sm} from model m. The likelihood of the dataset can then be written as

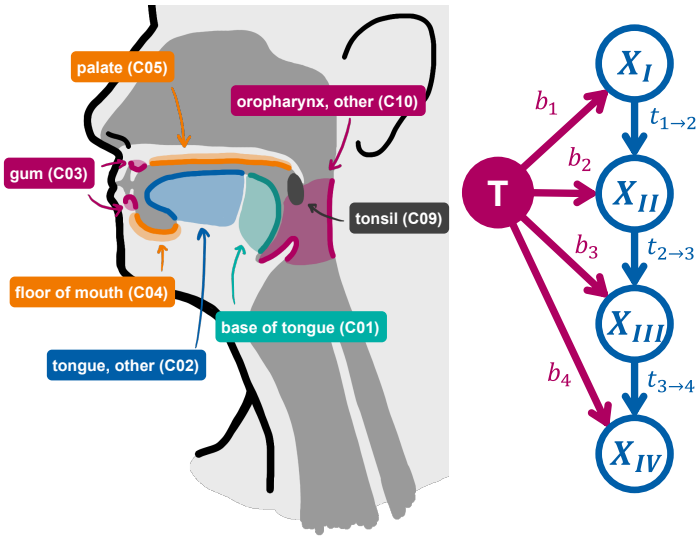


Figure 1: On the left: Anatomical sketch of the tumor subsites and corresponding ICD-10 codes considered in this work. The subsite "other parts of mouth" (C06) was not drawn. On the right: Parametrized graphical model of the lymphatic network considering four LNLs. Blue nodes represent the hidden states of LNLs X_v, while the red one is the tumor. Arcs represent possible routes of metastatic spread, associated with a probability.

$$P(\mathbf{D} \mid \boldsymbol{\theta}, \boldsymbol{\pi}) = \prod_s \prod_i \left[\sum_{m=1}^{M} \pi_{sm} P_m(\mathbf{X} = \boldsymbol{\xi}_i \mid \boldsymbol{\theta}_m) \right]^{N_{is}} \tag{6}$$

We now have two types of parameters, the probabilities of tumor spread for the different models, $\boldsymbol{\theta_m}$, and the mixing coefficients π_{sm}. Assuming a uniform prior in the interval $[0,1]$ for all parameters, the posterior distribution over the parameters $P(\boldsymbol{\theta}, \boldsymbol{\pi} \mid \mathbf{D})$ is given by the likelihood in equation 6 except for a normalisation constant. In this work, we use Markov chain Monte Carlo sampling (MCMC) via the `emcee` Python package [11] to sample model parameters from the posterior distribution. However, $P(\boldsymbol{\theta}, \boldsymbol{\pi} \mid \mathbf{D})$ itself is a multi-model distribution because one can permute the different models. To address this problem, we revert to an *expectation-maximization (EM)* algorithm where we iterate two steps until convergence of the mixing coefficients. In the E-step, we sample model parameters $\boldsymbol{\theta_m}$ using MCMC for given mixing

coefficients π_{sm}. In the M-step, we maximize the likelihood with respect to the mixing coefficients for given samples of $\boldsymbol{\theta_m}$

2.3 Multi-Centric Data

For the analyses in this work, we used five datasets from four different institutions, resulting in 1242 patients in total.

1. 287 oropharyngeal patients from the University of Zurich in Switzerland
2. 263 oropharyngeal patients from the Centre Léon Bérard in France
3. 289 oropharyngeal and oral cavity patients from the Inselspital Bern in Switzerland
4. 239 oropharyngeal and oral cavity patients from the Centre Léon Bérard in France
5. 162 oropharyngeal patients from the Hospital Vall d'Hebron in Spain (not yet public)

The data sets 1-4 are publicly available in the form of CSV tables [7, 12] and may be interactively explored in our **Ly**mphatic **Pro**gression e**X**plorer LyProX. For each patient, the primary tumor subsite is reported (among other patient and tumor characteristics) and each individual LNL is reported as metastatic or healthy, according to the available diagnostic modalities (in part pathology after neck dissection, otherwise clinical involvement).
In figure 2, we plot the prevalence of involvement in the four ipsilateral LNLs I, II, III, and IV stratified by the primary tumor's subsite. The figure illustrates the variations in LNL involvement between subsites within the oral cavity and oropharynx categories. The number of patients for each subsite is indicated in figure 3.

3 Results

We demonstrate the methodology for a mixture model with $M = 2$ components, considering the ipsilateral involvement of LNLs I, II, III, and IV and the primary tumor subsites

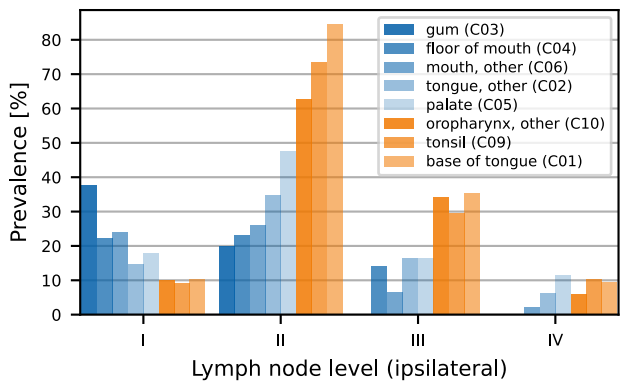


Figure 2: Prevalence of ipsilateral LNL involvement stratified by subsite. The subsites are sorted in ascending order by their prevalence of involvement in LNL II. Oral cavity subsites are plotted in shades of blue, oropharynx subsites in shades of orange.

shown in figure 2. In figure 3 we show the resulting mixture coefficients π_{sm}. The interpretation of this result is as follows: tumors of the base of tongue (C01) are fully described by component A, and tumors of the gum (C03) are fully described by component B. These two subsites are the most distinct regarding the involvement of LNLs I and II, and the result is thus intuitive. Component A may be interpreted as a model for oropharynx-like tumor spread, and component B as a model for oral cavity-like tumor spread. All other subsites are described as mixtures. tumors in the tonsil (C09) have LNL involvement similar to base of tongue tumors and are mostly assigned to component A. Instead, tumors of the palate (C05) are to similar degree assigned to components A and B, which is consistent with the anatomical location and the observation that the LNL involvement is in between oropharynx and oral cavity-type patterns (figure 2).

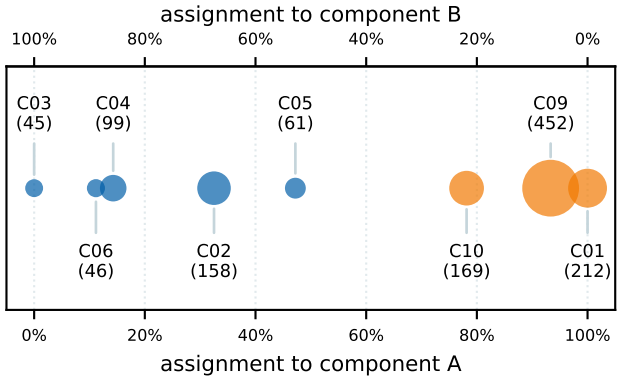


Figure 3: Assignment of each subsite to each of the two components. The further left a subsite, the more it is assigned to component A, the further right, the more to component B. The size of the marker (area) corresponds to the number of patients in the subsite.

The figure 4 illustrates the model's predictions of the overall prevalence of lymph node involvement in LNLs I, II and III (filled histograms) for selected subsites, obtained by summing the probabilities of states where the respecive level is involved. The mixture model is compared to two independent HMM models trained for oral cavity and oropharynx (by pooling the respective subsites). The mixture model and the independent oropharynx model perform similarly for tonsil tumors (C09), which is the largest patient group, dominating the independent oropharynx model and the component B in the mixture model. However, the mixture model better predicts the higher prevalence in level II for base of tongue and palate tumors.

4 Discussion

We have previously developed a model of lymphatic progression of HNSCC using HMM, which allows us to estimate the probability of occult lymph node metastases in clinically negative LNLs. Mixture models are a suitable method to incorporate the primary tumor location into the model, which allows us to account for differences in lymph node involvement for different subsites. Future work will extend the work

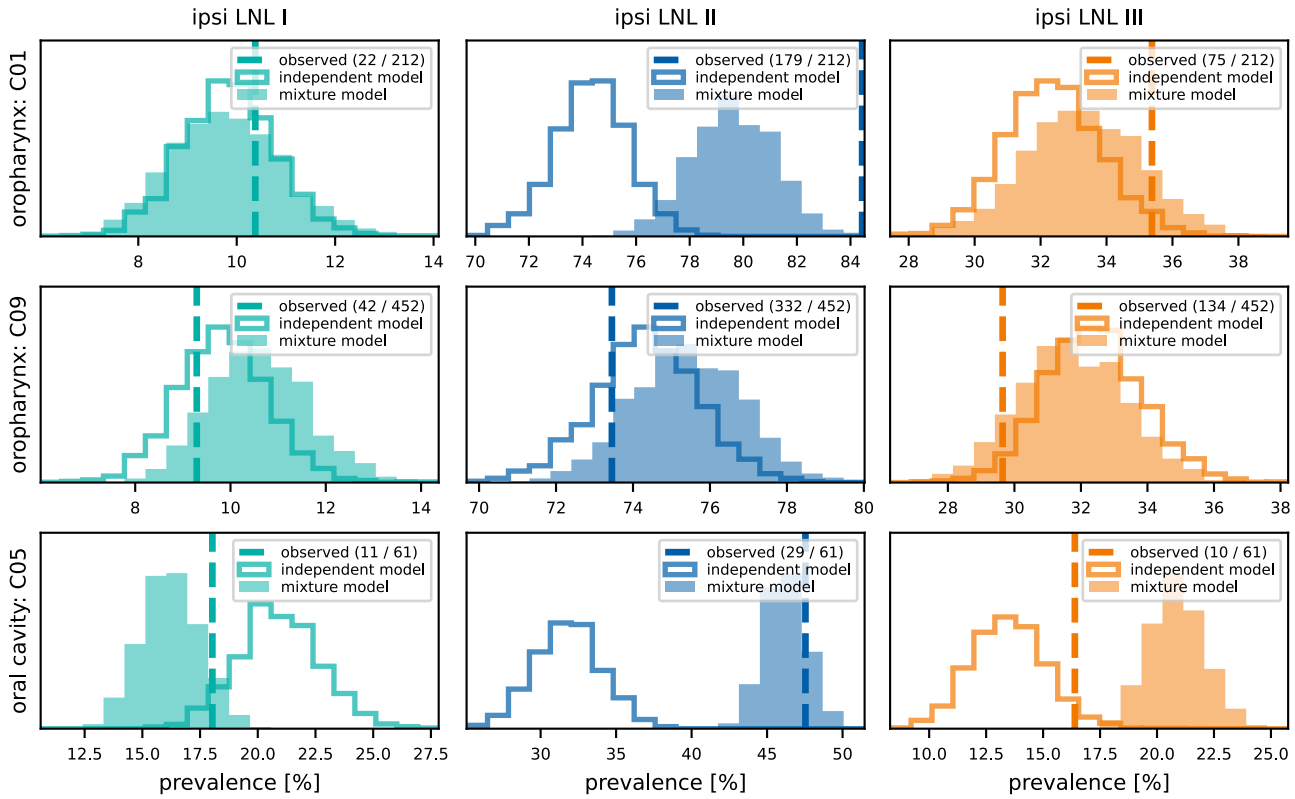


Figure 4: The prevalence of involvement as seen in the data (vertical dashed lines), predicted by an independent model for the oropharyngeal or oral cavity patients (outlined histograms), and predicted by the mixture model (filled histograms). Each row correpsonds to one subsite and each column to one LNL.

to tumors in the hypopharynx and larynx and optimize the number of model components.

Acknowledgements

This work was supported by the clinical research priority program (CRPP) "Artificial intelligence in oncological imaging" of the University of Zurich and by the Swiss cancer research foundation under grant KFS 5645–08–2022.

References

[1] R. Lindberg. "Distribution of Cervical Lymph Node Metastases from Squamous Cell Carcinoma of the Upper Respiratory and Digestive Tracts". *Cancer* 29.6 (June 1972), pp. 1446–1449. DOI: 10.1002/1097-0142(197206)29:6<1446::AID-CNCR2820290604>3.0.CO;2-C.

[2] J. Woolgar. "Histological Distribution of Cervical Lymph Node Metastases from Intraoral/Oropharyngeal Squamous Cell Carcinomas". *British Journal of Oral and Maxillofacial Surgery* 37.3 (June 1999), pp. 175–180. DOI: 10.1054/bjom.1999.0036.

[3] V. Snyder, L. K. Goyal, E. M. R. Bowers, et al. "PET/CT Poorly Predicts AJCC 8th Edition Pathologic Staging in HPV-Related Oropharyngeal Cancer". *The Laryngoscope* n/a.n/a (Jan. 2021). DOI: 10.1002/lary.29366.

[4] M. P. Strohl, P. K. Ha, R. R. Flavell, et al. "PET/CT in Surgical Planning for Head and Neck Cancer". *Imaging Options for Head and Neck Cancer* 51.1 (Jan. 2021), pp. 50–58. DOI: 10.1053/j.semnuclmed.2020.07.009.

[5] J. Biau, M. Lapeyre, I. Troussier, et al. "Selection of Lymph Node Target Volumes for Definitive Head and Neck Radiation Therapy: A 2019 Update". *Radiotherapy and Oncology* 134 (May 2019), pp. 1–9. DOI: 10.1016/j.radonc.2019.01.018.

[6] R. Ludwig, J.-M. Hoffmann, B. Pouymayou, et al. "A Dataset on Patient-Individual Lymph Node Involvement in Oropharyngeal Squamous Cell Carcinoma". *Data in Brief* 43 (Aug. 2022), p. 108345. DOI: 10.1016/j.dib.2022.108345.

[7] R. Ludwig, A. Schubert, D. Barbatei, et al. "A Multi-Centric Dataset on Patient-Individual Pathological Lymph Node Involvement in Head and Neck Squamous Cell Carcinoma". *Data in Brief* (Dec. 2023), p. 110020. DOI: 10.1016/j.dib.2023.110020.

[8] R. Ludwig, B. Pouymayou, P. Balermpas, et al. "A Hidden Markov Model for Lymphatic Tumor Progression in the Head and Neck". *Scientific Reports* 11.1 (Dec. 2021), p. 12261. DOI: 10.1038/s41598-021-91544-1.

[9] S. S. Batth, J. J. Caudell, and A. M. Chen. "Practical Considerations in Reducing Swallowing Dysfunction Following Concurrent Chemoradiotherapy with Intensity-Modulated Radiotherapy for Head and Neck Cancer". *Head Neck* 36 (2014), pp. 291–298. DOI: 10.1002/hed.23246.

[10] R. Ludwig. "Modelling Lymphatic Metastatic Progression in Head and Neck Cancer". PhD thesis. Zurich: University of Zurich, 2023.

[11] D. Foreman-Mackey, D. W. Hogg, D. Lang, et al. "Emcee: The MCMC Hammer". *pasp* 125.925 (Mar. 2013), p. 306. DOI: 10.1086/670067.

[12] R. Ludwig, J.-M. Hoffmann, B. Pouymayou, et al. "Detailed Patient-Individual Reporting of Lymph Node Involvement in Oropharyngeal Squamous Cell Carcinoma with an Online Interface". *Radiotherapy and Oncology* 169 (Apr. 2022), pp. 1–7. DOI: 10.1016/j.radonc.2022.01.035.

Data Collection Platform for Radiation Oncology Clinical Use and Medical Research in a Multivendor Setting

Diego García[1], David Brito[1]

[1]Department of Radiotherapy, Sociedad de Lucha Contra el Cáncer, Guayaquil, Ecuador

Abstract

Purpose: To develop a data collection platform for radiation oncology clinical use and Clinical Research, and to evaluate its implementation in a healthcare setting. This is accomplished by retrieve and display data obtained from Mosaiq (Elekta, Sweden) and Aria (Varian, USA) on a user-friendly interface, use the data to organize the workflow, build data collection forms and to generate automated reports and dashboards.

Introduction: In radiotherapy, is important to have an accurate an efficient way to access the relevant information about the patient's treatment course, to contextualize it on the workflow of the Department and to extract clinical parameters for Clinical research. This is critical in a mixed environment with various record and verify systems in place, as they do not communicate with each other in a native way. Here we present a locally developed platform to integrate data from Mosaiq, Aria and our local Electronic Health Record system (EHR) and display it in a user-friendly web interface.

Methods: Our system (Armoniq) was developed on a local server running PHP and MySQL to handle database queries and display the results, based on Data FAIR principles (Findable, Accessible, Interoperable and Reusable). The main conditions are that the system must have user authentication, to prevent unauthorized access and only read queries are allowed (SQL SELECT), to avoid data manipulation outside the native systems.

Results: The radiotherapy workflow module was deployed in mid-2017. So far, we have recorded the timeline of approximately 12500 cases, improving the dosimetry staff response time and replacing paper records. On the other hand, the data collection module has 186 variables meant to be used in 4 clinical retrospective studies. On average, the server handles 1500+ daily requests by 143 active users.

Conclusions: The platform provided a centralized and standardized system for data collection, storage, and retrieval, which improved the quality of patient care, streamlined the clinical workflow, and contributed to medical research in other areas of the clinic. The platform facilitated the collection and management of patient data for research purposes, while also serving as an essential tool for daily clinical practice.
Non-standardized data registration highlights the importance of standardization in healthcare data management. This work can serve as a model for other healthcare facilities seeking to improve their data management and contribute to medical research.

1 Introduction

The challenges faced by radiation oncology facilities worldwide include deficits in personnel and equipment, leading to compromised patient care and hindering medical research advancement. Insufficient qualified personnel and a scarcity of linear accelerators contribute to increased workloads, longer waiting times, compromised treatment planning, delayed interventions, and reduced treatment capacity, impacting the overall quality and accessibility of radiation therapy services. Furthemore, in a multivendor setting, interoperability problems and the absence of standardized tools exacerbate inefficiencies in data consolidation and obstruct strategic decision-making. The varied data formats across different systems impede a unified approach to optimizing resource and personnel allocation, equipment acquisition and treatment strategies.

In the dynamic field of radiation therapy and oncology, where accessing patient treatment information is crucial, the locally developed platform aims to transcend system incompatibility by seamlessly integrating data from disparate Record and Verify systems, such as Mosaiq, Aria, and the local Electronic Health Record system. The platform provides a user-friendly web interface, harmonizing data sources and empowering clinicians and researchers with a unified and comprehensive view of patient information, advancing patient care and scientific innovation.

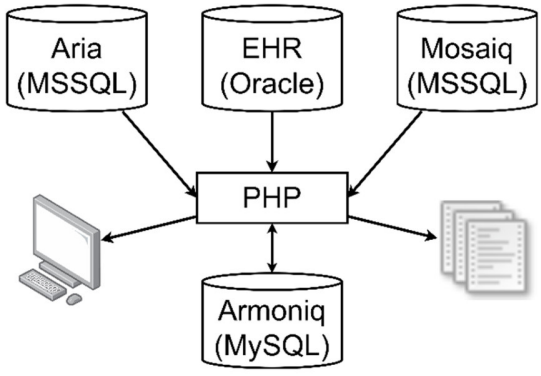


Fig 1. Basic software architecture diagram. Our server uses PHP to read data from external systems, save to a separate database and generate reports

2 Materials and Methods

The development of a comprehensive data collection platform, named Armoniq, was executed on a local server running PHP and MySQL (Fig.1). This infrastructure was chosen to efficiently handle database queries and display results. Armoniq was designed with adherence to the core tenets of the FAIR principles, ensuring that data within the system is easily findable, accessible, interoperable, and reusable.

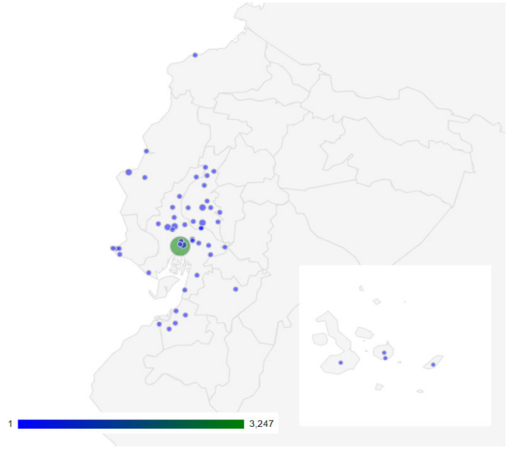


Fig. 2 Geographic distribution of the patients treated in the Radiation Therapy Department..

The main conditions governing the system's functionality include robust user authentication mechanisms to prevent unauthorized access, a critical consideration in handling sensitive patient data within the radiation oncology department. Furthermore, the system strictly allows read queries (SQL SELECT) exclusively, mitigating the risk of unintended data manipulation outside the native systems. This restriction enhances data security and integrity, aligning with best practices in healthcare information management.

This locally hosted system was seamlessly integrated into the radiation oncology department, replacing traditional paper records, and serving as a versatile tool for workflow management, task coordination, patient check-in, follow-up, and data mining from the two different Oncology Information Systems (OIS). The implementation underwent rigorous evaluation, incorporating feedback from healthcare providers and researchers, ensuring that Armoniq not only addressed the identified challenges but also complied with security protocols and enhanced decision-making processes within the multivendor context of the department.

3 Results

The implementation of the radiotherapy workflow module in mid-2017 marked a transformative milestone in the radiation oncology department. To date, the system has meticulously recorded the timelines of approximately 12,500 cases, providing a comprehensive and accessible repository of patient data. This deployment significantly contributed to the enhancement of dosimetry staff response times, streamlining the treatment planning process, and replacing traditional paper records with a more efficient and secure digital solution.

Simultaneously, the data collection module within Armoniq demonstrates its versatility as a powerful research tool. With a robust framework comprising 186 variables carefully designed for application in four distinct clinical retrospective studies, this module empowers researchers with the depth and breadth of information needed to drive advancements in the field of radiation oncology. The flexibility of the platform enables researchers to extract valuable insights, contributing to a growing body of knowledge within the department.

In terms of system usage, Armoniq has become an integral part of the daily operations, with the server consistently handling over 1,500 requests each day. This high demand is met by the active engagement of 143 users who rely on the platform for various tasks related to patient care, workflow management, and data analysis. The substantial user base and daily requests underscore the system's reliability and efficiency in supporting the multifaceted needs of the radiation oncology department. The successful integration and sustained utilization of Armoniq highlight its pivotal role in not only addressing the challenges posed by deficits in personnel and equipment but also in facilitating a data-driven approach to decision-making and research initiatives within the department.

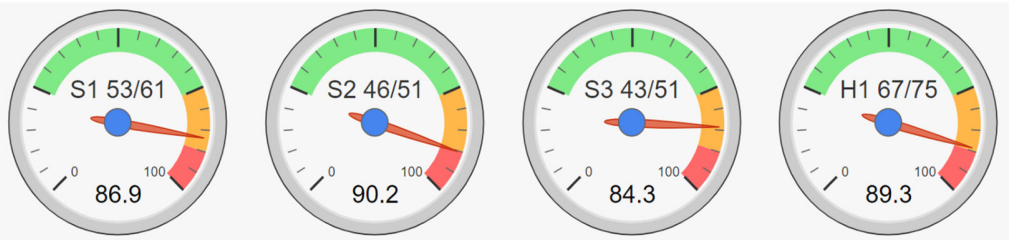


Fig. 4 Daily linac workload. The gauge shows the machine, number of activities planned / number of available schedule slots and the equivalent percentage of use

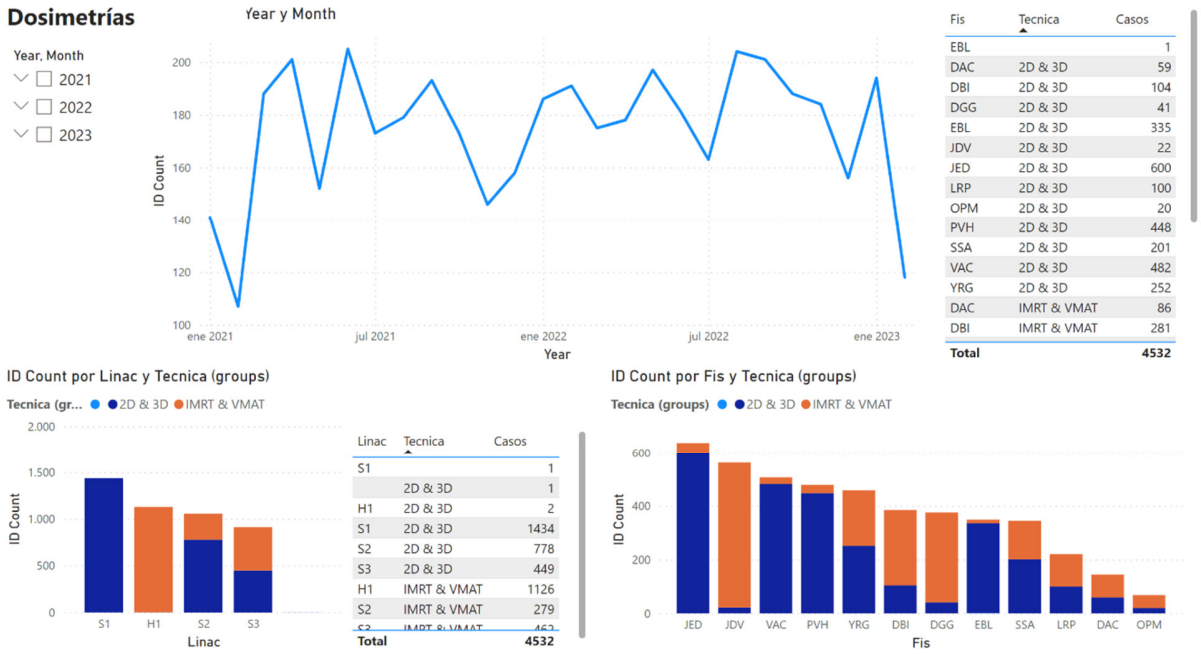


Fig. 5 Screenshot of the interactive dosimetry dashboard made using MS Power BI. The number of calculated plans can be filtered and ordered by dosimetrist, technique and linac. We use this data on the decision-making process for acquiring linacs and hire new personnel.

4 Discussion

While this work has demonstrated the positive impact of the data collection platform on patient care, clinical workflows, and medical research within the radiation oncology department, the platform is facing extinction due to external pressures from administration. This raises concerns about the sustainability and resilience of such initiatives in the face of broader institutional changes. The imposition of a single OIS by external vendors, driven by commercial strategies, and recent decisions to dismantle the platform highlight the vulnerability of healthcare systems to external influences, where short-term administrative goals may overshadow the long-term benefits of integrated and standardized data management solutions.

The discussions surrounding the potential extinction of the platform underscore the need for comprehensive evaluations of decisions impacting healthcare data infrastructure, considering not only immediate financial considerations but also the enduring value of interoperability, standardization, and continuity in patient care and research efforts. As the landscape of healthcare data management continues to evolve, this experience serves as a valuable lesson, emphasizing the importance of sustained commitment to data standardization and interoperability for the greater benefit of patient outcomes and scientific progress.

5 Conclusion

The implementation of the data collection platform brought about transformative changes in the radiation oncology department, providing a centralized and standardized system for data collection, storage, and retrieval. The positive impact of this platform extended across multiple facets of healthcare, demonstrating improvements in patient care, streamlined clinical workflows, and a noteworthy contribution to medical research within various areas of the clinic. By serving as a repository for patient data, the platform not only enhanced the efficiency of daily clinical practices but also emerged as a valuable tool for advancing medical research endeavors. This integration of patient data for research purposes underscores the platform's versatility, offering a dual benefit of supporting both clinical operations and research initiatives.

Despite challenges arising from non-standardized data registration across various OIS, Armoniq facilitated extensive data mapping and integration efforts. The challenges posed by non-standardized data registration further emphasize the critical need for standardization in healthcare data management. The absence of standardized practices creates obstacles not only in daily operations but also hampers the potential for meaningful research outcomes. Standardization is paramount for ensuring the accuracy, consistency, and reliability of healthcare data,

which is fundamental for evidence-based decision-making and the advancement of medical knowledge. The experiences encountered in addressing non-standardized data registration highlight the importance of establishing robust data management practices as an integral component of effective healthcare systems.

6 References

Bialke, M., Rau, H., Thamm, O. C., Schuldt, R., Penndorf, P., Blumentritt, A., … Hoffmann, W. (2018). Toolbox for Research, or how to facilitate a central data management in small-scale research projects. Journal of Translational Medicine , 16 (1), 16. https://doi.org/10.1186/s12967-018-1390-1

Bruland, P., & Dugas, M. (2017). S2O - A software tool for integrating research data from general purpose statistic software into electronic data capture systems. BMC Medical Informatics and Decision Making , 17 (1), 3. https://doi.org/10.1186/s12911-016-0402-4

Kirrmann, S., Gainey, M., Röhner, F., Hall, M., Bruggmoser, G., Schmucker, M., & Heinemann, F. E. (2015). Visualization of data in radiotherapy using web services for optimization of workflow. Radiation

Oncology, 10 (1), 1–12. https://doi.org/10.1186/s13014-014-0322-3
Mayo, C., Phillips, M., McNutt, T., Palta, J., Dekker, A., Miller, R., … Lawrence, T. (2018). Treatment data and technical process challenges for practical big data efforts in radiation oncology. Medical Physics , 45 (10), e793–e810. https://doi.org/10.1002/mp.13114

Mayo, C. S., Kessler, M. L., Eisbruch, A., Weyburne, G., Feng, M., Hayman, J. A., … Ten Haken, R. K. (2016). The big data effort in radiation oncology: Data mining or data farming? Advances in Radiation Oncology , 1 (4), 260–271. https://doi.org/10.1016/j.adro.2016.10.001

Pan, H. Y., Mazur, L. M., Martin, N. E., Mayo, C. S., Santanam, L., Pawlicki, T., … Smith, B. D. (2017). Radiation Oncology Health Information Technology: Is It Working for or Against Us? International Journal of Radiation Oncology Biology Physics, 98 (2), 259–262. https://doi.org/10.1016/j.ijrobp.2017.02.018

Pasalic, D., Reddy, J. P., Edwards, T., Pan, H. Y., & Smith, B. D. (2018). Implementing an Electronic Data Capture System to Improve Clinical Workflow in a Large Academic Radiation Oncology Practice. JCO Clinical Cancer Informatics , (2), 1–12. https://doi.org/10.1200/cci.18.00034

Phillips, M. H., Serra, L. M., Dekker, A., Ghosh, P., Luk, S. M. H., Kalet, A., & Mayo, C. (2020). Ontologies in radiation oncology. Physica Medica , 72 (March), 103–113. https://doi.org/10.1016/j.ejmp.2020.03.017

Ruth, C. J., Huey, S. L., Krisher, J. T., Fothergill, A., Gannon, B. M., Jones, C. E., … Mehta, S. (2020). An Electronic Data Capture Framework (ConnEDCt) for Global and Public Health Research: Design and Implementation. Journal of Medical Internet Research , 22 (8), e18580. https://doi.org/10.2196/18580

Strickler, M., Zbinden, C., Saccilotto, R., & Denecke, K. (2019). A Concept for a Data Dictionary System Supporting for Clinical Research. Studies in Health Technology and Informatics , 258 , 158–162. https://doi.org/10.3233/978-1-61499-959-1-158

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., … Mons, B. (2016). Comment: The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data , 3 , 1–9. https://doi.org/10.1038/sdata.2016.18

PARROT: A Web Platform for AI-Driven Radiation Therapy Planning

Margerie Huet Dastarac[1], Wei Zhao[2], Sylvain Deffet[3], Eléonore Longton[4], Sophie Cvilic[4], Edmond Sterpin[1], John Lee[1], and Ana Barragan Montero[1]

[1]Molecular Imaging, Radiation and Oncology (MIRO) Laboratory, Institut de Recherche Expérimentale et Clinique (IREC), UCLouvain, Belgium
[2]Institute of Information and Communication Technologies (ICTEAM), UCLouvain, Belgium
[3]Ion Beam Applications SA, Louvain-La-Neuve, Belgium
[4]Department of Radiation Oncology, Saint-Luc University Hospital, Belgium

Abstract Artificial Intelligence (AI) is gaining momentum in medical fields like radiation therapy. Models to delineate structures and predict optimal dose trade-off for patients are developed by the research community. Although some of these models are starting to be implemented clinically, an end-to-end workflow with a user-friendly graphical interface to visualize the result of successive AI models and to support decision of treatment indication is still prospective.

To address this problem, we present PARROT, a free and open-source web platform that facilitates the use of AI delineation and dose prediction models and the visualization of the models outputs. The treatment decision support shows clinical evaluation tools to compare dose distributions, among which normal tissue complication probabilities (NTCP) models.

1 Introduction

Artificial Intelligence (AI) is gaining momentum in medical fields like radiation therapy. Models to delineate structures and predict optimal dose trade-off for patients are developed by the research community. Although some of these models are starting to be implemented clinically, an end-to-end workflow with a user-friendly graphical interface to visualize the result of successive AI models and to support decision of final treatment indication is still prospective.

To address this problem, we introduce PARROT which stands for Platform for ARtificial intelligence guided Radiation Oncology Treatment. PARROT is a user-friendly, free and open-source web platform which allows users to visualize DICOM files, run AI models in a single mouse click, display and evaluate predictions easily. Several trained state-of-the-art dose prediction and contour segmentation models are available. Moreover, users can add their own models from the embedded code editor. The app consists of a front-end built in React JavaScript including the graphical interface with DICOM viewer, a selection panel of AI models, an editor of contours predictions and evaluation tools. The back-end is built with Flask to handle the predictions of AI models.

Download instructions are available on our website https://huggingface.co/spaces/AI4MIRO/parrot.ai and the source code can be found on the Gitlab repository https://gitlab.com/ai4miro/parrot.

2 Materials and Methods

The interface is organized in four screens, Figure 1.

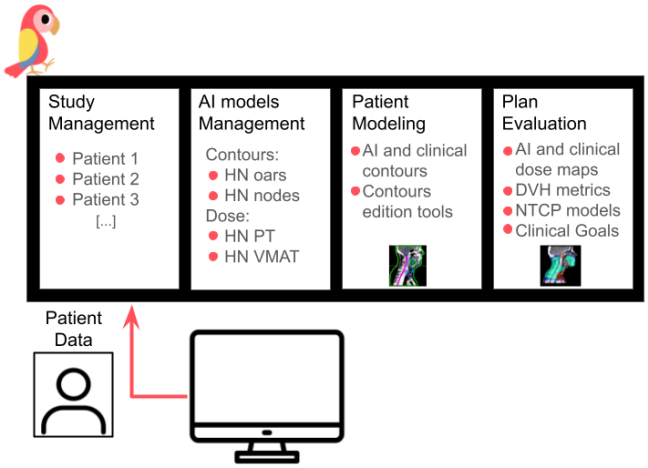


Figure 1: Overview of the PARROT platform.

2.1 Study Management

Users can import patient data from this screen to a Orthanc daemon—a lightweight, standalone DICOM server [1]. It runs locally and therefore ensures the confidentiality of patient files. The interactions with the Orthanc daemon are done through its RESTful API [2]. The list of patient data present in the Orthanc server is displayed with summarized information (Figure 2). Users can load the data into the app by clicking on the Load button. The left icon displays more complete information from DICOM tags (Figure 3) and users can delete a patient study from the server with the red arrow on the right.

2.2 AI Models Management

Eight segmentation models and six dose prediction models are made available and listed on the screen (Figure 4). These models are trained from state-of-the-art architectures such as nnUNet [3], SwinUNETR [4] and HDUNet [5] (Table 1). On the left hand side of each line, a standardized description of the model is available. The PDF description file contains details about the architecture with a link of the original research

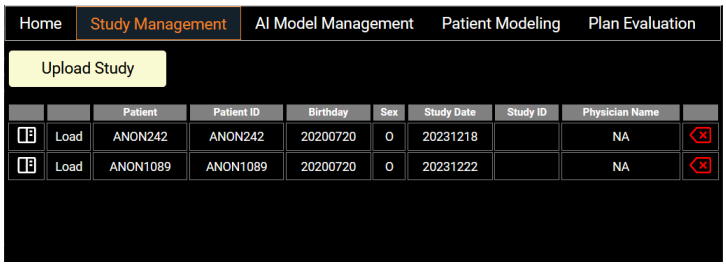


Figure 2: Study management screen.

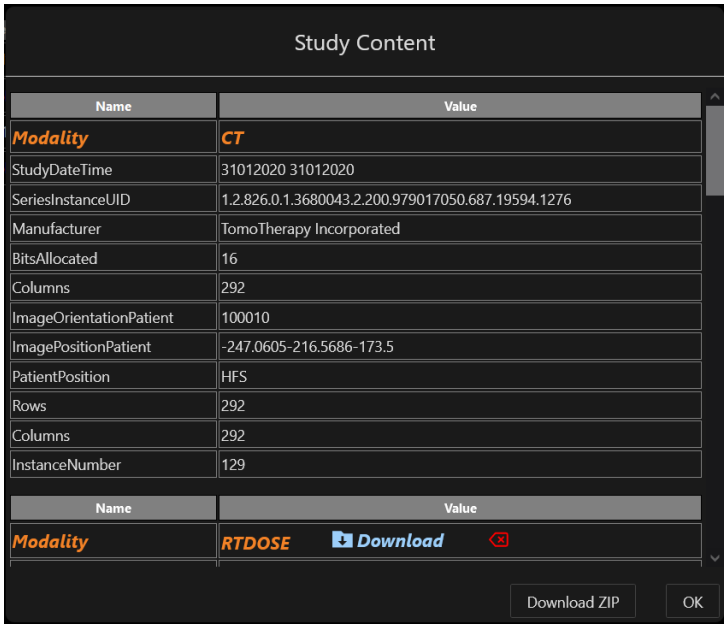


Figure 3: Content information of the study stored in DICOM tags.

article, details on the training process including dataset, and performance on our test set (Figure 5).

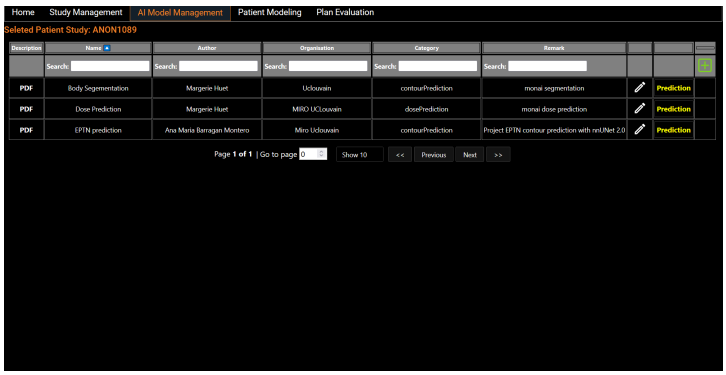


Figure 4: AI Models Management screen. A list of the embedded AI models are displayed.

To run a model prediction, the user simply clicks on the yellow Prediction button on the right hand side of the model line. This will open a Python scripting editor to run the code (Figure 6). Users can select a Python environment with a set of libraries from the drop-down menu in the upper right corner. By hovering on an item, a description of the libraries it contain is displayed. Once the prediction is done, users can upload the result to the study to visualize the output in the Patient Modeling screen (Figure 8). Users

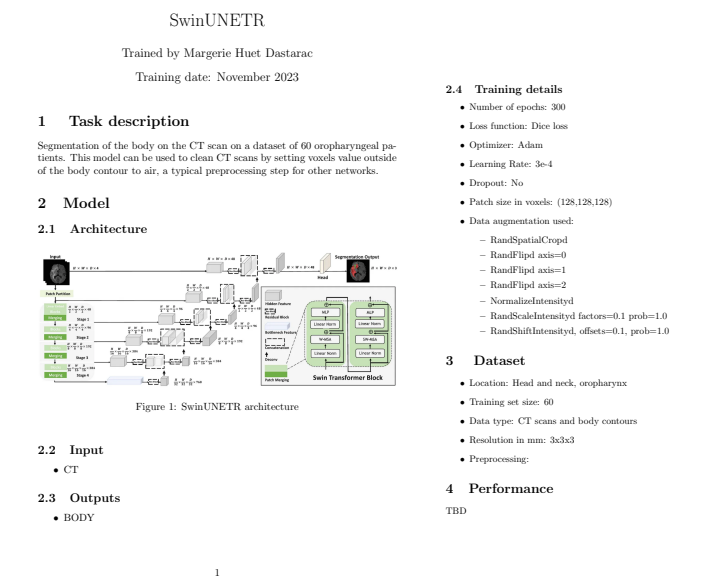


Figure 5: Example of PDF with description of a specific model provided in PARROT.

also have the flexibility to run their own models by loading their processing and inference scripts in the Python scripting editor. On our website, we provide a description and examples using the API to retrieve the patient information loaded on PARROT and upload the result.

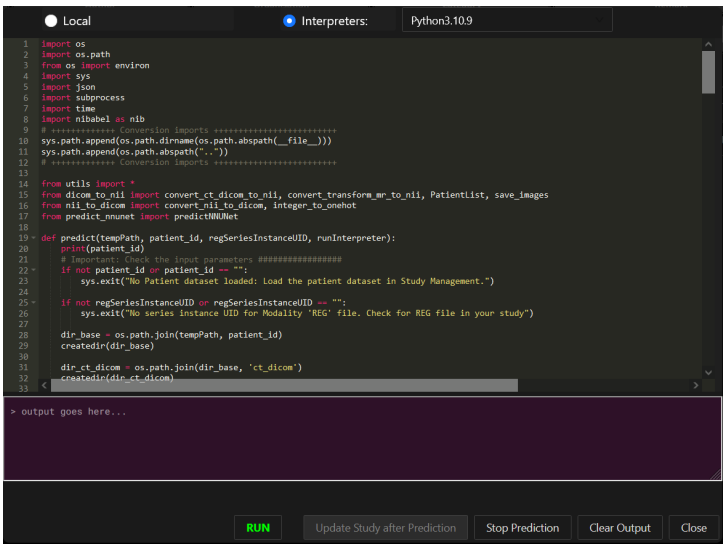


Figure 6: Python scripting editor. Users can edit their script, run and display output from this window.

We distinguish two categories of models: the segmentation models and the dose prediction models. A key difference between the two categories is that segmentation models require only the anatomy of the patient (CT scan or MRI image) whereas dose prediction models also require binary masks of tumor volumes (TV) and organs at risk (OARs). For the latter, we implemented a dialog that allows users to map structures present in the current patient data and the

input required for the AI dose prediction model, Figure 7.

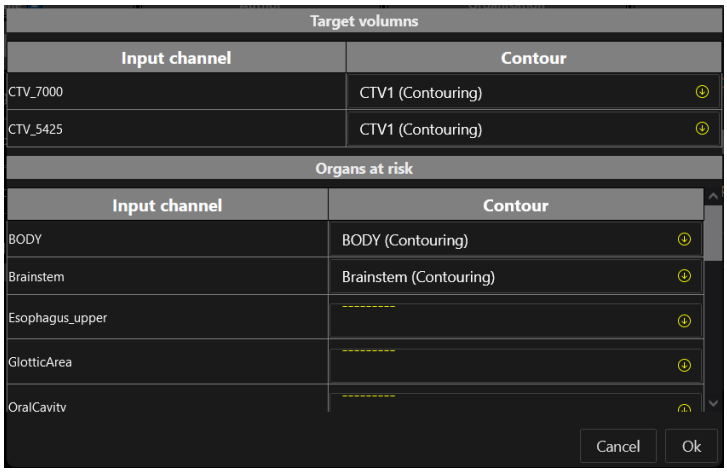


Figure 7: Mapping dialog between structures present in the current patient data and required inputs of the model. The left column displays the input channels required for that specific AI model and the right column a drop-down list of the contours available in the loaded patient study.

2.3 Patient Modeling

This screen allows an interactive visualization of contours, whether uploaded by the user or generated by AI models in the previous tab. Similarly to the visualization of treatment planning software, users can scroll through the anatomy and contours of the patient, zoom, pan and change window range and levels. Users can edit contours with the different tools that can be seen Figure 8. The platform maintains a detailed history of modifications, and users can provide descriptive labels for refining AI models.

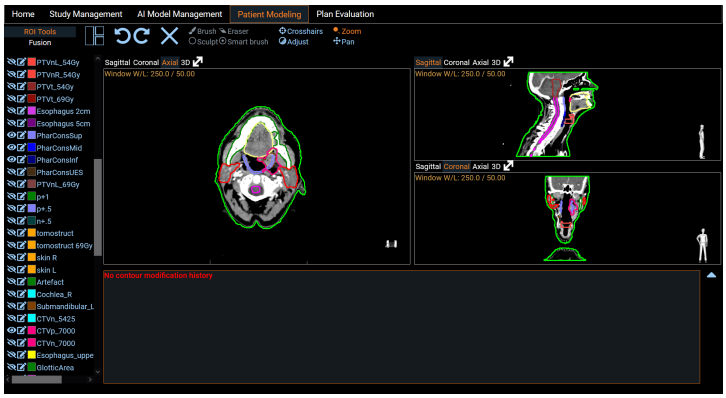


Figure 8: Patient Modeling screen.

2.4 Plan Evaluation

A comprehensive display of two dose distributions comparison is presented on the last screen. The two dose distributions can originate from the importation step or result from the inference of AI models. We provide a set of tools for comparison, such as dose-volume histograms (DVH) curves, dose statistics, and the fulfillment of clinical goals. Clinical goals can be uploaded following an Excel template so that it fits any clinical requirements of the users institutions. The platform incorporates treatment indications based on NTCP protocols. It has been shown in multiple studies that predicted dose distributions yield meaningful output when used with NTCP models for treatment selection decision support [6, 7]. Two NTCP protocols, validated by the clinical community have been integrated, one for head and neck cancer patients (with four NTCP models for xerostomia grade $\geq$ II and grade $\geq$ III and dysphagia grade $\geq$ II and grade $\geq$ III)[8] and one for esophagus cancer patients to estimate pulmonary complications probabilities [9]. This is a useful decision support tool for the clinical community.



Figure 9: Plan Evaluation screen.

3 Results

Three main use cases that can benefit from PARROT are the following.

3.1 Visualization of state-of-the-art AI models on patient of local institute

The integration of a novel AI model within a clinical setting necessitates a careful and meticulous analysis of the tool. It is imperative to conduct thorough testing using a local patient database to compare the performance of the model in its initial testing conditions and the specific context of the new healthcare institution. The PARROT framework facilitates seamless visualization of predictions, eliminating the need for manual coding.

3.2 Curate own AI segmentation models with the contour editing tools

Utilizing the features available on the viewer screen, clinicians have the capability to rectify contours produced by AI models and annotate the rationale behind the modifications. This process helps to identify systematic errors and limitations of the AI models, guiding efforts to curate the training set and retrain the model.

Segmentation		Dose prediction	
Location	Architecture	Location	Architecture
Head and Neck Nodes	nnUNetv2	Head and Neck PT	HDUNet
Head and Neck OARs	nnUNetv2	Head and Neck VMAT	HDUNet
Prostate OARs	nnUNetv2	Prostate IMRT	HDUNet
Breast OARs	nnUNetv2	Breast PT	HDUNet
Brain OARs	nnUNetv2	Esophagus IMRT	HDUNet
Brain Tumor	SwinUNETR	Esophagus PT	HDUNet
Esophagus OARs	nnUNetv2		
Body	SwinUNETR		

Table 1: List of provided AI models.

3.3 Decision support between PT and XT treatments

The dose evaluation screen of PARROT serves as a decision support tool for choosing between PT and XT treatments. Users can have a seamless overview of DVH curves, dose metrics, clinical goals reached and indication of NTCP models. In the current version, AI dose prediction models and corresponding NTCP models are provided for oropharyngeal cancer patients (PT versus VMAT) and esophageal cancer (PT versus IMRT).

4 Discussion

We believe that PARROT can have a positive impact on the safe integration of AI models in the clinic. On the one hand, PARROT can facilitate model comparison, sharing, and benchmarking by the research community. On the other hand, it can foster the dissemination of AI models to the clinics, standardizing and improving the medical practice. In addition, it can be a valuable tool to foster consensus between experts when updating or creating new contouring and planning guidelines. In the future, we would like to extend PARROT to integrate different methods of uncertainty quantification both for segmentation and dose prediction models. These uncertainty estimation tools can be used to curate AI models and datasets and flag out-of-distribution patients and therefore indicate clinicians that the AI model should not be used for that specific patient.

Currently, the platform supports CT and MR images, and automatically applies a registration field whenever a REG DICOM file is uploaded. But future releases will allow PET scan to be viewed and used by AI models.

5 Conclusion

PARROT is a a free and open-source web platform that facilitates the use of AI models in the radiation therapy field. The web platform allows users to visualize DICOM files, run AI models (contour segmentation and dose prediction models) in a single mouse click, display and evaluate predictions easily. PARROT offers clinical evaluation tools to compare dose distributions and support treatment decision. Among these tools, users can find dose-volume histograms curves, dose statistics, validation of clinical goals and outputs of normal tissue complication probabilities models. We believe that PARROT can have a positive impact on the safe integration of AI models in the clinic.

References

[1] S. Jodogne. “The Orthanc Ecosystem for Medical Imaging”. *JDI* 31.3 (May 2018).

[2] *REST API of Orthanc — Orthanc Book documentation*. https://orthanc.uclouvain.be/book/users/rest.html. Accessed: 2024-1-8.

[3] F. Isensee, P. F. Jaeger, S. A. A. Kohl, et al. “nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation”. *Nat. Methods* 18.2 (Feb. 2021).

[4] A. Hatamizadeh, V. Nath, Y. Tang, et al. “Swin UNETR: Swin Transformers for Semantic Segmentation of Brain Tumors in MRI Images”. *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*. Springer International Publishing, 2022.

[5] D. Nguyen, X. Jia, D. Sher, et al. “3D radiotherapy dose prediction on head and neck cancer patients with a hierarchically densely connected U-net deep learning architecture”. *PMB* 64.6 (Mar. 2019).

[6] M. Huet-Dastarac, S. Michiels, S. T. Rivas, et al. “Patient selection for proton therapy using Normal Tissue Complication Probability with deep learning dose prediction for oropharyngeal cancer”. *Med. Phys.* (May 2023).

[7] C. Draguet, A. M. Barragán-Montero, M. C. Vera, et al. “Automated clinical decision support system with deep learning dose prediction and NTCP models to evaluate treatment complications in patients with esophageal cancer”. *Radiother. Oncol.* 176 (Nov. 2022).

[8] N. V. voor Radiotherapie en Oncologie (NVRO). *LIPPv2*. https://nvro.nl/images/documenten/rapporten/2019-08-15__Landelijk_Indicatieprotocol_Protonentherapie_Hoofdhals_v2.2.pdf. [Accessed 25-01-2024].

[9] M. Thomas, G. Defraene, M. Lambrecht, et al. “NTCP model for postoperative complications and one-year mortality after trimodality treatment in oesophageal cancer”. *Radiother. Oncol.* 141 (Dec. 2019).

Dose evaluation on daily CBCTs for breast cancer patients

Peter Stijnman[1], Maureen Groot Koerkamp[1], Anette Houweling[1], Cornel Zachiu[1], and Bas Raaymakers[1]

[1]Department of Radiotherapy, University Medical Center Utrecht, Utrecht, Netherlands

Abstract In this work, an interfraction dose evaluation workflow based on daily CBCT data for breast cancer patients is shown. Using deformable image registration a synthetic CT is created from the planning CT and the daily CBCT and the contours delineated on the planning CT are propagated to the synthetic CT. This synthetic CT is used to calculate the daily delivered dose, from which a trigger system can be built to flag patients who might receive suboptimal treatment. Currently, the workflow has processed 456 patients with a total of 5157 fractions. The population-based data suggests that overall the delivered treatment meets a very high standard (i.e. 97.3% of patients receive an average CTV coverage higher than 95%). However, there are some patients for whom the treatment could be improved based on the patient-specific information that this workflow supplies. Furthermore, any uncertainties in the clinic about the impact of anatomical changes on the delivered dose can be verified with this workflow.

1 Introduction

The recent trend in radiotherapy is to move from a single treatment plan that is used throughout multiple fractions to an online adaptive treatment [1]. The result would be a better-tailored treatment that takes into account the daily patient anatomy. To achieve this goal, the treatment plan has to be re-optimized every day. This is a computationally expensive procedure and requires expertise at the treatment machine to run the online adaptive workflow and approve the newly optimized treatment.

As a stepping stone between planning a single treatment plan and performing fully online adaptive radiotherapy, we propose a workflow where the CBCT imaging, acquired for position verification, is used to calculate the planned dose on the daily anatomy to find patients that might receive suboptimal treatment, similar work has been shown in [2].

The benefit of the proposed approach is that it is relatively inexpensive to check if the current treatment plan still adheres to the prescribed dose constraints. These constraints can easily be checked at the treatment machine and a decision can be made to either go ahead with the treatment or to make a new treatment plan for the patient. The calculated daily delivered dose can provide more useful information about the treatment quality compared to only the anatomical information the CBCT provides.

2 Materials and Methods

Patient data is gathered using a daily SQL query on a server containing the CBCT DICOM-files. The CBCTs are filtered on the series description tag to only obtain the breast cancer patients. The number of breast cancer patients that are treated within our hospital on a daily basis is around 45.

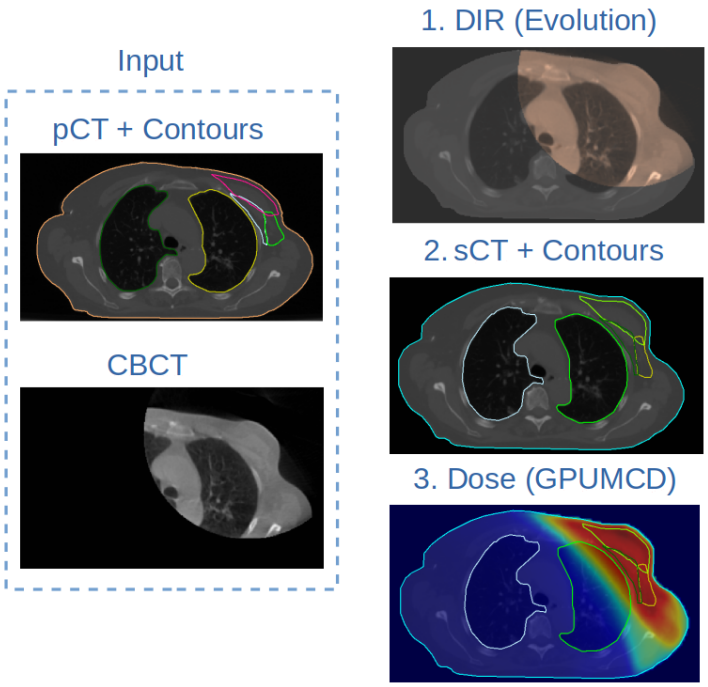


Figure 1: A schematic overview of the workflow. The planning CT with any relevant delineations and the daily CBCT are used as input. The first step is to perform a deformable image registration inside the CBCT FoV and a rigid image registration outside. The second step is to propagate the delineations. Finally, we can use the synthetic CT and the planned treatment to calculate the daily delivered dose.

A schematic overview of the workflow is shown in Figure 1. Before the treatment start, the daily CBCT is matched to the planning CT (pCT) in order to align the daily anatomy to that of with the anatomy as seen on the pCT. Using this matched CBCT and the pCT a daily synthetic CT (sCT) is created using deformable image registration, because the CBCT units are not intended for dose calculation. Since the CBCT only contains part of the anatomy, we perform the deformable image registration of the pCT to the CBCT only within the field of view (FoV) of the CBCT. Outside of the FoV of the CBCT the pCT is rigidly registered and stitched to the deformable registered pCT as a best estimate of the patient anatomy outside the CBCT FoV. Any air pockets that remain at the periphery of the sCT but are inside the newly generated body contour on the CBCT, are filled with a -120 HU value, corresponding to fatty tissue. This generated sCT is used

for the daily dose calculation. For both the deformable and rigid image registration Evolution, an in-house developed implementation, is used [3].

Using the deformation vector fields that are obtained from Evolution we propagate the contours that were delineated on the pCT to the sCT. For this patient group, we propagated the clinical target volumes (CTV) (i.e. breast and lymph node targets), gross tumour volumes (GTV), heart, and lung contours. For the dose calculation, we created a new body contour based on the sCT. Using this body contour the propagated CTVs are also adjusted to ensure that they are adapted to any changes in the body contour. This is done by first expanding the CTV in both the lateral and anterior direction with 1 cm and then cropping the expanded CTV on 0.5 cm below the body contour. For the lymph node and CTV boost contours no expansion and only the crop on the body contour is performed.

With the sCT and propagated contours available and the treatment plan unchanged, the daily delivered dose can be calculated. This is done using the GPUMCD dose engine from Elekta AB, Stockholm, Sweden. An application to call this dose engine from the command line has been developed in-house and was tested against dose calculations performed with the Monaco treatment planning system from Elekta AB. The dose is calculated with a variance of 3% on a 3 mm isotropic grid.

From the calculated daily delivered dose and the propagated contours the dose volume histograms (DVHs) can be computed. This is achieved by first interpolating the dose onto a 1 mm isotropic grid. The dose is then grouped in 2000 bins to obtain the DVHs that are comparable to DVHs obtained from Monaco. This information, together with the volume of the the contour, the prescribed, minimum, mean, and maximum dose within the contour is exported to a .csv file. From this file, any DVHs statistics can be calculated, for example the mean dose and the V95% are the most relevant statistics for the CTVs and the mean dose, V5Gy, and V10Gy for the organs at risk (OARs).

The DVH curves and statistics are written into a patient-specific report, making it is easy to track the daily delivered dose over the history of the treatment. Based on predetermined thresholds, for example, a CTV V95% coverage below 95%, a trigger can be implemented for clinicians to inspect the daily delivered dose or view any trends over the delivered fractions.

3 Results

At the time of writing, the interfraction dose for a total of 456 patients has been calculated, totaling 5157 fractions over 4 and a half months. On average it takes 6 minutes to perform the workflow for a single fraction. The majority of the time is spent on propagating the contours to the sCT, predominantly, because this has not yet been optimized for runtime.

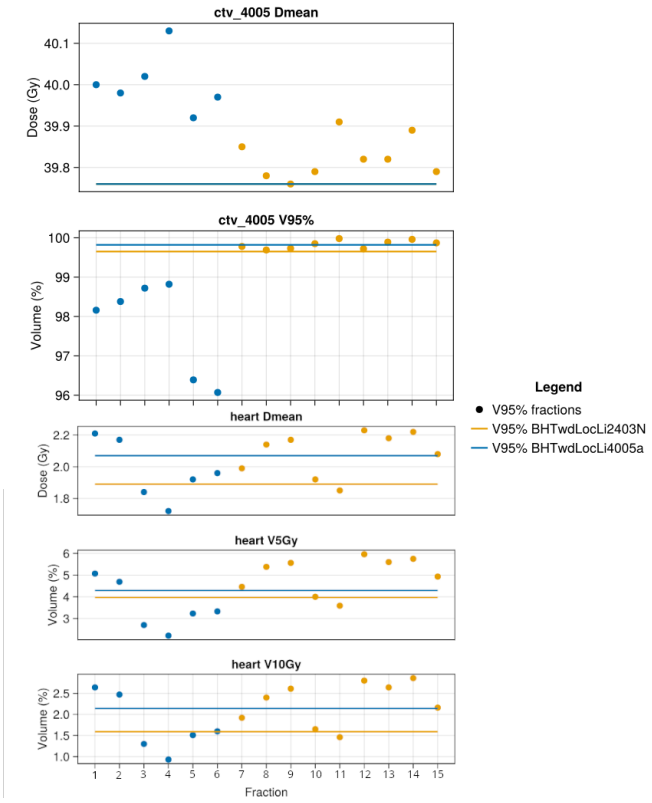


Figure 2: The mean dose and V95% for the CTV, and the mean dose, V5Gy and V10Gy per fraction for the heart of a single patient. The dots represent the separate fractions, while the lines represent the same values based on the planning CT and are used for comparison. This patient had a new treatment plan after 6 fractions to improve the CTV coverage.

To showcase the trends that can be observed from the daily dose, we highlight the CTV coverage and the heart dose for a single patient in Figure 2. This patient received a new treatment plan after 6 fractions because the CTV coverage was deemed too low to be clinically acceptable. From fraction 7 onwards the new treatment plan was started. As can be observed from Figure 2 the CTV coverage is improved drastically for this new treatment plan, while the heart dose is not significantly changed compared to the first treatment plan.

Rather than viewing the data for a single patient, we can also make an overview of the entire breast cancer patient population. This is especially valuable information for assessing the treatment quality and plan robustness that is delivered within the hospital. Such an overview is shown in Figure 3. From Figure 3, it can be observed that for the large majority of the patient population, a sufficiently high median CTV coverage is achieved on an interfraction basis. For the patients where this was not the case, a big reason was the lower number of fractions. The fractionation of the prescribed treatment ranges from 5 to 20 fractions. When the first 2 fractions of a 5 fraction treatment do not have desirable CTV coverage the mean coverage will be heavily impacted.

Another interesting aspect that we can investigate is the robustness of different radiation plan types. As an example, patients where lymph node structures are irradiated are treated

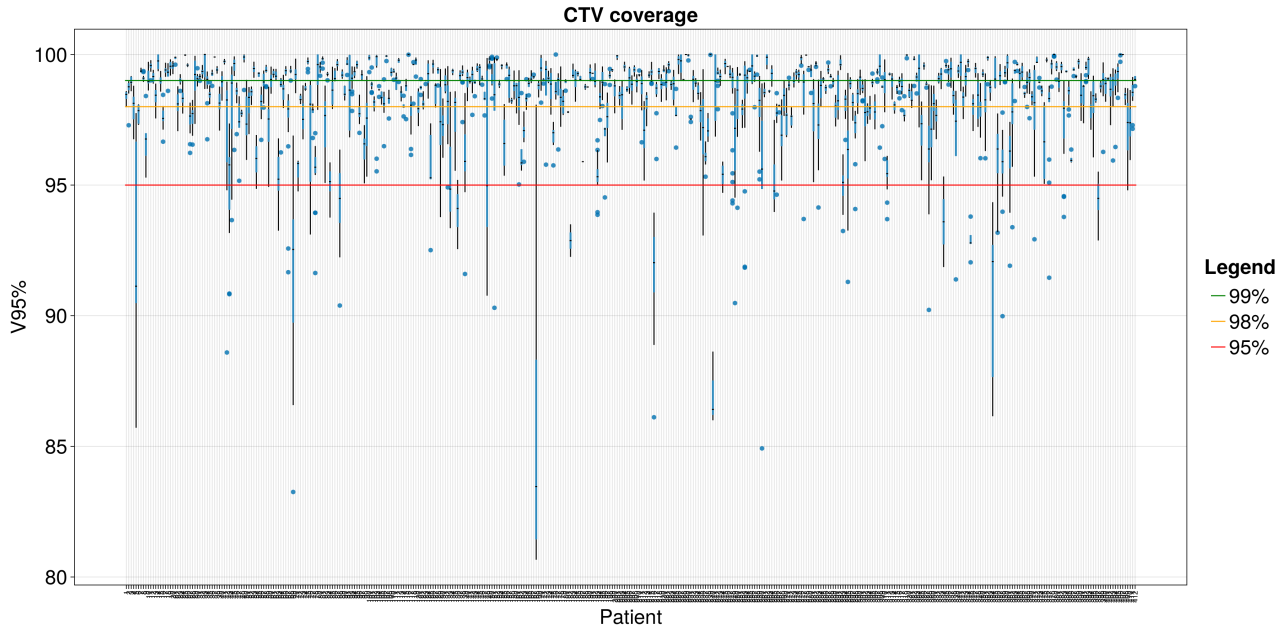


Figure 3: An overview of the CTV coverage for the patient population. Each boxplot represents the CTV V95% coverage for a single patient, showing the median, IQR, and the outliers. The green, yellow, and red lines are to indicate the 99%, 98%, and 95% coverage levels respectively. The target for the clinic is to have coverage above 95%. Currently, we check fractions below this threshold to evaluate potential causes for the lower coverage.

with a full volumetric modulated arc therapy (VMAT) plan type, while patients who do not receive radiation of lymph node structures are treated with a combination of tangential radiation fields (∼80% dose) and VMAT (∼20% dose) since it is a more robust method and the delivery time is shorter. The patients for both treatment plan types can be split into 4 groups: 1) this workflow would warrant a check and the clinic did not make a new treatment plan; 2) where this workflow would warrant a check and the clinic did make a new treatment plan; 3) no trigger, but more than one treatment plan; and 4) no trigger and no new treatment plan. If we were to trigger a check with this workflow based on a CTV V95% coverage lower than 95%, we would obtain the results that can be observed from Figure 4.

What we observe from Figure 4 is that the treatment for most patients goes as planned and no changes need to be made. We observed a similar result based on the data from Figure 3. For some patients, we do observe that a new treatment plan was made while our workflow would not trigger a check for that patient. In these cases, often the reason for a new treatment plan was discomfort in positioning the arms above the head. Furthermore, there are some cases where our workflow would trigger a check of the plan quality while the clinical workflow did not notice a need for a new treatment plan. This can be because of a few reasons, however, it should be noted that a single fraction with a CTV V95% coverage lower than 95% is quite a strict threshold. Patients from groups 1 and 3 are the most interesting cases for an evaluation, because there might be patients that receive a suboptimal treatment that are not caught by the clinical workflow.

Finally, Figure 5 indicates an example where a difference in

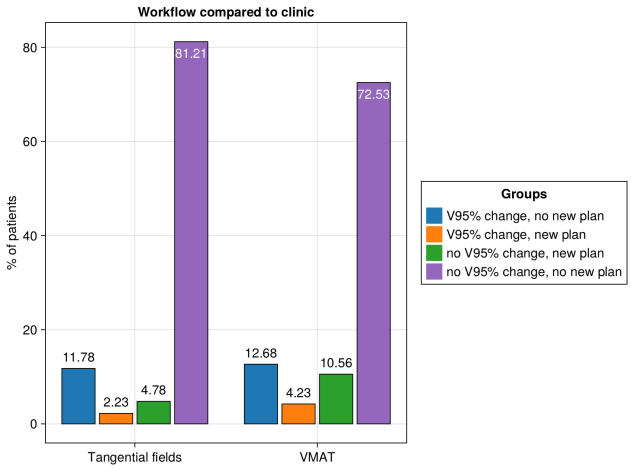


Figure 4: Clinically triggered replanning compared to workflow-initiated check for breast cancer patients with and without lymph node irradiation. We divided the patients into four groups. Either both the workflow and clinic agree and a new treatment plan needs to be made or both agree that the current treatment is sufficiently effective. Or the workflow and the clinic do not agree. The trigger for the workflow-initiated check is a primary CTV coverage lower than 95% of the prescribed dose.

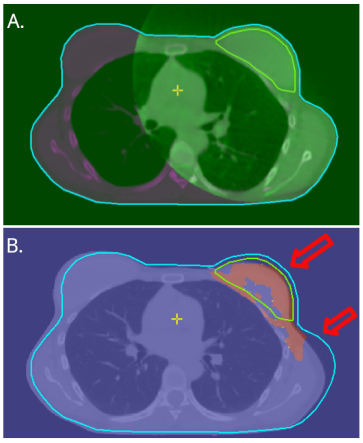


Figure 5: An example of a difference between the daily anatomy and the pCT resulting in a difference in CTV coverage. A.) the daily sCT with the CBCT overlapped in a copper tint with the new body and CTV contours. B.) the pCT with the daily delivered dose clipped on 95% of the prescribed dose overlapped and the same contours as A.). The red coloured areas indicate where the dose is higher than 95% of the prescribed dose and blue areas are below this threshold. The red arrows indicate locations where the body contour has been increased.

the daily anatomy results in underdosage of the CTV. For this patient no new plan was made as this discrepancy between the daily anatomy and the pCT was predominantly there only on the last two out of fifteen fractions. However, this type of anatomical difference has been observed with multiple patients.

4 Discussion

The impact of the developed workflow on the clinic is multi-faceted. Primarily, a better-informed decision to make a new pCT to check if a new treatment plan is required can be made. Mainly because rather than basing this decision on the daily anatomical differences, the daily delivered dose can be incorporated in this decision. Furthermore, because the dosimetric impact of the daily anatomical changes is now known, the radiotherapy technicians who are performing the treatment might inquire clinical physicists and radiation oncologists fewer times, and as such, saving time and unburdening them.
Another impact of the developed workflow is the ability to obtain dosimetric information on a large population. This information can be used to investigate potential sources of errors inside the clinic, either with the patient setup or the plan quality for different plan types.
Ultimately, the biggest impact would be to detect patients with this workflow that would receive suboptimal treatment and have gone unnoticed through the clinic otherwise. Especially for hypofractionated treatments, since being able to check the daily delivered dose and acting upon any under or overdosing will have a large impact.
Currently, there are still some edge cases that this workflow does not handle properly. Examples of such cases are tissue expanders that create an artefact on both the pCT and the daily CBCT. This artefact is delineated for a density override on the pCT, but is not properly propagated onto the sCT. Another example is patients where the target volumes do not entirely fit inside the FoV of the CBCT. This can result in issues with constructing a proper body and target contour. However, these issues can easily be identified through visual inspection of the propagated contours, sCT, and calculated dose.
As for the computation time per fraction, on average about 6 minutes, this can be significantly reduced. The deformable image registration and the dose calculation are already properly optimized, however, propagating and adjusting the contours (i.e. cropping the CTVs) still take several minutes. This is the result of utilizing generic software tools that read and write from disk unnecessarily often for this rather than a custom-made solution that does a single pass over all the contours.
Another improvement of the workflow we are investigating, is the extension to other treatment sites. We already expanded towards head and neck cancer patients, however, the definitive workflow for those patients is still being developed.
Finally, in future work, we continue to move towards a daily adaptive treatment on the CBCT linacs. Utilizing the information that is obtained from this workflow to perform a re-optimization of the treatment plan on the daily anatomy to improve the efficacy of the treatment for each fraction.

5 Conclusion

In this work, we proposed an interfraction dose calculation workflow using CBCTs for breast cancer patients. The daily delivered dose that is calculated using this workflow can be used as additional information to ensure the radiation treatment is effective. This effectiveness can be evaluated on a patient-specific level or on a population level, to draw conclusions for different treatment plan types or to obtain insights into what anatomical changes could cause significant dosimetric changes. Finally, the next step for this workflow is to extend it towards other treatment sites than solely breast cancer.

References

[1] D. Winkel, G. Bol, P. Kroon, et al. “Adaptive radiotherapy: the Elekta Unity MR-linac concept”. *Clinical and translational radiation oncology* 18 (2019), pp. 54–59. DOI: 10.1016/j.ctro.2019.04.001.

[2] V. Hamming, S. Andersson, and J. Maduro. “Daily dose evaluation based on corrected CBCTs for breast cancer patients: accuracy of dose and complication risk assessment”. *Radiat Oncol* 17.205 (2022). DOI: 10.1186/s13014-022-02174-4.

[3] B. D. de Senneville, C. Zachiu, M. Ries, et al. “EVolution: an edge-based variational method for non-rigid multi-modal image registration”. *Physics in Medicine & Biology* 61.20 (2016), p. 7377. DOI: 10.1088/0031-9155/61/20/7377.

Harnessing Radiotherapy Data for Quality Care and Outcome Research

Rishabh Kapoor[1,2], William Sleeman IV[1,2], and Jatinder R Palta[1,2]

[1]Department of Radiation Oncology, Virginia Commonwealth University, Richmond, VA, USA
[2]National Radiation Oncology Program, Veteran Health Administration, Richmond, VA, USA

Abstract The our work presents the development and validation of the Health Information Gateway and Exchange (HINGE) software platform, designed to harness radiotherapy data for quality care and outcome research in oncology. HINGE addresses challenges associated with healthcare data, including volume, variety, velocity, veracity, and value, by offering scalable architecture, diverse data format handling, real-time processing, data quality prioritization, and insightful data analysis. Key design features include data standardization, integration with EHR, Treatment Planning System (TPS), and Treatment Management System (TMS), and a user-friendly interface with disease-specific templates. The platform underwent validation using a comprehensive dataset, showing promising results in reducing physician burden and enhancing data accuracy. Challenges encountered in building HINGE, such as consensus among physicians, clinical workflow variations, and system integrations, were overcome through collaboration and iterative improvements. The platform's deployment architecture ensures reliability, security, and performance, with ongoing evaluation for VA compliance. Overall, HINGE represents a significant advancement in facilitating data-driven decision-making in radiation oncology.

1 Introduction

Manual abstraction, collation, curation, and subsequent analysis of healthcare data for quality and outcome assessment of patient treatments are onerous, expensive, and impractical. Advances in computer storage, computing power, and the ability to electronically mine data from disparate sources (e.g., demographics, genetics, imaging, treatment, clinical decisions, and outcomes) have the power to enable big data research in radiation oncology. Developing an ecosystem to make routine clinical data "Artificial Intelligence (AI) ready" will allow us to generate new knowledge from large scale analysis of historical patients' treatments and outcomes to improve care for the future patients. A fundamental barrier in reaching this vision is the lack of strategic development of IT infrastructures that facilitate data aggregation out of clinical systems via federated or centralized infrastructures for data access. Another barrier is to organize and standardize knowledge and information in a way that is computer understandable with standardized vocabularies, established taxonomies and interoperability standards.

We designed and built an IT infrastructure; Health Information Gateway and Exchange (HINGE) software platform that address the challenges associated with the 5 Vs of healthcare data - Volume, Variety, Velocity, Veracity, and Value [2]. In terms of Volume, the scalable architecture of HINGE allows it to handle the vast amounts of data generated in healthcare, including data from EHRs, Treatment planning and dosimetry and Treatment Management Systems in real time. Regarding Variety, the HINGE software is designed to handle diverse data formats, including structured, unstructured, and semi-structured data. When it comes to Velocity, the IT infrastructure supports data processing and analysis when the data are sent over to the HINGE central server for analysis. For Veracity, the HINGE software prioritizes data quality and reliability using the mapping techniques with established ontology and data standardization efforts of professional societies. Finally, in terms of Value, the HINGE software infrastructure focus on extracting meaningful insights and knowledge from healthcare data. This is done by efficient data analysis and mining techniques like ontology-based search tools that are designed to gather data for cohort analysis, machine learning, outcome assessment, and informed decision making..

2 Materials and Methods

The HINGE software, designed for radiotherapy data management, employs several key design features. In terms of data standardization, it facilitates Extract, Transform, and Load (ETL) processes for disparate clinical, treatment, and process data by deploying 'smart' disease-specific templates in its user interface. These templates enable structured data entry, utilizing disease-specific nomenclatures, and include critical data elements for quality assurance analyses.

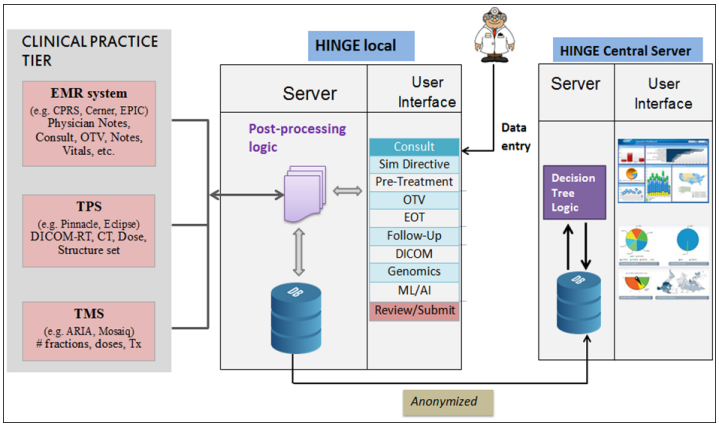


Figure 1 Overview of the architecture of Health Information Gateway and Exchange (HINGE) software platform.

The software integrates with various data sources, including VHA's EHR (VISTA), through an external interface that fetches patient details, ensuring seamless data interoperability. For treatment planning system (TPS) integration, HINGE utilizes the IHE-RO standard, importing DICOM-RT data and adhering to TG-263 standardized nomenclature for target and organ-at-risk structures [1] (Figure 1). To connect with treatment management systems (TMS), a Docker container named HINGE-Broker is implemented, accessing TMS databases and exposing a web API for remote queries. The data retrieved from TMS populates HINGE templates, providing a discretized view of treatment delivery data (figure 2). The software ensures ease of use for physicians through disease-specific templates, designed with UI/UX

considerations to reduce clicks, facilitate lateral data entry, and automate calculations. Validation checks, error messages, and disease-specific fields enhance data entry accuracy. For EHR integration, HINGE utilizes HL7 FHIR standards, playing a crucial role in interoperability profiles such as IHE-RO and CodeX [4]. These profiles allow the extraction of treatment summary information from radiotherapy treatment management systems, integrating it seamlessly with EHRs. The data elements interfaces from the TMS to HINGE templates is shown in Figure 2.

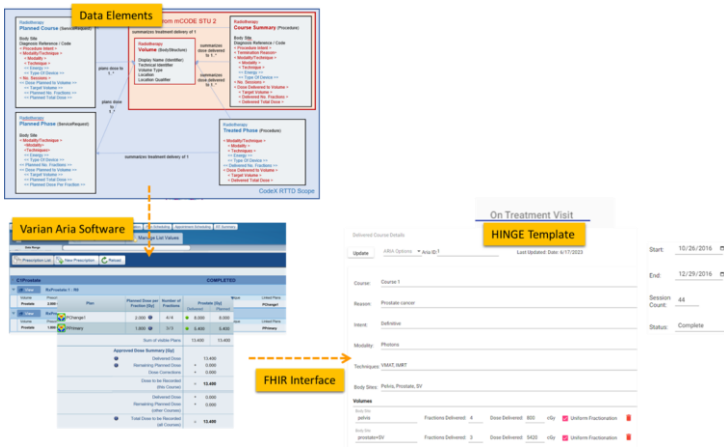


Figure 2: Consensus treatment summary data elements defined with CodeX and IHE-RO effort is implemented in the Varian Aria software and interfaced to the HINGE software via the FHIR interfaces

HINGE Information Technology and Deployment Architecture: The HINGE web applications are deployed on the VA cloud platform hosted by Amazon Web Services (AWS), utilizing ECS containers and Fargate services. MongoDB serves as the backend database for template data storage, Angular web application framework powers the front-end user interface, and AWS S3 is employed for storing medical imaging and treatment planning objects in DICOM format. The architecture, depicted in Figure 3, employs EC2 instances for MongoDB and incorporates essential data backup solutions, two availability zones, and a load balancer system across developmental, pre-production, and production environments. Continuous cybersecurity measures, including vulnerability analysis and remediation, ensure security. The overall cost for hosting and cybersecurity measures is estimated to be less than half a million dollars (FY 2022). This architecture provides a robust and scalable foundation, leveraging AWS services to support reliable, secure, and high-performance operation for radiation oncology practitioners and researchers.

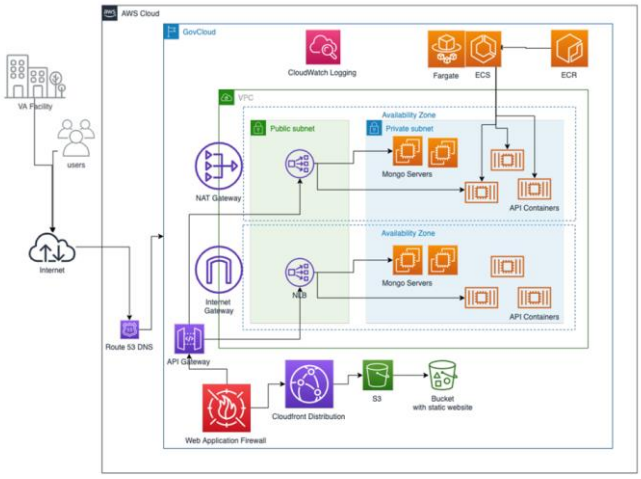


Figure 3: High-level cloud architecture for the HINGE platform

3 Results

The validation of the HINGE platform involved a meticulously collected dataset of 1660 patient clinical and dosimetry records from 40 VA Radiation Oncology clinics, tailored for a pilot study focusing on prostate cancer, non-small cell lung cancer (NSCLC), and small cell lung cancer (SCLC). Each patient's record, comprising 80-100 clinical data elements, was entered into disease site-specific HINGE templates. Leveraging an auto-population feature from electronic medical records, physicians experienced minimal data entry requirements, with the completion of templates and generation of textual narrative notes for initial consultations or follow-up care taking less than 10 minutes based on preliminary analysis. The HINGE platform intelligently populated relevant template sections, utilizing comprehensive patient data from electronic medical records. This streamlined process aimed to reduce the burden on physicians, allowing them to focus more on patient care and decision-making. [3]

To analyze and score the data, the platform employed decision tree logic within a dashboard software (shown in Figure 4). Visual representations, including plots, graphs, and charts, provided insights into the performance of each VA practice for various Clinical Quality Measures (CQMs). [4] In-depth manual checks validated the accuracy and functionality of the decision tree logic, with meticulous analysis and visualization tools used to identify and address discrepancies or issues. The platform underwent further refinement through workshops with VA Subject Matter Expert physicians. Training sessions and iterative improvements based on physician feedback enhanced the quality of templates and generated discrete data. This collaborative approach ensured alignment with the specific needs and workflows of the VA healthcare system.

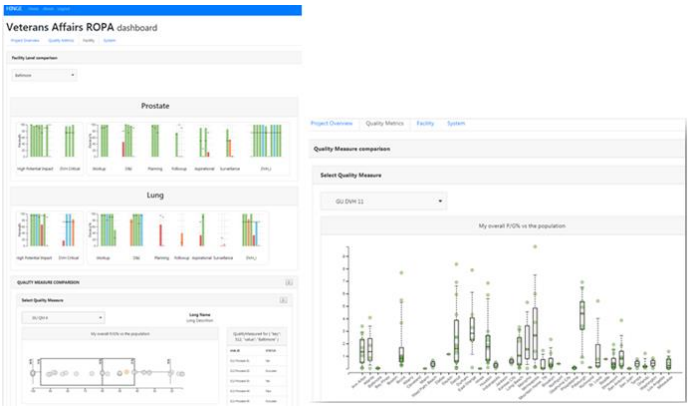


Figure 4: Screen capture of the HINGE dashboard application showing data from 40 VA practices

4 Discussion

The development of the HINGE software for comprehensive radiation oncology data management encountered and successfully addressed several challenges. Achieving consensus among VA subject matter expert physicians posed a significant hurdle due to diverse preferences and practices. The solution involved relying on VA quality measures aligned with evidence-based guidelines and fostering extensive communication to identify and prioritize essential data elements. Variation in clinical workflows across disease sites and clinics necessitated designing adaptable templates, aligning with accrediting bodies' requirements. Integration challenges included harmonizing data syntax and semantics from diverse systems, addressed through an HL7 FHIR interface for EHR integration. Standardizing data from Treatment Planning Systems (TPS) proved challenging, resolved by adopting TG-263 naming conventions and exploring machine learning for automation. The lack of radiation oncology-specific ontologies prompted the deployment of disease-specific templates, mapped to standard ontologies for interoperability. Template usability and user experience were crucial, demanding iterative feedback loops and usability testing to balance comprehensive data capture with user-friendly interfaces. Overcoming these challenges underscores the HINGE software's effectiveness in facilitating robust radiation oncology data management.

5 Conclusion

While the HINGE platform has made significant progress in establishing a robust data collection framework, it is currently undergoing evaluation for compliance with VA cybersecurity standards based on the NIST framework. This rigorous assessment encompasses infrastructure, data protection measures, access controls, encryption protocols, data backups, disaster recovery planning, incident response planning, and vulnerability management. Once deemed compliant, the platform will be deployed within the VA system, providing a secure tool for data entry, analysis, and decision support in radiation oncology.

HINGE notably reduces the burden of data collection, ensuring data integrity by eliminating manual transcription, enhancing traceability, and minimizing errors. The system's versatility extends to creating decision support models for clinicians, fostering improved patient care quality and safety. In the context of value-based healthcare, HINGE quantifies care quality and value, utilizing standardized templates. While alternative approaches, like natural language processing (NLP), face challenges due to the lack of universal standards in radiation oncology documentation, HINGE excels in structured data extraction. Future HINGE releases will integrate Patient Reported Outcomes (PRO), extending data capture to community providers for comprehensive big data outcome research. The paper concludes by advocating collaborative efforts across stakeholders and funding agencies to replicate HINGE's success in other medical domains, emphasizing the importance of defining structured data collection methodologies and clinically meaningful measures for continuous quality improvement.

References

[1] Mayo CS, Moran JM, Bosch W, Xiao Kapoor, R, et.al American Association of Physicists in Medicine Task Group 263: Standardizing Nomenclatures in Radiation Oncology. Int J Radiat Oncol Biol Phys. 2018 Mar 15;100(4):1057-1066. doi: 10.1016/j.ijrobp.2017.12.013. Epub 2017 Dec 15. PMID: 29485047; PMCID: PMC7437157..

[2] Kapoor R, Sleeman WC 4th, Nalluri JJ, Turner P, Bose P, Cherevko A, Srinivasan S, Syed K, Ghosh P, Hagan M, Palta JR. Automated data abstraction for quality surveillance and outcome assessment in radiation oncology. J Appl Clin Med Phys. 2021 Jul;22(7):177-187. doi: 10.1002/acm2.13308. Epub 2021 Jun 8. PMID: 34101349; PMCID: PMC8292697.

[3] Kapoor R, Sleeman WC 4th, Ghosh P, Palta J. Infrastructure tools to support an effective Radiation Oncology Learning Health System. J Appl Clin Med Phys. 2023 Oct;24(10):e14127. doi: 10.1002/acm2.14127. Epub 2023 Aug 25. PMID: 37624227; PMCID: PMC10562037.

[4] Mayo CS, Feng MU, Brock KK, Kudner R, Balter P, Buchsbaum JC, Caissie A, Covington E, Daugherty EC, Dekker AL, Fuller CD, Hallstrom AL, Hong DS, Hong JC, Kamran SC, Katsoulakis E, Kildea J, Krauze AV, Kruse JJ, McNutt T, Mierzwa M, Moreno A, Palta JR, Popple R, Purdie TG, Richardson S, Sharp GC, Satomi S, Tarbox LR, Venkatesan AM, Witztum A, Woods KE, Yao Y, Farahani K, Aneja S,

Gabriel PE, Hadjiiski L, Ruan D, Siewerdsen JH, Bratt S, Casagni M, Chen S, Christodouleas JC, DiDonato A, Hayman J, Kapoor R, Kravitz S, Sebastian S, Von Siebenthal M, Bosch W, Hurkmans C, Yom SS, Xiao Y. Operational Ontology for Oncology (O3): A Professional Society-Based, Multistakeholder, Consensus-Driven Informatics Standard Supporting Clinical and Research Use of Real-World Data From Patients Treated for Cancer. Int J Radiat Oncol Biol Phys. 2023 Nov 1;117(3):533-550. doi: 10.1016/j.ijrobp.2023.05.033. Epub 2023 May 26. PMID: 37244628; PMCID: PMC10741247.

[5] William C Sleeman IV, Sriram Srinivasan, Preetam Ghosh, Jatinder Palta, Rishabh Kapoor; Big data applications in radiation oncology: challenges and opportunities; Artificial Intelligence in Cancer Diagnosis and Prognosis, Volume 3: Brain and prostate cancer; 8-1-8-28; 2022/10/1

TrajectoryLog.NET: Applications of an Integrated Trajectory Log File Analysis API

Matthew Schmidt[1], Frank Marshall[1], Sarah Cradick[1], Justin Mikell[1], Phillip Wall[1], Nels Knutson[1], Tom Mazur[1], and Yao Hao[1]

[1]Department of Radiation Oncology, Washington University in St. Louis School of Medicine

Abstract: Machine log file analysis has shown effective utility within a quality assurance program for both periodic linear accelerator testing and patient-specific plan verification. Limitations of existing machine log analysis implementations include the lack of integration with treatment planning system application programming interfaces and incompatibility with emerging ring delivery system linear accelerator platforms. An application programming interface (API) has been developed, TrajectoryLog.NET, to overcome these limitations. This API was benchmarked against a standard python trajectory log file analysis tool and employed in the creation of two analysis applications – a Respiratory Gating QA Analysis tool and a patient specific delivery parameter report. In benchmarking, all parameters were within 5.0 x 10^{-5} deg and 5.0 x 10^{-6} cm for rotational and translational axes of motion respectively. The API was also used to determine the agreement in gantry angle and MLC position between gated and non-gated QA and in determination of the dose rate hysteresis in a gated beam being an average of 5.5 control points (110 ms).

1 Introduction

Many delivered radiation therapy treatments employ a sophisticated combination of dose delivery with varying Multi-Leaf Collimator (MLC) positions in Intensity Modulated Radiation Therapy (IMRT) or variable dose rate, gantry position and speed, and MLC position in Volumetric Modulated Arc Therapy (VMAT) [1]. With the adoption of modulated treatment planning and delivery, robust quality assurance (QA) programs commonly include treatment-specific measurements in addition to periodic geometric and dosimetric linear accelerator (LINAC) QA tests [2]. Both periodic and treatment-specific QA of a dynamic treatment delivery can include dosimetric measurements on an array or phantom, software calculation-based evaluations, log-file analysis or a combination of any of these methods [3].

Analysis of machine log files (trajectory logs) have been utilized in the evaluation of pre-treatment QA, ongoing during treatment QA, and periodic LINAC QA [4,5]. Modern linear accelerators have the capacity to write a trajectory log describing the beam delivery with control points (CP) spaced every 20ms. Current implementations to the use of trajectory logs for LINAC and patient-specific IMRT QA have limitations in both integration with commercial software solutions and workflow automation. Modern treatment planning systems and QA software implement application programming interfaces (APIs) to facilitate powerful custom analysis application designs and streamlined workflows for the evaluation, reporting, and trending of QA data [6,7]. Our institution implements daily trajectory log file analysis, but differing axes of motion have rendered the application incompatible with Ring Delivery System (RDS) LINAC analysis.

To address these limitations and to extend our institutions current solution to facilitate trajectory log analysis for (RDS) LINACs, a Trajectory Log file reader has been developed and distributed as an open-source API.

2 Materials and Methods

TrajectoryLog.NET is an open-source API built and compiled in C# as a .NET class library. The primary function of the API is to consume and interpret binary Trajectory log files in accordance with the standardized format provided by the linear accelerator vendor (TrueBeam V2.7, Varian Medical Systems, Palo Alto, CA) [8]. The TrajectoryLog.NET API currently exposes the following methods in its early inception:

a) EnableDebug: Allows the client to stream the reading and interpretation of the trajectory log binary as a standard output message.
b) DisableDebug: Disables the debug output.
c) LoadLog: Reads the trajectory log and returns a strongly-typed TrajectoryLogInfo class object. *Requires formal parameter for the file name <string>*.
d) ToCSV: Exports trajectory log file as a comma separated value (CSV) file. *Requires formal parameter for the trajectory log <TrajectoryLogInfo>*.
e) ToPDF: Prints standard PDF output for trajectory log analysis. *Requires formal parameter for the trajectory log <TrajectoryLogInfo>*.
f) BuildFluence: Returns 2 dimensional array of fluence from the MLC positions. *Requires formal parameters for the MLC positions <List<float[]>> and the MLC Model <MLCModelEnum>*.

To validate the interpretation of the Trajectory log files with this API, 77 trajectory log files were acquired from a C-arm LINAC (N=35) and RDS LINAC (N = 42) to test the conversion of 3D-conformal radiotherapy (3DCRT) (N = 8), IMRT (N = 32), and VMAT (N = 37) plans. The API exported the trajectory log results in CSV format and each trajectory log file was compared at each CP with a popular open-source python-based trajectory log analysis toolkit, Pylinac [9].

QA software analysis tools leveraging TrajectoryLog.NET were also developed to meet clinical needs that highlight both the TPS integration advantages and enabling log-file based patient-specific QA for RDS LINACs. First, a software application was developed to analyze images and deliveries with and without Respiratory Gating influenced beam delivery. The application evaluates the impact of gat-

ing on machine output and performance via integrated images and trajectory logs acquired from those gating and non-gating plans. The output variations are reported via a gamma analysis index [10] facilitated through the Portal Dosimetry API (PDSAPI V15.6, Varian), while trajectories for gantry angle position, dose rate, and MLC position are plotted within the application for review.

A second application was developed to report the comparison of the RDS trajectory log file actual treatment delivery with the control point parameters from the TPS (Eclipse V15.6, Varian). This application generates a report comparing planned and delivered trajectories, utilizing a comparison of various axes of motion and fluence comparisons from MLC positions.

3 Results

Of the 77 analyzed treatment field trajectory logs, the gantry angle position, collimator angle position, positions of all 120 MLC leaves (57 leaves for RDS machine configuration), and the meterset value were compared to the log file interpretation from Pylinac for each 20ms control point interval during the treatment. All rotational values (collimator and gantry angles) were within 5.0 x 10^{-5} degrees. Linear (MLC) axes were all within 5.0 x 10^{-6} cm and the meterset values were within 5.0 x 10^{-5}MU. The deviation in gantry angle, meterset, and the central MLC leaves are plotted below to identify any drift in the data from incorrect byte counting or data-type precision differences in the file conversion [Figure 1].

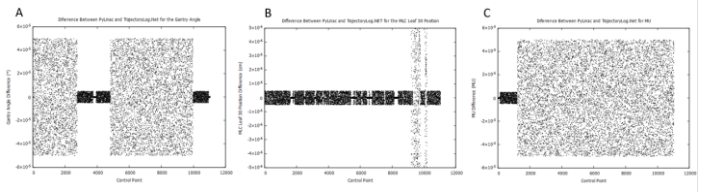


Figure 1: Trajectory log actual value differences when interpreted by TrajectoryLog.NET and Pylinac for (a) Gantry Angle position, (b) Center MLC leaf position, and (c) MU at each control point.

The user interface for the Respiratory Gating QA Analysis application is seen below [Figure 2].

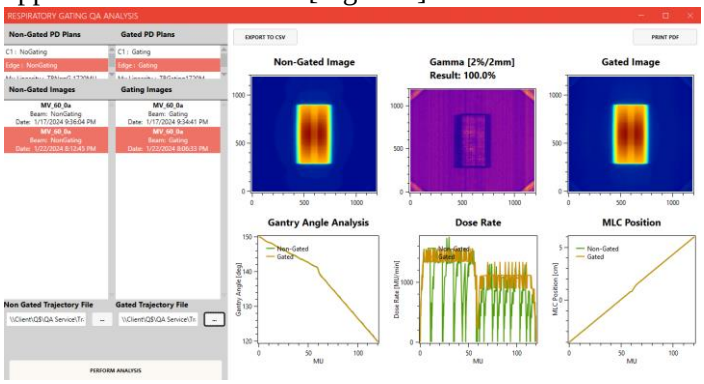


Figure 2: Respiratory Gating QA Analysis utilizing TrajectoryLog.NET for MLC Analysis

The application reported gamma index passing rates are between 87.2% and 100.0% using a 1%/1mm dose difference and distance-to-agreement criteria and 98.0% to 100.0% with 2%/2mm [Table 1]. The gamma analysis is reported with a global dose difference and an region-of-interest defined as the image within a dose threshold of 5%.

	Passing Rate [%]	
TrueBeam 1	1%/1mm	2%/2mm
1/12/2024	98.2	99.5
1/17/2024	87.2	98.0
1/24/2024	92.9	98.1
TrueBeam 2		
1/17/2024	100.0	100.0
1/24/2024	100.0	100.0
TrueBeam 3		
1/17/2024	100.0	100.0
1/124/2024	100.0	100.0

Table 1: Gamma analysis results for 2 Truebeam linear accelerators when acquiring integrated imaging with and without a gated beam.

The Respiratory Gating QA Analysis application can also print an analysis report PDF and export CSV data for further investigation. From this exported data, the gantry angle position of the gated delivery remains within 0.02 degrees of the non-gated, and the central MLC position remains within 0.01 cm. The dose rate of the gated field requires an average of 5.5 control points (110 ms) to return to within one standard deviation of the non-gated field dose rate.

Figure 3 shows the analysis report PDF prototype for a patient-specific log analysis for pre-treatment QA. The report compares various motion axes and reconstructs a fluence from plan parameters and the MLC actual positions during treatment.

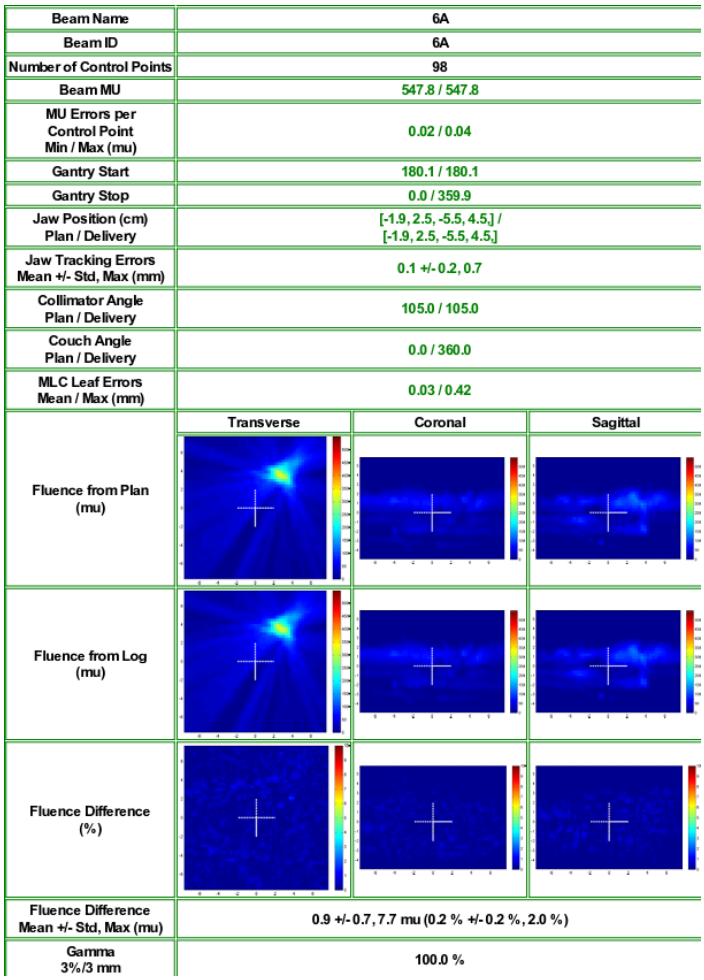


Figure 3: Trajectory log report implementation for patient-specific pretreatment QA and ongoing delivery verification.

4 Discussion

While numerous Trajectory Log analysis tools currently exist, complete integrations with commercial TPS and RDS systems have yet to be explored. Small variations exist within the solutions tested for interpreting trajectory log files, but based on the magnitude of these variations, the cause is more likely a consequence of different implementations of the CSV writing methods as opposed to differences in the data-types between python and C# floating point precision (16-bytes vs 4-bytes for python and C#, respectively). With the increased implementation of workflows leveraging TPS APIs, performing this analysis within the same software framework allows for enhanced integration with clinical tools. This .NET Framework solution is available open-source with installation and use instructions at https://github.com/WUSTL-ClinicalDev/TrajectoryLog.NET, and installed within the nuget package management system for streamlined implementation into Visual Studio.

Two separate applications leveraging the TrajectoryLog.NET API show the APIs immediate utility in both periodic linear accelerator QA and patient-specific QA. Both applications still require further investigation. The Respiratory Gating QA Analysis application would benefit from further exploration of various trajectories and baselining of expected dose rate hysteresis (as seen in the center plot in Figure 2) as the gating window is entered.

Initial implementation of patient-specific trajectory log QA analysis has several benefits to our institution's current solution. With integration of the log analysis into the same framework as the TPS API, DICOM file export is no longer required since the reference trajectories can be acquired directly from planned control points. This simplified workflow also allows for the report implementation as an *on-demand* evaluation during weekly chart checks. Additionally, utilizing an independent API allows for new machine configurations, such as RDS and future linac models, to be included in the API without impacting the client application.

5 Conclusion

TrajectoryLog.NET is a new application programming interface for the consumption and interpretation of binary machine logs acquired during treatment delivery. Trajectory log analysis is a valuable component to quality assurance for various scenarios and can be implemented in an automated fashion with custom software. To show the benefits to integrated trajectory log file interpretation, two sample applications have been developed to evaluate trajectory parameters during a gated delivery and perform basic parameter validation between planned and treated parameters for a new RDS LINAC configuration.

References

[1] Molineu, A., Hernandez, N., Nguyen, T., Ibbott, G. and Followill, D. (2013), Credentialing results from IMRT irradiations of an anthropomorphic head and neck phantom. Med. Phys., 40: 022101. https://doi.org/10.1118/1.4773309

[2] Ezzell, G.A., Galvin, J.M., Low, D., Palta, J.R., Rosen, I., Sharpe, M.B., Xia, P., Xiao, Y., Xing, L. and Yu, C.X. (2003), Guidance document on delivery, treatment planning, and clinical implementation of IMRT: Report of the IMRT subcommittee of the AAPM radiation therapy committee. Med. Phys., 30: 2089-2115. https://doi.org/10.1118/1.1591194

[3] Miften, M., Olch, A., Mihailidis, D., Moran, J., Pawlicki, T., Molineu, A., Li, H., Wijesooriya, K., Shi, J., Xia, P., Papanikolaou, N. and Low, D.A. (2018), Tolerance limits and methodologies for IMRT measurement-based verification QA: *Recommendations of AAPM Task Group No. 218*. Med. Phys., 45: e53-e83. https://doi.org/10.1002/mp.12810

[4] Chuang KC, Giles W., Adamson W. (2020), On the use of trajectory log files for machine & patient specific QA. Biomedical Physics & Engineering Express., 7: https://doi.org/10.1088/2057-1976/abc86c

[5] Dharanipathy, R. et al., (2013), Catching errors with patient-specific pretreatment machine log file analysis. Practical Radiation Oncology, 3(2): 80-90. https://doi.org/10.1016/j.prro.2012.05.002

[6] Hao, Y., Schmidt, M.C., Wu, Y. and Knutson, N.C. (2020), Portal dosimetry scripting application programming interface (PDSAPI) for Winston-Lutz test employing ceramic balls. J. Appl. Clin. Med. Phys., 21: 295-303. https://doi.org/10.1002/acm2.13043

[7] Schmidt, M.C., Raman, C.A., Wu, Y., Yaqoub, M.M., Hao, Y., Mahon, R.N., Riblett, M.J., Knutson, N.C., Sajo, E., Zygmanski, P., Jandel, M., Reynoso, F.J. and Sun, B. (2021), Application programming interface guided QA plan generation and analysis automation. J Appl Clin Med Phys, 22: 26-34. https://doi.org/10.1002/acm2.13288

[8] Varian Medical Systems (2013). TrueBeam 2.7 Trajectory Log File Specifications.

[9] Zaila, A., Adili M., Bamajboor S. (2016), Pylinac: A toolkit for performing TG-142 QA related tasks on linear accelerator. Physica Medica 32(3): 292-293. https://doi.org/10.1016/j.ejmp.2016.07.122

[10] Low, D.A. and Dempsey, J.F. (2003), Evaluation of the gamma dose distribution comparison method. Med. Phys., 30: 2455-2464. https://doi.org/10.1118/1.1598711

PARADIM: a platform to support research at the interface of AI and medical imaging

Yannick Lemaréchal[2], Gabriel Couture[2], François Pelletier[1], Pierre-Luc Asselin[1], Ronan Lefol[1], Samuel Ouellet[1], Leonardo Di Schiavi Trotta[1], and Philippe Després[1,2]

[1]Département de physique, de génie physique et d'optique, Université Laval, Québec (Québec), Canada.
[2]Centre de recherche de l'Institut universitaire de cardiologie et de pneumologie de Québec-Université Laval, Québec, Québec, Canada.

Abstract This paper describes a digital infrastructure designed to support AI research in medical imaging, with a focus on Research Data Management best practices. The platform (PARADIM, Platform for the annotation, reuse and analysis of medical images - *Plateforme d'Annotation, de Réutilisation d'Analyse D'Images Médicales* in French) is rooted in the FAIR principles, through strict compliance with the DICOM standard, and aims at fulfilling several needs arising from the research community at the intersection of AI and medical imaging: robust data curation procedures, responsible data access and effective data governance, robust de-identification and annotation pipelines, and streamlined data operations (e.g. model training, analysis). The platform, built from open-source components, is designed to facilitate and automate the execution of large-scale, computationally-intensive pipelines (e.g. automatic segmentations, dose calculations, AI model training). PARADIM fills a gap at the interface of AI and medical imaging, where data and digital infrastructures were historically not given the attention and resources they deserve.

1 Introduction

Considerable resources have been allocated to advancing AI in recent years. Meanwhile, investments have been relatively scarce in elements fueling AI, namely data and the fabric that supports it, digital infrastructures. Research and development efforts towards data and infrastructure appears essential for AI to fulfill its promises. This paper describes a digital infrastructure designed to support AI research in medical imaging, with a focus on Research Data Management (RDM) best practices. The platform (PARADIM, Platform for the annotation, reuse and analysis of medical images - *Plateforme d'Annotation, de Réutilisation et d'Analyse D'Images Médicales* in French) is rooted in the FAIR principles [1], through strict compliance with the DICOM standard, and aims at fulfilling several needs arising from the research community at the intersection of AI and medical imaging: robust data curation procedures, responsible data access and effective data governance, robust de-identification and annotation pipelines, and streamlined data operations (e.g. model training, analysis). The development of PARADIM is inspired by the data-centric AI movement – stating that models are mature but quality data to train them are lacking – and by MLOps/DataOps approaches where research activities are integrated as much as possible in the operations of the healthcare enterprise. The platform, built from open-source components, is designed to facilitate and automate the execution of large-scale, computationally-intensive pipelines (e.g. automatic segmentations, dose calculations, AI model training). It is also well-suited for federated learning approaches, where models (but not data) are exchanged across institutions [2]. The platform stems from work presented at the 19th ICCR, which featured the use of Orthanc [3] as a research DICOM server [4].

2 Materials and Methods

2.1 Overview

The PARADIM architecture is presented in Figure 1, showing the main technological components of the platform as well as high-level functionalities. The components are deployed in two distinct environments (blue and orange zones in the figures), corresponding to network areas belonging to the medical institutions on one hand and to other service providers on the other (in this case VALERIA, institutional Research IT services at Université Laval). From a legal perspective, at least in the jurisdiction PARADIM is deployed in (Québec, Canada), data containing Personal Health Information (PHI) can only reside in the medical network area; they must be de-identified before they can be transferred outside this domain. The de-identification component of PARADIM (Karnak) therefore lives in this area. De-identified images are then transferred to the main PARADIM ecosystem (orange frame in Figure 1) where they are stored in a DICOM server (Orthanc) [3]. An authentication/authorization layer (Kheops) manages access to images, either through a web portal (identities managed by institutional AzureAD) or through an API (revocable tokens). This component (Kheops) is the main entry point to images, by humans or machines, and allows the creation of image collections that can be enriched with annotations, typically by medical experts contouring regions of interest. The rest of the platform is dedicated to data operations (e.g. model training), which are orchestrated by a custom-built job dispatcher handling containerized applications. Components are deployed in an institutional OpenShift (RedHat) Kubernetes environment, and the majority of them offer a web-based interface and REST API functionalities that are used for orchestration and connectivity.

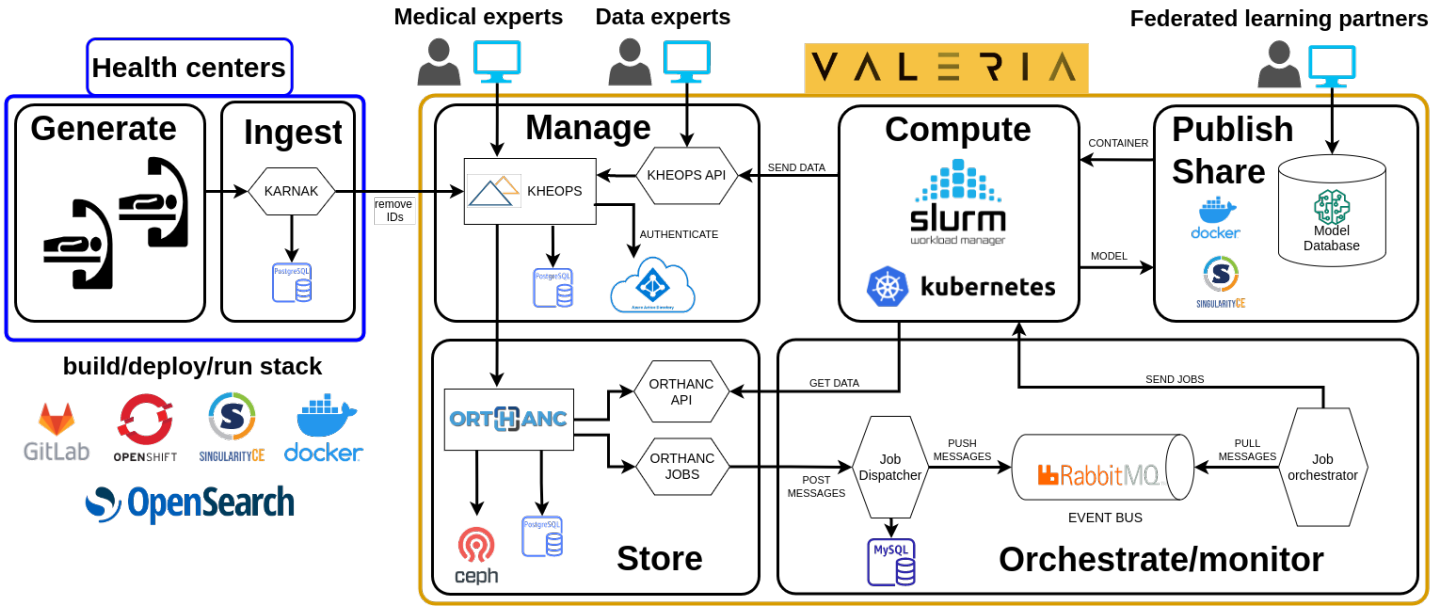


Figure 1: High-level system architecture of PARADIM.

3 Data de-identification

The Karnak software developed by the Osirix Foundation [5] is used in PARADIM as a DICOM de-identification gateway. Karnak provides several functionalities related to data ingestion and de-identification, which is not trivial considering the large number of DICOM tags that can contain PHI. De-identification, as offered by Karnak, typically consists in applying the default de-identification profile defined in the DICOM standard and to add customizations (e.g. keep the original value of a particular field, replace with a fixed value, or shift dates by a random number of days). It is also possible to apply a mask to images, which is useful to remove information engraved in images. De-identification profiles can be defined and reused in Karnak, depending on the requirements of each project (some project might actually need PHI to conduct research activities [6]). Karnak can generate new IDs for patients (based on hash+salt) or use a list provided by users mapping identities. The Clinical Trial Study DICOM module is used to document project information upon ingestion of images in PARADIM (including ERB approval number and type of consent associated with images). This marking of images on ingestion facilitates long-term management with respect to ERB and/or legal approvals.

3.1 Storage backend

Basic data management functionalities in PARADIM are provided by an Orthanc server [3], which natively manages DICOM instances and can be customized and extended in a modular way. In PARADIM, the Python plugin is used to notify the arrival of new studies and series, and to add new endpoints to the existing API. The DICOMweb plugin provides an unified interface to communicate with Orthanc, while the PostgreSQL plugin is used to store the main DICOM tags in a robust database for increased performance. Additional elements were added to the main DICOM tags list to interface with external components such as the OHIF viewer [7]. The AWS S3 plugin was configured to use the Ceph (RedHat) storage provided by our institutional research IT services. DICOM instances stored in Orthanc are immutable; additional information such as annotations, segmentations or derived information are stored in distinct DICOM objects within a study.

3.2 Data governance

Orthanc is well-suited to act as a DICOM data store but has limited functionalities to implement data governance rules, including user management, authentification and authorizations. For these purposes, the Kheops software component developed by the Osirix foundation was used [5]. Kheops uses an album system to constitute image collections, allowing DICOM studies to belong to several albums. Album parameters and metadata generated by Kheops are stored in a dedicated PostgreSQL database. Each album is managed by a data steward who controls access and rights, including adding and removing data or users, and the permission to share or download data. In PARADIM, identities are managed through our institutional AzureAD instance providing single sign-on (SSO) and multifactor authentification (MFA) functionalities. Revocable access tokens can be generated within Kheops to allow services to access specific data collections. Kheops therefore acts as an auth/autz layer on top of Orthanc, with functionalities mainly related to the curation of image collections. An NGINX reverse proxy was set in front of Orthanc to receive network requests and set up Cross-Origin Resource Sharing (CORS) rules that will limit sharing to certain authorized external applications only, including Kheops.

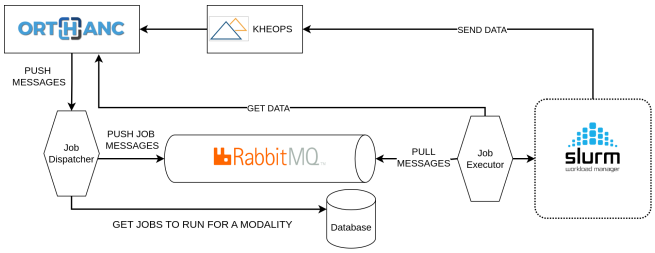


Figure 2: Job dispatching and execution in PARADIM.

3.3 Data analysis

Given the relative complexity of curating data and training AI models, it is important to rely on robust, traceable and ideally automated methods to generate results, even more so that research activities at the interface of AI and medical imaging require significant and expensive computational resources. PARADIM was designed with this in mind, and provides functionalities to automatically trigger compute jobs, and capture execution artefacts (e.g. logs) and metadata on all data operations, from annotations to versions of software used in analysis. These traces are kept in DICOM SR objects and stored along data within studies.

PARADIM users are required to provide their code in application containers that respects the Open Container Initiative specifications. These containers are stored in a private registry, and require only two parameters for execution: one input and one output directories. Internally, user-provided containers are converted to Apptainer images. PARADIM enforces the use of the DICOM format as input and output for all applications. Code snippets and utility functions packaged in libraries (PyOrthanc [8], PyGRPM [9]) are provided to users to ease the capture of execution artifacts and to convert to other formats (e.g. NifTI) if necessary, as some libraries natively require them.

For job execution, two software components were developed: a job dispatcher and a job executor. The first one is connected to the Orthanc API, which triggers contextual messages upon the reception of new images (e.g. a new lung CT study was received). These are sent to a message bus (RabbitMQ) acting as a buffer preventing the overload of compute instances. Messages accumulated in the bus are consumed one by one by the job executor, which 1) gets the relevant patient, study or series associated with the SeriesInstanceUID 2) generates a Slurm script containing the appropriate parameters for job execution. The data and the Slurm script are sent to compute resources and the job executor notifies the job dispatcher that a new job was submitted. Figure 2 illustrates the process.

In addition to jobs triggered by new images received on the platform, it is also possible to launch job manually, either from the command line or through a GUI developed with Streamlit.

3.4 Deployment environment

Components of PARADIM were deployed on our institutional OpenShift (RedHat) platform, using a DevOps-type continuous development and release approach orchestrated by GitLab. Files in YAML format define the different Kubernetes resources to deploy as well as the required resources. It also includes the number of replicas to deploy to ensure high availability, as well as the network configuration, such as the ports to open. All applications can be scaled except the RabbitMQ component to avoid concurrent access problems. Resources can be quickly redeployed or terminated using the GitLab task scheduler. Probes were developed to periodically enquire the state of applications. In case of failure, notifications are sent by email using the SysDig tool provided by our institution. Sensitive variables and information, such as database access, are managed with the GitLab Secrets Manager and dynamically inserted into configuration files during deployment.
Deployed instances, called pods in the Kubernetes nomenclature, are accessible inside the platform through services, and from outside through routes. For example, Orthanc has no external route, and Kheops accesses it through a service. However, Kheops is open to the external network (and users) through routes. The platform lives inside the institutional network, accessible either from campus or via a VPN connection. In all cases, connections are made with the TLS security protocol, with institutional certificates.

4 Results

This section presents use cases of PARADIM, for which the platform was instrumental in automating tasks or in capturing expensive information such as image annotations.

4.1 Annotation pipeline

One of the first use of PARADIM was the capture of image annotations by medical specialists. This information can be considered expensive, in terms of time and expertise, and it is crucial to preserve it on the long term, along with contextual details (e.g. who, when, how). Annotations often act as ground truths in AI project and in this regard they deserve robust methods for preservation and documentation. The current annotation pipeline in PARADIM is based on the use of 3D Slicer [10] and its QuantitativeReporting plugin [11]. 3D Slicer is used to fetch data from the platform (through authenticated DICOMweb transfers), to capture the name of the operator (stored in the ContentCreatorName DICOM attribute), to perform lesion contouring tasks, and to encode everything in DICOM objects containing as much metadata as possible (e.g. dates, software versions). This pipeline was successfully used at our institution to create lesion segmentations on CT for a cohort of approximately 2000 patients in lung cancer. It allows us to keep track of who made the

annotations, when, and in which context. Considering that machines are poised to increasingly perform this type of task in the future, keeping a record on the origin of information is of paramount importance. For the particular case of lung lesion contouring in CT, capturing the annotator identity led us to unveil discrepancies in contouring methods used by different individuals – an important information that can have an impact on results generated.

4.2 Automated tasks

AI models are rapidly becoming commodities that can be consumed from various outlets (e.g. HuggingFace, MONAI), and integrated into complex pipelines. PARADIM facilitates this process and provides functionalities to automate it. As a proof of concept, the TotalSegmentator model [12] was integrated in a pipeline triggered each time a CT study of the thorax is received on the platform. This automatically created DICOM RTSTRUCT objects in the study, with rich metadata (e.g. execution log, software version). A database keeps track of objects created by a particular model, so that subsequent calls will not generate calculations unless the model version (in the container registry) is different. This MLOps/DataOps approach allows for large scale computational endeavours and continuous monitoring of model performance.

4.3 Monte Carlo calculations

Another use case, different from AI-based tasks, is the automatic recalculation of dose maps for brachytherapy cases [13]. The code was provided in a Docker container, stored in the registry, converted to an Apptainer format, and made available in the job dispatcher database. Each RTPLAN modality delivered on the platform triggers a calculation using this container (which also requires RTSTRUCT and CT images to generate results). The entire recalculation pipeline can be re-triggered manually on demand. This kind of automation streamlines the production of results and fosters the transition from research to the real world, through robustness and reproducibility. From a user perspective, the complexity related to storage and computing resources dispatching is masked.

5 Discussion

PARADIM fills a gap at the interface of AI and medical imaging, where data and digital infrastructures were historically not given the attention they deserve. Several improvements are planned for the future, including a fully web-based annotation pipeline (3D Sclicer client currently required), an interface with tools to extract semantic information from radiology reports and better support for whole-slide imaging in digital pathology. Scaling resources on-demand to external providers (cloud bursting) is also planned to deal with large computational tasks.

6 Conclusion

The PARADIM platform was designed to store, manage, enrich and responsibly access medical images in a data science context. It relies on the DICOM standard for maximal interoperability, and facilitates the execution of complex pipelines. It is used by several research teams for various tasks, and keeps track of as much information as possible on data operations, aligned in this regard with the FAIR principles of research data management.

References

[1] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, et al. “The FAIR Guiding Principles for scientific data management and stewardship”. en. *Scientific Data* 3.1 (Mar. 2016). Number: 1 Publisher: Nature Publishing Group, p. 160018. DOI: 10.1038/sdata.2016.18.

[2] L. Mullie, J. Afilalo, P. Archambault, et al. “CODA: an open-source platform for federated analysis and machine learning on distributed healthcare data”. *Journal of the American Medical Informatics Association* (Dec. 2023), ocad235. DOI: 10.1093/jamia/ocad235.

[3] S. Jodogne. “The Orthanc Ecosystem for Medical Imaging”. en. *J. Digit. Imaging* 31.3 (May 2018), pp. 341–352.

[4] P. Després. *Harnessing the potential of data in clinical PACS with an open-source DICOM server*. eng. Publisher: Zenodo. Sept. 2019. DOI: 10.5281/zenodo.3450560.

[5] *OsiriX Foundation*. en.

[6] S. M. Moore, D. R. Maffitt, K. E. Smith, et al. “De-identification of Medical Images with Retention of Scientific Research Value”. eng. *Radiographics: A Review Publication of the Radiological Society of North America, Inc* 35.3 (2015), pp. 727–735. DOI: 10.1148/rg.2015140244.

[7] E. Ziegler, T. Urban, D. Brown, et al. “Open Health Imaging Foundation Viewer: An Extensible Open-Source Framework for Building Web-Based Imaging Applications to Support Cancer Research”. *JCO Clinical Cancer Informatics* 4 (Apr. 2020), p. CCI.19.00131. DOI: 10.1200/CCI.19.00131.

[8] G. Couture, Y. Lemaréchal, and P. Després. *PyOrthanc*. Sept. 2022. DOI: 10.5281/zenodo.7086219.

[9] Y. Lemaréchal, R. Lefol, P.-L. Asselin, et al. *PyGRPM: a Medical physics library containing many utilty functions*. Aug. 2023. DOI: 10.5281/zenodo.8250064.

[10] A. Fedorov, R. Beichel, J. Kalpathy-Cramer, et al. “3D Slicer as an image computing platform for the Quantitative Imaging Network”. en. *Magn. Reson. Imaging* 30.9 (Nov. 2012), pp. 1323–1341.

[11] A. Fedorov, D. Clunie, E. Ulrich, et al. “DICOM for quantitative imaging biomarker development: a standards based approach to sharing clinical data and structured PET/CT analysis results in head and neck cancer research”. en. *PeerJ* 4 (May 2016), e2057.

[12] J. Wasserthal, H.-C. Breit, M. T. Meyer, et al. “TotalSegmentator: Robust Segmentation of 104 Anatomic Structures in CT Images”. *Radiology: Artificial Intelligence* 5.5 (Sept. 2023). Publisher: Radiological Society of North America, e230024. DOI: 10.1148/ryai.230024.

[13] S. Ouellet, Y. Lemaréchal, F. Berumen-Murillo, et al. “A Monte Carlo dose recalculation pipeline for durable datasets: an I-125 LDR prostate brachytherapy use case”. en. *Physics in Medicine & Biology* 68.23 (Nov. 2023). Publisher: IOP Publishing, p. 235001. DOI: 10.1088/1361-6560/ad058b.

pyCERR – A Python-based Computational Environment for Radiological Research

Aditya P. Apte, Aditi Iyer, Eve LoCastro, Sharif Elguindi, Jue Jiang, Jung Hun Oh, Harini Veeraraghavan, Amita Shukla-Dave, Joseph O. Deasy

Department of Medical Physics, Memorial Sloan Kettering Cancer Center, New York, USA

Abstract A Python-native, GNU-GPL copyright software framework, pyCERR, for commonly used radiological image data formats is presented in this work. It enables researchers to organize, access and transform metadata from high dimensional, multi modal datasets to build cloud-compatible workflows for Artificial Intelligence (AI) modeling in radiation therapy and image analysis. pyCERR provides an extensible data structure and a Viewer to allow for rapid prototyping. Various modules are provided to facilitate analyses including metrics for comparing image segmentation, Image Biomarker Standardization Initiative compliant radiomics, radiotherapy Dose Volume Histogram-based features, convolutional neural network-based models for image segmentations, normal tissue complication and tumor control models for radiotherapy and deformable image registration. pyCERR integrates with Jupyter notebooks, GitHub, the Napari viewer, and cloud services. In summary, pyCERR provides an end-to-end radiological image analysis framework to facilitate reproducible research and seamless integration with cloud platforms and clinical systems.

1 Introduction

A computational Environment for Radiological Research (CERR) platform[1], developed in MATLAB programming language has been widely used and adapted for various medical physics applications over the past decade due to its robust and easy to use data structure and visualization capabilities. These include prototyping Intensity Modulated Radiation Therapy (IMRT) optimization algorithms[2], outcomes modeling [3, 4], film dosimetry QA[5], quality assurance of RT and imaging data by leading quality assurance centers[6], CNN-based training and inference[7]. CERR's flexible data structure has been extended by several plugins such as segmentation benchmarking software developed by AAPM TG 211 [8]. While CERR has been tremendously useful in clinical research, it is limited by the requirement of MATLAB license for development. This is especially crucial in cloud-environments where open-source programming languages offer ease of use as well as readability of the underlying codebase. Python has emerged as the language of choice for research, serving as a glue language for various software libraries. Python's scientific computing and visualization libraries now offer functionalities that are on par or better than MATLAB and its toolboxes for medical image processing.

pyCERR maintains the flexible, readable data structure and graphical user interfaces of CERR platform while providing relevant tools in a Python programming environment, enabling access to cutting edge AI modeling and image processing libraries for local and cloud friendly workflows.

2 Materials and Methods

The components of pyCERR can be broadly categorized as (1) Data container (2) Visualization and (3) Analysis modules.

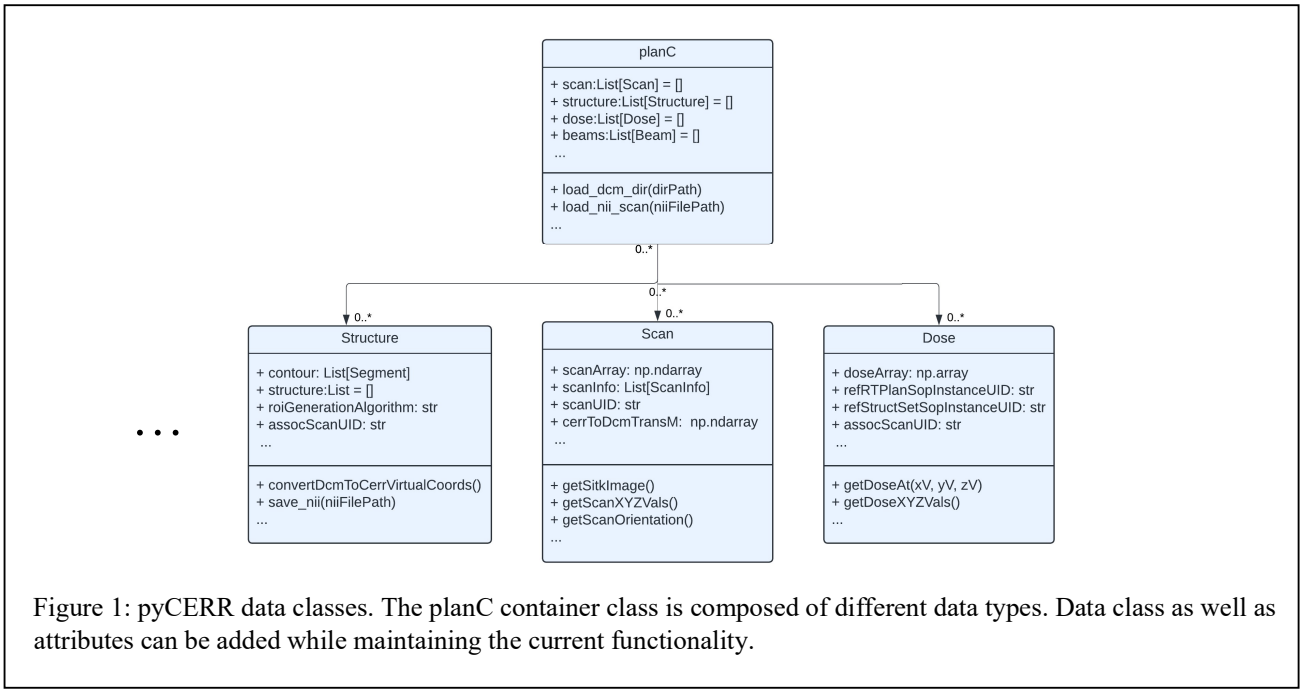


Figure 1: pyCERR data classes. The planC container class is composed of different data types. Data class as well as attributes can be added while maintaining the current functionality.

Data container: The central idea of pyCERR is to provide consistent data structure for metadata from different file formats. The plan container class planC houses data classes containing attributes for various data types. These include (a) Scan for imaging data such as CT, MR, PET, (b) Structure for image segmentation, (c) Dose for RT dose, (d) Beams for RT Plans as well as derived data types such as deformable registration. Figure 1 shows schematic of data classes including some of their attributes and routines for data access and transformation.

Modules are provided for importing and exporting imaging data between DICOM and NifTI formats as well as conversion to SimpleITK[9] Image object. Pydicom[10] is used to read and write DICOM, Nibabel[11] to write to NifTI and SimpleITK to read ITK-compatible formats such NifTI, MHA, NRRD. Data attributes in planC container can be readily accessed for analyses, edited and saved to DICOM and NifTI file formats.

pyCERR currently imports metadata from DICOM RTSTRUCT and SEG segmentation modalities, scan modalities CT, MR, PET and radiotherapy dose modalities RTDOSE and RTPLAN. pyCERR also supports import of Scans, Segmentations and Doses stored in various research friendly formats such as NifTI, NRRD, MHA etc. to planC container. Segmentation is stored as polygonal contours and is converted to as binary mask and vice-versa. pyCERR provides utilities to convert PET images to SUV units using normalizations such as Body Weight, Body Surface Area, Lean Body Mass [ref] approaches. pyCERR supports time series dynamic scan sequences such as DCE MRI as well as diffusion weighted imaging. Image volumes are organized by time or b-values. MRI images are converted to real world units using vendor-specific private tags.

Visualization: Multi-dimensional image viewer Napari[12] is perfectly suited for pyCERR's multi-modal data structure. As Napari is written in Python programming language, it saves users from passing data back and forth between viewers in other languages such as ImageJ[13] and 3D Slicer[14]. Napari ecosystem consists of fast-growing list of plugins geared towards image analysis. These include microscopy datasets and Segment Anything Model[15]. To our knowledge, pyCERR is the only software platform that enables visualization of radiotherapy data in Napari. The viewer module in pyCERR provides routines to visualize scan, segmentation, and dose data types. Currently, pyCERR provides widgets to window scans and to segment them by leveraging Napari's labeling tools.

Analysis modules: Organization of imaging data in plan container enables efficient access to the underlying data elements and their relationships for building modules for analyses. Modules for data transformation, radiomics, dosimetry features, segmentation evaluation and outcomes explorer are included with pyCERR. Data transformations include tools for post-processing results of image segmentation such as smoothing contour polygons, connected components filtering, resampling and copying segmentation and doses from one scan to another. Radiomics module in pyCERR provides IBSI1 compliant scalar features and IBSI2 compliant filters. Feature and filter computation is entirely Python-based and makes use of vectorized code using Numpy arrays for speed. Filter response of IBSI2 filters is stored in planC which facilitates visualization and analysis. pyCERR provides dose volume histogram (DVH) based features such as Vx, Dx, MOHx, MOCx.

3 Results

pyCERR package is distributed under an open-source, GNU GPL license at https://github.com/cerr/pyCERR. pyCERR can be installed from the GitHub repository by using Python package installer pip. Tests for various modules including IBSI compliant feature calculation, conversion of data between various formats are distributed with pyCERR. These are run at every code update to GitHub repository. Notebooks demonstrating various use cases are provided at https://github.com/cerr/pyCERR-Notebooks to provide a starting point for analyses. Following are some common use cases:

(1) Copying segmentation from ome image to other: In image analysis applications involving feature extraction, AI and multi-modal images, segmentation from one modality is used on another scan in the same physical space. pyCERR provides efficient routines to copy over segmentation between scans.

Define Paths for scans and segmentation

```
origNiiScan = r'path/to/original_scan.nii'
origNiiStr = r'path/to/original_seg.nii'
resampNiiScan = r'path/to/smore_resampled_scan.nii'
outputNiiStr = r'path/to/output_resampled_seg.nii.gz'
```

Import scans and segmentation to planC

```
from cerr import plan_container as pc
planC = pc.load_nii_scan(origNiiScan)
planC = pc.load_nii_scan(resampNiiScan, planC)
assocScanNum = 0 # index of the original scan
structName = 'tumor'
planC = pc.load_nii_structure(origNiiStr, assocScanNum, planC, labels_dict={1: structName})
```

```
print("Number of scans:" + str(len(planC.scan)))
print("Number of structures:" + str(len(planC.structure)))
```

Copy segmentation from the original scan to the resampled scan

```
from cerr.dataclasses import structure as cerrStruct
structNum = 0
scanNum = 1
planC = cerrStruct.copyToScan(structNum, scanNum, planC)
print("Number of structures:" + str(len(planC.structure)))
```

Figure 2: Notebook demonstrating copying segmentation between scans.

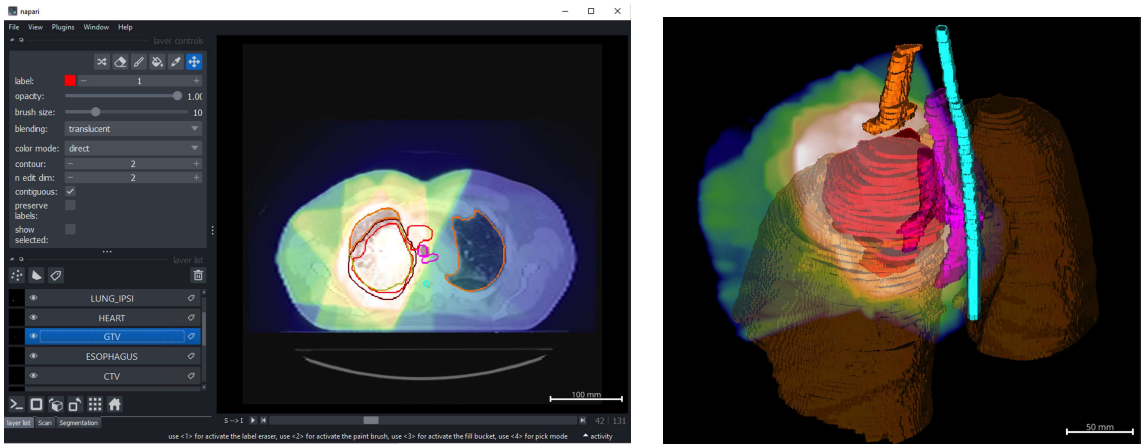


Figure 3: Radiotherapy Does distributions can be visualized along with structures in 2D and 3D. Display settings such as Window, transparency, gamma can be adjusted from GUI or command, thus making it suitable for capturing screenshots in batch mode for large number of datasets.

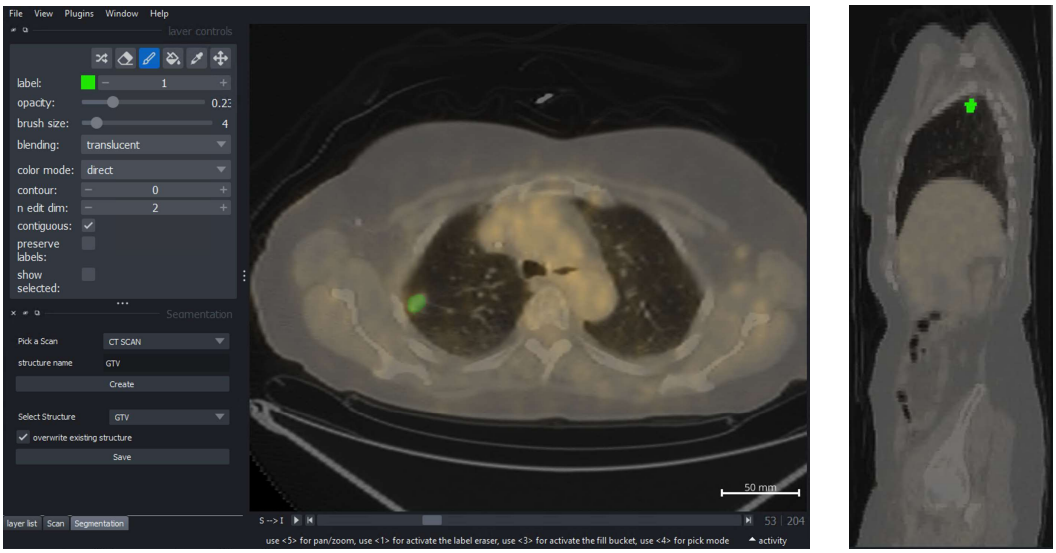


Figure 4: Contouring tools in pyCERR. Segmentation Widget provides the ability to create and edit structures. This is useful to edit contours resulting from auto-segmentation as well as to generate new structures. Napari provides Paintbrush, Eraser and Fill tools for editing segmentation / labels.

(2) Visualizing a batch of datasets

pCERR's Viewer module provides routines to overlay imaging data from the planC container such as scans, doses and structures as demonstrated in figure 3. These routines can be used locally in an interactive mode or on a remote server to capture screenshots. https://github.com/cerr/pyCERR-Notebooks/blob/main/batch_visualize_scan_seg_ex1.ipynb notebook demonstrating capturing screenshots for a batch of DICOM datasets on Google Colab platform.

(3) Image segmentation:

Manual segmentation is required to generate ground truth data for model training as well as editing the output of AI segmentation for downstream use. pyCERR provides segmentation widget which can be used on all of the orthogonal views while overlaying multiple images such as PET and CT as demonstrated in figure 4.

(4) Radiomics and DVH-based metrics calculation:

pyCERR provides "*radiomics*" and "*dvh*" modules to support feature extraction from scan and dose datasets. The *radiomics* module provides routines for IBSI-1 and IBSI-2 compatible feature extraction. The dvh module provides ability to compute commonly used metrics with user defined dose bin width.

```python
import os
from cerr import plan_container as pc
from cerr.radiomics import ibsi1

featList = []
idDict = {}
writeHeader = True
for pt_dir in all_pat_dirs[:5]:
    _, id = os.path.split(pt_dir)
    print("Data dir :" + id)
    planC = pc.load_dcm_dir(pt_dir)
    scanNum = 0
    structNum = 0
    featDict = ibsi1.computeScalarFeatures(scanNum, structNum, settingsFile, planC)
    idDict['id'] = id
    featDict = {**idDict, **featDict}
    featList.append(featDict)
    ibsi1.writeFeaturesToFile(featDict, csvFileName, writeHeader)
    writeHeader = False
```

(a)

```python
from cerr import dvh
structNum = 0
doseNum = 0
dosesV, volsV, isErr = dvh.getDVH(structNum, doseNum, planC)
binWidth = 0.025
doseBinsV,volHistV = dvh.doseHist(dosesV, volsV, binWidth)
percent = 70
dvh.MOHx(doseBinsV,volHistV,percent)
```

(b)

Figure 5: (a) Radiomics feature extraction for a batch of datasets: The ***radiomics*** module provides flexibility of selecting scan and segmentation objects from planC as well as specification of settings via JSON configuration file (b) DVH feature extraction: The ***dvh*** module provides routines to obtain RT dose within the segmentation, accumulation of dose values into histogram and calculation of scalar metrics.

4 Discussion

Napari is a relatively new project compared to more established viewers such as ITK-SNAP[16] and 3DSlicer. While the underlying Napari features are being continually updated, it is the only Python-based viewer with such as a rich feature set. Napari being a generic image viewer, supports visualization of various data types such as shapes, vectors, labels, images and allows defining custom callback events. This enables users to customize pyCERR's visualization tools as per their use cases. Additionally, pyCERR will be able to make use of incremental features on Napari's development roadmap including support for orthogonal view planes, grouping of display layers and viewing remote datasets.

5 Conclusion

pyCERR provides comprehensive tools for radiological and radiotherapy research, and leverages open-source formats and tools, including a rapid prototyping framework to train and infer from AI models. pyCERR hence facilitates reproducible research.

References

1. Deasy JO, Blanco AI, Clark VH. CERR: a computational environment for radiotherapy research. Med Phys. 2003;30(5):979-85. Epub 2003/05/30. doi: 10.1118/1.1568978. PubMed PMID: 12773007.
2. Tiwari PB. Automating intensity modulated radiation therapy treatment planning by using hierarchical optimization2015.
3. Chao KSC, Deasy JO, Markman J, Haynie J, Perez CA, Purdy JA, Low DA. A prospective study of salivary function sparing in patients with head-and-neck cancers receiving intensity-modulated or three-dimensional radiation therapy: Initial results. International Journal of Radiation Oncology Biology Physics. 2001;49(4):907-16. doi: Doi 10.1016/S0360-3016(00)01441-3. PubMed PMID: WOS:000167327200001.
4. Deasy JO, Moiseenko V, Marks L, Chao KC, Nam J, Eisbruch A. Radiotherapy dose–volume effects on salivary gland function. International Journal of Radiation Oncology* Biology* Physics. 2010;76(3):S58-S63.
5. Deasy J, Apte A, editors. Data analysis tools for 3D dosimetry: the use of CERR as a platform to integrate and compare measurements and treatment planning information. Journal of Physics: Conference Series; 2010: IOP Publishing.
6. Ulin K, Yorty J, Hanusik R, Urie M, Bosch W, Apte A, Khullar D, Deasy J, Fitzgerald T. Use of CERR at the quality assurance review center to assess protocol compliance of radiation therapy treatment plans submitted in digital format. International Journal of Radiation Oncology, Biology, Physics. 2008;72(1):S673.
7. Apte AP, Iyer A, Thor M, Pandya R, Haq R, Jiang J, LoCastro E, Shukla-Dave A, Sasankan N, Xiao Y, Hu YC, Elguindi S, Veeraraghavan H, Oh JH, Jackson A, Deasy JO. Library of deep-learning image segmentation and outcomes model-implementations. Phys Medica. 2020;73:190-6. doi: 10.1016/j.ejmp.2020.04.011. PubMed PMID: WOS:000534270200023.
8. Shepherd T, Berthon B, Galavis P, Spezi E, Apte A, Lee J, Visvikis D, Hatt M, De Bernardi E, Das S. Design of a benchmark platform for evaluating PET-based contouring accuracy in oncology applications. European Journal of Nuclear Medicine and Molecular Imaging. 2012;39:S264-S.
9. Lowekamp BC, Chen DT, Ibáñez L, Blezek D. The design of SimpleITK. Frontiers in neuroinformatics. 2013;7:45.
10. Mason DL, et al, pydicom: An open source DICOM library. pydicom: An open source DICOM library.
11. Brett M, Markiewicz CJ, Hanke M, Côté M-A, Cipollini B, McCarthy P, Jarecka D, Cheng C, Halchenko Y, Cottaar M. nipy/nibabel: 3.2. 1. Zenodo. 2020.
12. napari: a multi-dimensional image viewer for Python.
13. Abràmoff MD, Magalhães PJ, Ram SJ. Image processing with ImageJ. Biophotonics international. 2004;11(7):36-42.
14. Kikinis R PS, Vosburgh K. 3D Slicer: A Platform for Subject-Specific Image Analysis, Visualization, and Clinical Support. In: Jolesz F (eds) Intraoperative Imaging and Image-Guided Therapy. 2014. doi: https://doi.org/10.1007/978-1-4614-7657-3_19.
15. Ma J, He Y, Li F, Han L, You C, Wang B. Segment anything in medical images. Nature Communications. 2024;15(1):654.
16. Yushkevich PA, Gao Y, Gerig G, editors. ITK-SNAP: An interactive tool for semi-automatic segmentation of multi-modality biomedical images. 2016 38th annual international conference of the IEEE engineering in medicine and biology society (EMBC); 2016: IEEE.

A system of governance for use of scripts in a commercial treatment planning system

Kenton Thompson[1], Keith Offer[2], Glen Osbourne[1], Catherine Lawford[2], Alan Turner[1], Sandro V Porceddu[3,4,5], and Tomas Kron[2,4,6]

[1]Department of Radiation Therapy, Peter MacCallum Cancer Centre, Melbourne, Australia
[2]Department of Physical Sciences, Peter MacCallum Cancer Centre, Melbourne, Australia
[3] Department of Radiation Oncology, Peter MacCallum Cancer Centre, Melbourne, Australia
[4]Sir Peter MacCallum Department of Oncology, University of Melbourne, Melbourne, Australia
[5] University of Queensland, Faculty of Medicine, Queensland, Australia
[6]Centre for Medical Radiation Physics, University of Wollongong, Wollongong, Australia

Abstract Scripting has become common practice to enhance radiotherapy treatment planning and extract additional information for decision making and research. However, scripts are not part of the commercial planning software that has been approved for clinical use by an appropriate authority. As such we have developed a process that requires scripts to be discussed by a multidisciplinary group, undergo appropriate testing and be signed off by a member of the department's executive.
Since 2019 sixteen scripts have been authorized for use in our Varian Eclipse treatment planning system. While most of them are 'plug in' and 'read only' the process has also allowed for the use of stand-alone scripts and scripts that can write to the Eclipse database. The governance process developed encourages participation of all professional disciplines and ensures all scripts that can impact the treatment plan are documented and verified.

1 Introduction

Modern radiotherapy (RT) treatment planning is a complex process involving many steps that often are institution specific. Commercial RT treatment planning systems (TPS) allow all required steps to be performed, however they often require numerous manual interactions, such as mouse clicks, keystrokes and data entries. As such, work can be streamlined using scripting, which also helps standardising planning procedures and reduces the risk of errors. It is also an integral part of full automation of treatment planning classes.[1, 2]
Additional benefits of scripting can result in data analytics which is a powerful tool to check and document practice and evaluate its effectiveness.
However, there are risks introduced by the use of scripts - in particular if they are developed in-house and there is an indication that increasing reliance on computers reduces random errors but may increase systematic ones.[3] RT planners may not know all the available scripts and their functionality and limitations. Running scripts can slow the system down or make some data temporarily inaccessible. Also the output of the script must be verified under various conditions to ensure the results are reliable. In some cases TPS offerings scripts can not only read but also write to the database, which introduces additional risks such as incorrect data being written or corrupting information in the data base. As such, it is important to develop a governance system that can ensure that scripts:

- have functionality relevant for the department
- are tested prior to implementation
- staff appropriately trained
- actions documented
- have their use monitored.

The present paper describes the governance system used in our large multi-campus RT department to develop, implement and monitor script use in a commercial TPS.

2 Materials and Methods

Our institution uses Varian Eclipse (Palo Alto, CA) for external beam radiotherapy planning for Varian linear accelerators (linacs) operating across 5 campuses. Approximately 70% of the 7000 new patients treated annually receive VMAT based treatments with a significant number of stereotactic treatments (approximately 500). The TPS is cloud based and introduced across the campuses in 2018.
All images used for planning are imported and archived in a PACS system (Vue, Philips, Amsterdam, NL)), the plans are transferred to the linacs using ELEKTA MOSAIQ (Stockholm, SE) and patient clinical details are entered into EPIC electronic medical records system. Our department has an active research program and participates in clinical trials that can benefit from data extraction.

While scripting had been conducted in an ad-hoc basis for many years, with the introduction of the cloud-based system in 2018 it was decided to put a standardised governance structure to minimise risk and ensure sufficient oversight. To achieve this goal, a scripting governance committee (SGC) was founded with representation from clinicians, physicists, therapists and planners. The committee initially met at least six times a year and now meets on demand to review requests for scripts, monitor progress of script development and organise and review testing. The committee is chaired by the clinical coordinator radiation therapist, with all scripts requiring the final validation report to be signed off by the

director of physical sciences before deployment onto the clinical system.

The workflow is illustrated in figure 1.

As part of continuous monitoring to provide feedback to the committee on usage, all scripts log their use and additional information such as the patient they were run on and the user who ran the script as well and time and date.

3 Results

The process of introducing a script in the planning environment consists of several steps:

- Request for a new script or functionality to the SGC by any member of the department.
- Review of the request by SGC, assessment of feasibility and prioritization.
- Assignment of staff for script testing and implementation
- Scoping of the script to specify exact functionality and behaviour
- Realisation of the script and installation in a non-clinical Eclipse environment for testing
- Testing and validation
- Review of the testing results and sign off by the director of Physical Sciences
- Deployment of the script to the clinical Eclipse environment

The frequency of SGC meetings is determined by the number of scripts requested and in development.

At present 16 scripts are accessible within the planning environment with the vast majority requested by radiation therapists who are also the profession performing treatment planning. Scripts are categorized by their functionality and their interaction with the TPS and its database. In Eclipse scripts are classified into two main categories of scripts:

1. Plugin scripts can access a single patient's worth of data at a time
2. Standalone scripts can access all patients in the database

Further, scripts are read-only by default, but can be marked as 'write-enabled' to allow them to change some parts of the data within the TPS for the open patient, such as changing contours, fields, plans etc. These different categorisations allow for scripts to be given the minimal permissions required for operation, limiting risk.

Fourteen of the 16 scripts are plug in and as such only affecting one patient at the time. Of these three do not just read data but also have the capability of writing data such as deleting empty contours or creating new contours based on geometric operations from existing contours or dose distribution.

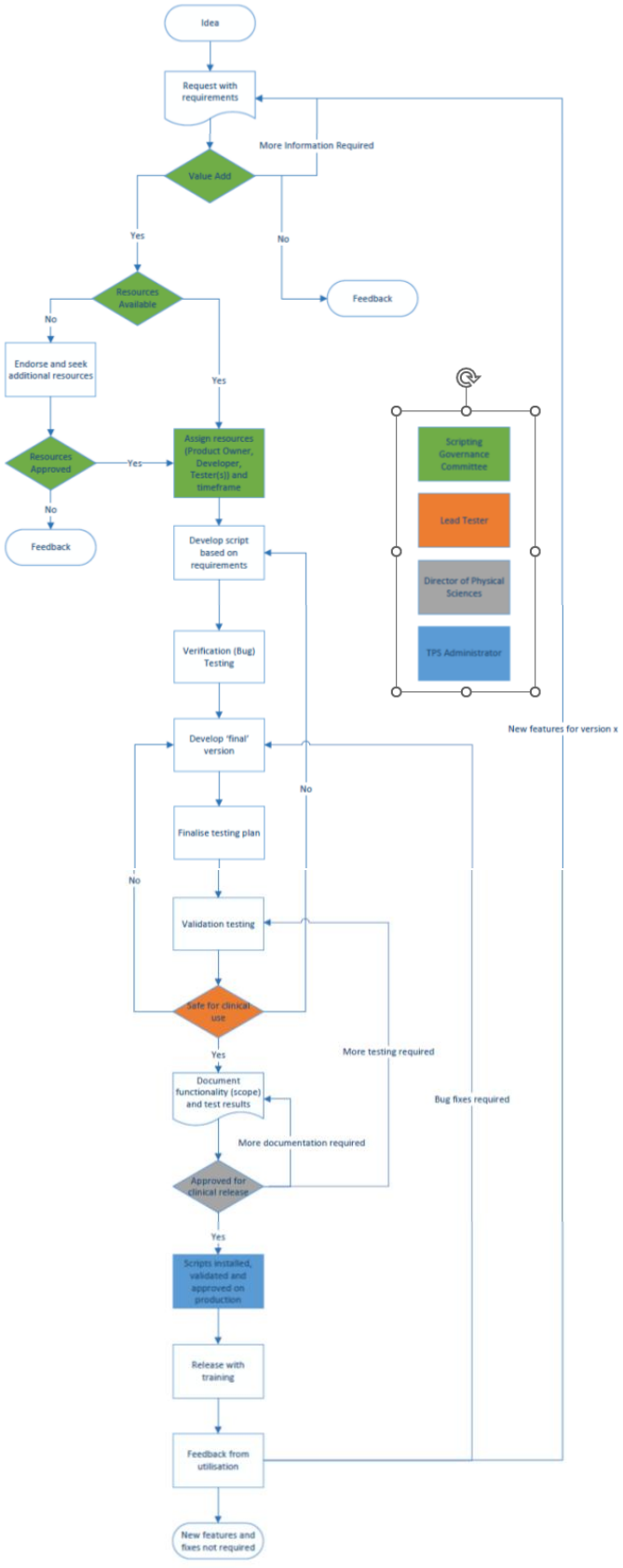


Figure 1: Workflow for script development

The two stand-alone scripts operate on multiple patients and plans based on a list of plans provided by the user. These are powerful tools for quality improvement and research. Examples are exporting dose, plan and structure metrics or planning objectives.

The use of all scripts is monitored both to provide feedback on script usage and estimation. It was found that monitoring execution time of scripts as a surrogate for system load was poor, as often scripts were left running after calculations were complete.

Monitoring use revealed that scripts have been run more than 225,000 times across over 18,000 patients by over 400 users. Scripts are most often run by radiation therapists (~76%) and physicists (~23%) and a small number of executions by radiation oncologists (<1%). The scripts executed the most often are those that aid plan checking [4] and comparing plans against internal protocols.

4 Discussion

Governance in clinical processes has always been an important part of practice. There has been awareness of this in radiotherapy for some time [5] and the recent growth of machine learning and artificial intelligence has put this questions firmly on the agenda [6]. However, there is relatively little published about governance of in-house developed software beyond the validation stage.

Shortcomings of our current processes pertain to the limited involvement of clinicians and the lack of structured monitoring of the performance of scripts in the case of software and hardware changes. In a busy clinical environment, it is also difficult to find the time for the SGC to meet and there is concern about the length of time required to bring something that is ‘obviously needed’ into the clinical environment.
We are considering a more formal risk assessment process [7] to inform the testing of scripts. However, on balance, even ‘obviously needed’ scripts need to be validated to a sufficient depth to ensure correct operation.

5 Conclusion

An important tool in the automation of processes is scripting. We have developed a process that formalises script development and employment in a clinical TPS and provides a governance structure for associated processes. There is no doubt that scripting can improve efficiency and standardization of radiotherapy treatment planning and typically also leads to a reduction of errors. However, given the potential risks a robust system of governance is required to control, validate, and monitor the development and use of scripts.

References

[1] Teruel JR, Malin M, Liu EK, McCarthy A, Hu K, Cooper BT, et al. Full automation of spinal stereotactic radiosurgery and stereotactic body radiation therapy treatment planning using Varian Eclipse scripting. J Appl Clin Med Phys. 2020;21:122-31.
[2] Wang X, Li C, Li G, Zhao J, Bai S. A Script-Based Automatic Intensity Modulated Radiation Therapy Planning Method With Robust Optimization for Craniospinal Irradiation. Pract Radiat Oncol. 2023;13:e209-e15.
[3] Fraass BA. Impact of complexity and computer control on errors in radiation therapy. Annals of the ICRP. 2012;41:188-96.
[4] Covington EL, Chen X, Younge KC, Lee C, Matuszak MM, Kessler ML, et al. Improving treatment plan evaluation with automation. J Appl Clin Med Phys. 2016;17:16-31.
[5] McNee SG. Clinical governance: risks and quality control in radiotherapy. Br J Radiol. 2001;74:209-12.
[6] Stogiannos N, Malik R, Kumar A, Barnes A, Pogose M, Harvey H, et al. Black box no more: a scoping review of AI governance frameworks to guide procurement and adoption of AI in medical imaging and radiotherapy in the UK. Br J Radiol. 2023;96:20221157.
[7] Cardan RA, Covington EL, Popple RA. Code Wisely: Risk assessment and mitigation for custom clinical software. J Appl Clin Med Phys. 2021;22:273-9.

QuAAC: An interoperable routine quality assurance data format

James Kerns[1] and Randle Taylor[1]

[1]Radformation, New York, NY, United States of America

Abstract A new data format is proposed for the interoperability and exchange of routine machine quality assurance data between in-house and commercial software. The format is flexible for multiple workflows, is human-readable, and uses common standards available today.

1 Introduction

Routine machine quality assurance (QA) is a constant consideration for the clinician. Recommendations from governing bodies [1, 2] include tests that should be performed at various frequencies. Depending on the complexity of the QA, the clinician may be interacting with multiple softwares to complete it, whether a simple spreadsheet or dedicated applications for advanced tasks such as image analysis or statistical process control. This data can also be in multiple formats such as simple text files, spreadsheets, or databases.

The clinician is often faced with the task of organizing this QA data for one of several purposes such as preparing for an audit by governing authority, creating an off-site data backup, or refactoring the QA program as would be the case when introducing a new application or workflow to the QA process. This process can be daunting considering the above issues as well as the cognitive load.

This work aims to present a new data format specifically aimed at QA data called the Quality Assurance Archive & Communication (QuAAC) standard. The goal is to provide a consistent representation of QA data that can be used for the purposes of backup, communicating from one in-house QA application to another, transitioning from an in-house QA flow to a commercial one, or transitioning from one commercial system to another. Each of these tasks can be difficult given the myriad data paradigms employed. In a similar manner to how the digital imaging and communication in medicine (DICOM) standard brought a standardization for communication, this format also aims to bring those same benefits for interoperability of QA data.

2 Format Specification & Design

2.1 Rationale

The project was born out of the need to store and exchange routine QA data in a vendor-neutral format, much as DICOM did [3–5]. Clinics that move between commercial vendors of QA equipment and software are often faced with the challenge of converting their data from one paradigm to another, depending on the vendor. This can also happen when migrating from an in-house solution to a commercial solution. There currently exists no reasonable standard for the storage and exchange of this QA data. This project aims to fill this gap.

One of the gaps is that of storing "interpreted" data as well as raw data. For example, when evaluating image quality metrics for a computed tomography (CT) unit such as Houndsfield (HU) uniformity, the metrics are almost always based on digital data that was acquired (such as a scan of a phantom). The clinician then interprets, either manually or via a software application, what the metric value is. In this case, both the raw data and "interpreted" data are important and should nominally be stored and referenced together. This is not limited to DICOM data however. Some QA may require manual inputs for mechanical tests, or printouts of manual safety checks such as a photocopy of a checklist performed by service personnel. The raw data in these cases may be text, screenshots, or other binary data that is important for the proper documentation of the QA performance.

2.2 Use Cases

The use cases for such a standard are many and revolve around the concept of measurements to track the constancy and performance of a machine along with any relevant metadata and ancillary data that such data is based upon. Wheras DICOM is patient-centered, this standard is unit-centered. Some examples of data the standard is meant to represent are:

- Manual measurements of radiation output (ion chamber readings, kVp measurements)
- Imaging metrics based on a DICOM image (HU uniformity, scaling)
- Manual measurements of mechanical performance (couch flex, laser drift)
- Safety output results from simulated failures (door interlocks, emergency stops)

These examples demonstrate a wide range of measurements, both manual and digital, with and without associated data that need to be recorded as part of a routine QA workflow.

2.3 Principles

The QuAAC project has several principles that drive its design and patterns:

- It should store all data relevant to the QA workflow
- It should store both raw and interpreted data
- It is not a replacement for DICOM
- It is not a QA platform
- It does not prescribe the QA workflow
- It is not for patient data
- Data is stored in a way that is readable to humans when possible
- It should use existing standards whenever possible

2.4 Comparison to DICOM

This project can be compared to DICOM as there are many similarities. The rationale of the project is similar. There is no known standard for communicating QA data between vendors nor from in-house formats to vendor formats. An obvious question is: Why not use private tags of the existing DICOM standard? Although private tags are an option, the data and format do not cleanly align with individual tags. Further, the tags and data would have to be specified just as they are for this project. A new standard also allows for using formats and options that are commonly-accepted today. Given the desire to have the data be human readable and avoid proprietary formats, simple text formats are preferred. It's also unclear how this standard will evolve. By starting with a DICOM definition only, the standard may be limited. Finally, the DICOM standard very often centralizes the patient as the foundational starting point [6]. Routine QA data applies across patients and does not share this starting point.

DICOM has 16 different parts, most of which do not apply to the requirements of the interoperability of routine QA data. The scope of QuAAC is similar to sections 3.5 [6] and 3.6 [7] of the DICOM standard.

This project proposes specific implementations, but moreso proposes a paradigm for data serialization. With this definition, it is possible to have a DICOM-specific variant of the QuAAC standard, where private tags are defined that match the overall standard.

2.5 Comparison to an API

Usage of an application programming interface (API) would in theory allow clinics to communicate with other applications. This is commonly used for server-to-server communication. This standard is orthogonal to APIs because even if both parties expose an API, there is no standard for what that API should return and how. An API could be implemented that conforms to this standard, but until both parties agree on what could and should be exchanged, the original problem remains.

3 Format

The QuAAC format proposal is to allow users to write QA data to one or more files in an easy and readable way. We propose the use of the YAML markup language as it is commonly used in many contexts, although the standard is not limited to this file type. Datetimes use the ISO 8601 definition, and data integrity uses the cyclic redundancy check (CRC) algorithm to verify the integrity of all components within the format.

As the format is explained, we will use a scan of a phantom on a CT scanner as an example.

It contains 4 top-level keys which are shown in listing 1: `version`, `data points`, `equipment`, and `files`. The `version` corresponds to the version of the QuAAC standard used, `data points` store individual datum, `equipment` references the equipment the data was measuring on, or ancillary equipment that assisted in the measurement, and `files` reference raw data that the data points were based upon, if any. The `version`, `data points`, and `equipment` keys are mandatory while the `files` section is not required if there was no separate raw data information.

Listing 1: Root format specification

```
version: 1.0
data points:
  - ...  # array of data points
equipment:
  - ...  # array of equipment
files:
  - ...  # array of files
```

The definition of the data points section is the main and longest component and is shown in Listing 2. The `hash` key is a CRC32 hash of the content of all keys in the same data point. It is worth noting that these hashes are not meant to be cryptographically secure and only function as data integrity and uniqueness checks. The remaining keys all contain information relevant to the data point. It is important to note not all keys are required. E.g. the `reviewer name` would not be required if the `review status` is unreviewed.

The `primary equipment` references the machine or unit the QA is being performed on and the `ancillary equipment` is an array of equipment that is used to assist in capturing the data point, such as an ion chamber or water tank. Both of these keys reference entries in the `equipment` top-level key.

The `files` section is an array of references to entries in the `files` top-level key.

Listing 2: Data points section

```
data points:
  - hash: 374284055
    name: CatPhan Uniformity (Top)
    performer name: Kerns, James
    performer id: jkerns
    perform datetime: 2024-01-22T13:36:47.2620
    measurement value: 43.113
    measurement type: number
    measurement unit: HU
    reference value: 40
    description: Top HU ROI of CTP486 module
    procedure: Use CatPhan 604
    performer comments: normal scan
    primary equipment: 2349277898
    ancillary equipment:
      - 90452529
    review status: reviewed
    review datetime: 2024-01-22T13:36:47.2620
    reviewer name: Kerns, James
    reviewer id: jkerns
    reviewer comments: Looks good
    files:
      - 3665905493
      - 1919210666
    parameters:
      energy: 120kVp
      mAs: 30
      protocol name: Pelvis Spotlight
      phantom: CatPhan 604
  - ...  # further entries
```

The equipment section contains the detailed information for the `primary equipment` and `ancillary equipment` keys and the definition is shown in Listing 3:

Listing 3: Equipment section

```
equipment:
  - name: Somatom 2
    type: CT scanner
    model: Somatom Force
    manufacturer: Siemens
    serial number: 12345
    hash: 2349277898
  - name: CatPhan 604 (TB1)
    type: Phantom
    model: 604
    manufacturer: PTW
    serial number: A3405
    hash: 90452529
  - ... # further entries
```

The top-level `files` section contains the detailed information referenced from the data points and is shown in Listing 4:

Listing 4: Files section

```
files:
  - hash: 3665905493
    name: Screenshot of parameters
    type: image/png
    encoding: base64
    compression: gzip
    content: gmewip3bypNBvQyQCymjQ...
  - hash: 1919210666
    name: CT DICOM ZIP archive
    type: application/dicom
    encoding: base64
    compression: gzip
    content: 4sIAKxAsGUC/+y7dVBc0d...
  - ... # further entries
```

4 Discussion

4.1 Advantages

The advantages of this standard include using human-readable storage medium (text), usage of common standards where applicable (ISO 8601 [8], YAML, etc), and the ability to communicate between vendors or programs. The scope of the standard is also limited to routine QA data and thus does not aim to replace any existing standard (e.g. DICOM) and is compatible with existing forms of storage and communication. QuAAC is not trying to re-invent the wheel.

Another advantage comes from references to equipment and files. It is not uncommon to have multiple QA data points be interpreted from a single raw data source. In our ongoing example many data points can be interpreted from a single CT scan of the CatPhan including geometry, HU linearity, HU uniformity, and more. These are individual data points that all reference one data set. By using reference keys, the data storage format does not duplicate this DICOM information yet retains all the individualized information. This logic is similar for equipment, where multiple data points may be taken on one piece of equipment.

Creating a format that is easy to store and read is critical to both adoption and transparency. With the viewpoint that the QA data is ultimately the clinician's data and not that of the vendor, such a standard encourages vendors and other clinics to embrace the ease of transferring data to and from their applications, much the same way DICOM does.

4.2 Limitations

In the current state, the standard has not been tested across multiple vendors. A consortium of clinics and/or vendors would likely be needed to gain widespread adoption. There is also often a lack of initiative on the part of some vendors who might perceive the standard as encroaching on their proprietary data format, or reducing the level of vendor lock-in. Dispelling such perceptions requires the viewpoint that the QA data belongs to the clinician.

There is also no software library available yet to load from and save to this standard. However, it is currently in progress and will be released as an open-source project alongside this standard. The standard will also be incorporated into the Pylinac library so users of that software can export their data into the QuAAC format easily [9].

The QuAAC standard is meant to encompass all routine QA data generated by clinicians working with radiation-generating equipment. However, there may be missing use cases as radiotherapy is the first domain addressed currently. More research will be necessary to ensure these are accounted for, such as in the diagnostic and nuclear medicine domains.

Finally, conformance with such a standard adds a burden to all involved. This is not unlike DICOM, but given the overwhelming advantages and appreciation by all for it, the burden is well worth the effort to give more flexibility and power to the clinician to manage and maintain their own data, wherever they choose to put it.

References

[1] E. E. Klein, J. Hanley, J. Bayouth, et al. *Task group 142 report: Quality assurance of medical acceleratorsa*. 2009. DOI: 10.1118/1.3190392.

[2] J. Hanley, S. Dresser, W. Simon, et al. *AAPM Task Group 198 Report: An implementation guide for TG 142 quality assurance of medical accelerators*. Oct. 2021. DOI: 10.1002/mp.14992.

[3] NEMA. *PS3.1 - Introduction and Overview*. 2014.

[4] W. D. Bidgood, S. C. Horii, F. W. Prior, et al. *Understanding and Using DICOM, the Data Interchange Standard for Biomedical Imaging*. 1997. DOI: 10.1136/jamia.1997.0040199.

[5] M. Mustra, K. Delac, and M. Grgic. "Overview of the DICOM standard". Vol. 1. 2008.

[6] NEMA. "PS3.5 - Data Structures and Encoding". *The DICOM Standard* (2014c 2014).

[7] NEMA. "PS3.6 - Data Dictionary". *The DICOM Standard* (2014c 2014).

[8] I. O. for Standardization and I. O. for Standardization. *ISO - ISO 8601 Date and time format*. 2019.

[9] J. R. Kerns. "Pylinac: Image analysis for routine quality assurance in radiotherapy". *Journal of Open Source Software* 8 (92 Dec. 2023), p. 6001. DOI: 10.21105/joss.06001.

MCMimic: a machine learning approach for fast prediction of Monte Carlo proton dose and LET

Samuel Ingram[1,2], Karen Kirkby[2,1], Ranald Mackay[1,2], and Matthew Clarke[1,2]

[1]Christie Medical Physics and Engineering, The Christie NHS Foundation Trust, Manchester, United Kingdom
[2]Division of Cancer Sciences, University of Manchester, Manchester, United Kingdom

Abstract Monte Carlo is an important technique to achieve accurate dose and LET distributions but comes with the cost of long simulation times. In this study, machine learning is utilised to build a generator able to predict both dose and LET distributions simultaneously for proton beam therapy. An adapted pix2pix generative adversarial network is trained using Monte Carlo distributions in a fully supervised learning approach. This results in the ability to predict both dose and LET at increased speeds that are similar to analytical approaches. At 1%/1mm γ, the model can predict dose at 98.36% and LET at 67.29% pass rates. When using combinations of dose and LET, typically done for radiobiological review, γ pass rates are over 90%. Integration of the model into clinical networks will allow for it's use directly from the treatment planning system. Machine learning is well suited to learning to predict Monte Carlo results.

1 Introduction

A patient's radiotherapy treatment is carefully designed to deliver optimal dose distributions that maximise tumour control whilst minimising the probability of normal tissue toxicity. This is typically achieved using commercial treatment planning systems (TPS) that often utilise proprietary analytical dose calculation systems that are very fast. This is a necessary characteristic as treatment planning is a highly iterative process which can take many iterations to achieve a clinically viable plan. However, it is widely accepted that Monte Carlo, a simulation approach, is the gold standard for accuracy. Monte Carlo can be problematic for treatment planning as it is prohibitively computationally intensive, with traditional, central processing units (CPUs) bound options taking many hours to days to calculate a single iteration and is, therefore, either entirely omitted or reserved purely for final dose checks after the planning process has been completed. There are a growing number of accelerated Monte Carlo options which utilise graphics processing units (GPUs) to speed up the simulation process, but these can often require prohibitively expensive hardware to achieve the necessary speeds required for treatment planning optimisation and patient throughput.

The issue with not being able to utilise Monte Carlo during the planning process is that the inaccuracies of the analytical calculation approaches must be acknowledged by clinicians when reviewing a plan. In proton beam therapy two groups of patients are of specific concern where quantitatively accounting for these inaccuracies would be especially impactful. Firstly, patients with high-density surgical implants experience increased proximal dose to the implant. However, commercial TPSs do not correctly model sharp changes in density, thus under-representing this phenomenon. Monte Carlo would allow for a more accurate depiction of these effects and provide confidence to the reviewing clinician of the dose delivered to neighbouring sensitive organs (e.g., spinal cord). Secondly, Monte Carlo can also provide other physical metrics which describe how the radiation is dosing the tissue. One such metric is the linear energy transfer known to alter the radiobiological effectiveness of the administered dose. Both dose inaccuracies and lack of linear energy transfer calculation result in clinicians having to make a clinical judgement without tangible quantitative values. This can result in requiring treatment plans to be re-made or if there are safety concerns it may result in the decision to decrease the dose to avoid potential normal tissue toxicity. If Monte Carlo dose distribution and linear energy transfer can be modelled at the time of planning these decisions will be better informed.

Creating Monte Carlo distributions using machine learning is a growing area of study. There has been success shown using long short-term memory (LSTM) networks [1], U-Net [2], generative adversarial networks (GANs) [3] and transformers [4]. However, these approaches have not been extended to also creating Monte Carlo LET distributions. In this study, an adapted pix2pix GAN [5] has been used to generate both Monte Carlo dose and LET simultaneously from a model called MCMimic. This method differs from the previous approach of using a GAN [3] which used a discovery cross-domain generative adversarial network (DiscoGAN) which relies on unsupervised learning, for the pix2pix framework which supports fully supervised learning. The rationale is that utilising a fully supervised learning approach simplifies the training process which is beneficial given the model needs to predict both dose and LET distributions rather than just dose. This alteration in approach and the simultaneous generation of dose and LET distributions would provide potential value in clinical decision making as it could be generated at speeds required in planning. Furthermore, the model was developed to generate a single proton spot which allows for the fast creation of influence matrices at Monte Carlo dose accuracy. This is useful for evaluating alternative optimisation approaches and can't be readily obtained from some of the commercial TPSs.

2 Materials and Methods

2.1 Training Data

The training dataset consisted of 6,400 single spots simulated using the Monte Carlo algorithm MCsquared - Fast Monte Carlo dose calculation algorithm [6]. These single spots were generated within one of ten patient CT scans at random locations and energies to obtain a patient-representative spread of density heterogeneities for the simulation. All scans were interpolated to 1 mm voxel size and all data is kept at this resolution for training. The patient cohort was chosen to cover all areas of the body to ensure the model remains robust to the various density changes seen across different sites. The number of primaries for the MCsquared simulations was adjusted to achieve a maximum of 0.5% statistical uncertainty. For LET the MCSquared 'deposited energy' scorer was used which scores a dose-weighted average of the energy loss divided by the length travelled of particles through each voxel and accounts for LET_D from secondary particles [7]. A region of interest (ROI) was created around the single-spot of size 256x64x64 to minimise the data size required for training and inference to only what is pertinent for both dose and LET. This resulted in a targeted 2x256x64x64 output array for the model, where two indexes for the 1st dimension are dose and LET respectively.

2.2 Network

To train a generative model a modified pix2pix GAN was used. Like the targeted output array, the input is also a 2x256x64x64 array. The first index for the input array is a 256x64x64 analytical Stopping Power (SP) calculation which is based on an approximation of the Bethe equation [8]. The second index for the input array is a 256x64x64 relative stopping power (RSP) representation of the CT scan. The GAN comprises a generator and a discriminator model trained in parallel. To moderate observed instability between the generator and discriminator models, the discriminator model had a minimum threshold applied to the loss for the optimisation step to be applied. This biased the training to allow the generator model to create "challenging" outputs before the discriminator would be able to improve and helped stabilise the training.

2.3 γ-Analysis

To evaluate the predicted dose against the MCSqaured Monte Carlo dose a local 3D gamma analysis was performed using the PyMedPhys (version 0.39.3) Python module at 1%/1mm with a 10% lower threshold cutoff. Gamma analysis was carried out for dose, LET, Dose*LET and the McMahon Relative Biological Effectiveness (RBE) calculation [9]. Where the former is just the element-wise multiplication of the dose and LET distributions and the latter is given by $Dose \times (1 + 0.055 \times LET)$.

2.4 Comparison Calculation Systems

AUTOMC is an in-house developed CPU Monte Carlo solution [10] built upon GATE v8.1/Geant4 v10.3.3 and simulates plans on a 2x2x2mm dose grid to an uncertainty of 0.8% within the 90% isodose contour. AcurosPT is a commercial Monte Carlo produced by Varian and is running on CPU for this study. AcurosPT was run using a default 2% uncertainty and 0.01 density threshold. Proton Convolution Superposition (PCS) is a commercial analytical dose calculation system produced by Varian and is running on the CPU.

2.5 Compute Specifications

All MCMimic work was conducted on a Ubuntu 22.04 (LTS) workstation with an Intel Core i9 10980XE CPU, 128 GB RAM and 2x NVIDIA A6000 (48 GB VRAM). MCSquared Monte Carlo simulations were done on the CPU. Input data generation, model training and inference were done using the GPU. Model training utilised both GPUs to speed up the process.

3 Results

3.1 Model Accuracy

To examine the accuracy of MCMimic a per-spot γ-analysis was carried out for a range of spots with differing energies and density paths. The MCSquared Monte Carlo was used as the reference dose and LET. The results (Figure 1) show that the model produced accurate representations of dose across all test cases with an average 1%/1mm γ pass rate of 98.36%. The produced LET distributions were less accurate with a lower γ pass rate of 67.29% with a larger variance in the prediction. However, when coupled with dose, a typical approach to giving biological significance to LET, the γ pass rates increase to 91.94% and 98.04% for Dose*LET and McMahon RBE approaches respectively. This is due to radiobiological approximations placing higher weighting on the dose than LET resulting in good predictions from the model.

3.2 Prediction Efficiency

The time required to produce the outputted dose and LET distributions were compared to a range of other solutions (details in Section 2.4). The times recorded from the other solution was to generate dose distributions only, although AUTOMC can generate LET distributions this was not scored during the comparison tests. As MCMimic always gives you both dose and LET distributions there are no timing differences for creating both distributions. Table 1 and Table 2 show that MCMimic predictions are able to produce the distributions at a significantly faster rate than the other Monte Carlo systems included in the comparison and have similar timings to the commercial analytical solution (PCS). The two

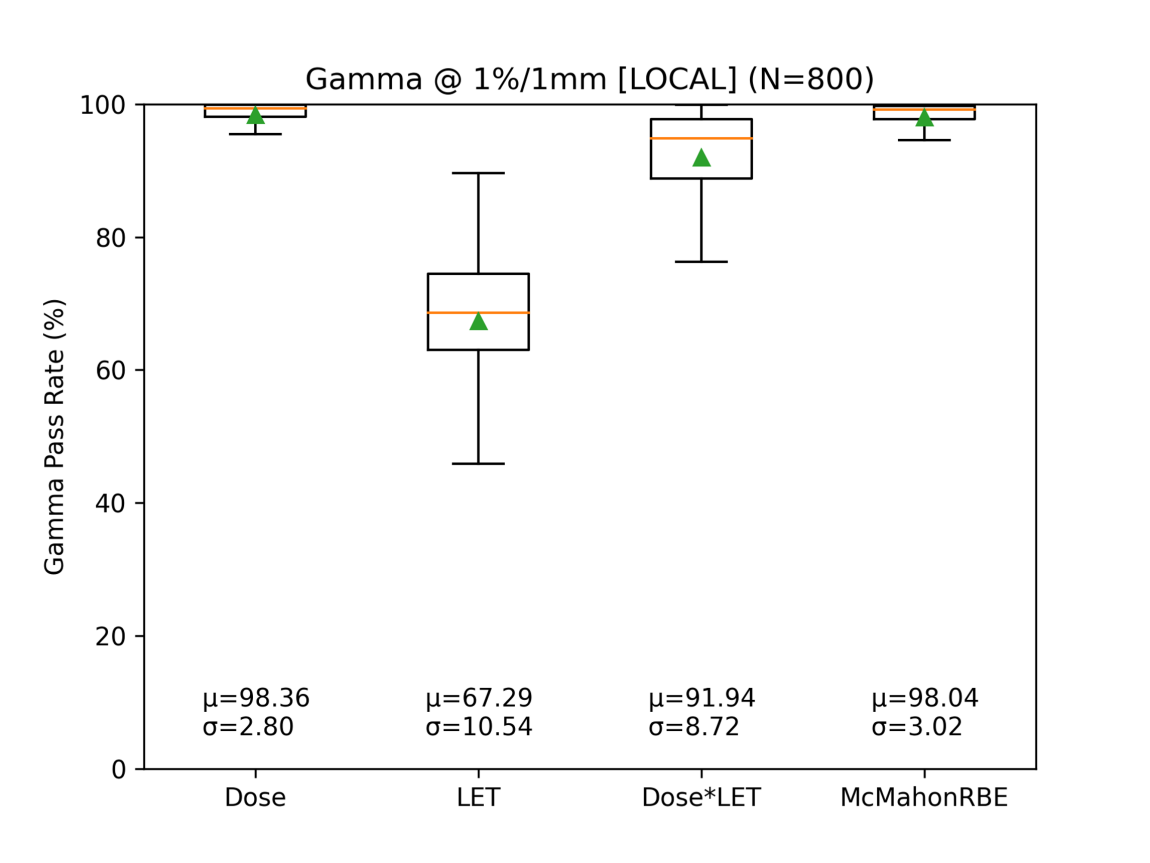


Figure 1: Box plots for the γ pass rates across a range of output distributions from the model based on 800 test spots. Dose and LET are explicitly outputted and Dose*LET and MCMahonRBE are simple combinations of the two distributions. The mean values are shown by the green triangles.

Method	Total Time (s)
AutoMC	5,330.7
AcurosPT (CPU)	126.0
PCS (CPU)	28.0
MCMimic	24.6

Table 1: Nominal 2-field Craniopharyngioma example consisting of 2,104 spots.

Method	Total Time (s)
AutoMC	59,348.7
AcurosPT (CPU)	3,600.0
PCS (CPU)	122.2
MCMimic	245.6

Table 2: Nominal 4-field Cranio-Spinal Irradiation example consisting of 26,809 spots.

nominal examples show that MCMimic plan generation time scales fairly linearly based the number of spots that need predicting, which indicates it would be a good candidate for distributed compute (multiple processors) if predictions were required at faster speeds and would scale well. Furthermore, as each spot is generated independently with its own distributions you can record the results as an influence matrix with no additional compute time required.

3.3 System Integration

For integration into the clinical workflow, careful consideration was taken on ensuring the necessary connectivity and resources were implemented within the clinical network so that MCMimic could be evaluated within a realistic setting. The integration workflow consists of using the Eclipse Scripting API (ESAPI) to trigger a DICOM daemon to push the CT and plan files to a job manager on our Scientific Compute Platform. The job manager will launch and run MCMimic with the sent CT and plan files and record the activity for monitoring purposes. The resultant dose and LET distributions will be sent back to the ESAPI which will inject the results back into Eclipse as new dose maps (LET will be shown as a dose due to not being supported in the Eclipse TPS). From the user perspective, no operations outside of Eclipse will be required by them and the resultant distributions can be viewed with the tools and views that are familiar to the end user.

4 Discussion

Modern proton beam therapy is now typically delivered using a beam scanning system where a set of magnets move the beamlet to the desired x and y position defined withing the proton treatment plan. This x and y position is true at the isocentre but includes a slight angulation to achieve this from as the beam is central before the scanning magnets. This results in beams not being perfectly perpendicular and requiring a correction method to address this. Whilst the angulation is small, given the large number of proton spots in a typical plan (1,000 - 30,000), the effect can result in an observable dosimetric difference, most notably at the patient surface. A possible correction for this, which is being evaluated, is a modification to the SP calculation using a 2D Gaussian that could be offset from the centre and, when convolved along the depth axis of the beamlet, can produce a representation of the beam steering which encodes the necessary information for the model. Whilst the alteration is small, the additional steering information being encoded results in the machine learning problem becoming substantially harder for the model to learn and requires significantly more training data. Further to this, a range of mathematical approximations for ray tracing are being evaluated along with augmenting the RSP maps to allow for perpendicular doses to be mapped back to their angled counterparts.

An alternative approach to machine learning that is growing in availability is to run the Monte Carlo simulations directly on GPUs rather than CPUs. This significantly speeds up the simulation process and can achieve similar speeds to the machine learning approaches. However, these solutions can require more GPU compute than predicting on a pre-trained machine learning model. A benefit of machine learning is that after training the model running it, also called inference, is much quicker and doesn't require the same resources. Whilst the current work has been done showing inference on GPUs it can also be done, at a slower rate, on CPUs. Furthermore, models can the quantised which can reduce the required memory to run the model and expand the places it can be run from (e.g. edge computing). Finally, with Monte Carlo, there is a trade-off between simulation quality and speed, whereas in machine learning the predictions are based on the quality of the training data, which can be simulated to very high quality given the training is only required once per version of the model.

A counterpoint to using machine learning is that the models can only be as good as the data it has been trained upon. If the model encounters anatomy that has not been seen before it could respond incorrectly. Whereas Monte Carlo remains the gold standard as it looks to simulate the underlying Physics. This means care should be taken in the way we utilise machine learning models, especially in the context of informing clinical decision-making. There is also less familiarity with machine learning within the medical physics community and the complexity of the required models used in 3D generation can make it challenging to understand and reinforce trust.

5 Conclusion

Machine learning is a viable method for predicting Monte Carlo distribution. The pix2pix framework is a good option for the presented learning problem. Understanding the limitations of the MCMimic model will be an important step to implementing it safely within a clinical context. In the meantime, it remains a valuable research tool that can quickly calculate dose and LET distributions along with influence matrices that can be used in optimisation studies.

References

[1] A. Neishabouri, N. Wahl, A. Mairani, et al. "Long short-term memory networks for proton dose calculation in highly heterogeneous tissues". *Medical Physics* 48.4 (2021), pp. 1893–1908. DOI: 10.1002/mp.14658.

[2] C. Wu, D. Nguyen, Y. Xing, et al. "Improving proton dose calculation accuracy by using deep learning". *Machine Learning: Science and Technology* 2.1 (2021), p. 015017. DOI: 10.1088/2632-2153/abb6d5.

[3] X. Zhang, Z. Hu, G. Zhang, et al. "Dose calculation in proton therapy using a discovery cross-domain generative adversarial network (DiscoGAN)". *Medical Physics* 5 (2021). DOI: 10.1002/mp.14781.

[4] O. Pastor-Serrano and Z. Perkó. "Millisecond speed deep learning based proton dose calculation with Monte Carlo accuracy". *Physics in Medicine & Biology* 67.10 (2022), p. 105006. DOI: 10.1088/1361-6560/ac692e.

[5] P. Isola, J.-Y. Zhu, T. Zhou, et al. "Image-to-Image Translation with Conditional Adversarial Networks". *arXiv* (2016).

[6] K. Souris, J. A. Lee, and E. Sterpin. "Fast multipurpose Monte Carlo simulation for proton therapy using multi- and many-core CPU architectures". *Medical Physics* 43.4 (2016), pp. 1700–1712. DOI: 10.1118/1.4943377.

[7] D. Wagenaar, L. T. Tran, A. Meijers, et al. "Validation of linear energy transfer computed in a Monte Carlo dose engine of a commercial treatment planning system". *Physics in Medicine & Biology* 65.2 (2020), p. 025006. DOI: 10.1088/1361-6560/ab5e97.

[8] D. R. Grimes, D. R. Warren, and M. Partridge. "An approximate analytical solution of the Bethe equation for charged particles in the radiotherapeutic energy range". *Scientific Reports* 7.1 (2017), p. 9781. DOI: 10.1038/s41598-017-10554-0.

[9] S. J. McMahon, H. Paganetti, and K. M. Prise. "LET-weighted doses effectively reduce biological variability in proton radiotherapy planning". *Physics in Medicine & Biology* 63.22 (2018), p. 225009. DOI: 10.1088/1361-6560/aae8a5.

[10] A. H. Aitkenhead, P. Sitch, J. C. Richardson, et al. "Automated Monte-Carlo re-calculation of proton therapy plans using Geant4/Gate: implementation and comparison to plan-specific quality assurance measurements". *The British Journal of Radiology* 93.1114 (2020), p. 20200228. DOI: 10.1259/bjr.20200228.

Evaluation of the GATE-RTion/Geant4 simulation speed and dosimetric commissioning for independent dose calculation in carbon ion pencil beam scanning radiotherapy

Yihan Jia[1,2], Lisa Hartl[1,3], Martina Favaretto[1], Markus Stock[1,4], Loïc Grevillot[1], and Andreas F. Resch[1]

[1]MedAustron Ion Therapy Center, Wiener Neustadt, Austria
[2]Department of Radiation Oncology, Medical University of Vienna, Vienna, Austria
[3]Faculty of Electrical Engineering and Information Technology, Vienna University of Technology, Vienna, Austria
[4]Department of Oncology, Karl Landsteiner University of Health Sciences, Wiener Neustadt, Austria

Abstract Independent dose calculation (IDC) is an effective method to verify the planned dose in the treatment planning system before the treatment field is delivered to the patient. The open-source IDC IDEAL v1.1, using GATE-RTion v1.0/Geant4.10.3 as dose engine, is designed for the verification of dose calculation in treatment planning system for patients treated with proton and carbon ion pencil beam scanning. This study focuses on carbon ion only and investigates the GATE-RTion/Geant4 parameter settings, e.g. Geant4 reference physics lists, electron production cut, in search of the optimal simulation setting balancing speed and accuracy. Over the clinical energy range of 120-400 MeV/u, key pencil beam parameters were simulated, and compared with reference dosimetric commissioning data set of two clinical carbon ion beamlines at MedAustron Ion Therapy Center. Twenty spread out Bragg peaks of cubic targets of field size of 3-17 cm, modulation length of 3-10 cm and center at depths of 4-24 cm in water were simulated as verification. With the setting of QGSP_BIC_EMZ and high electron production cut, the simulation speed was improved by a factor of 10 compared with Shielding_EMZ, while achieving good dose calculation accuracy: range in water difference better than 0.2 mm at isocenter and better than 0.3 mm at 50 cm reduced air gap with range shifter; spot size in air difference better than 5%; point dose difference in spread out Bragg peak better than 7%. Considering the simulation speed and accuracy of the investigated GATE-RTion/Geant4 setting, IDEAL was commissioned from a dosimetric standpoint for carbon ion IDC.

1 Introduction

The state-of-the-art pencil beam scanning radiotherapy delivers intricate treatment fields meticulously devised by the treatment planning system (TPS) through the beam delivery system. The optimized field covers the planned target volume (PTV) of tumor longitudinally and laterally with pencil beams of various energies and lateral shifts. In MedAustron Ion Therapy Center, the routine quality assurance (QA) guarantees that key beam characteristics fall within tolerance level, e.g. the range and Bragg peak width, the spot size, and the beam intensity of the pristine monoenergetic pencil beam[1, 2]. The uncertainty of beam range and lateral position contributes to the 3D positioning accuracy of every spot in patient. Spot position uncertainty, along with the spot size uncertainty, the intensity uncertainty and the patient positioning uncertainty, affects the accuracy of the delivered dose in patient. Experimental patient specific QA (PSQA) verifies the dose of the planned field at various positions in water. Compared with the experimental PSQA, *in silico* independent dose calculation (IDC) within the actual patient geometry is expected to improve the specificity of the verification process, not to mention the reduction of beam time.

The first part of this manuscript would give an overview of the GATE-RTion/Geant4 settings under investigation. Next, the simulation speed of the settings was assessed. The dosimetric accuracy of the optimal setting was evaluated against measurements over the clinical energy range, with regard to range in water, spot size in air and absolute dose at reference depth in water for monoenergetic pencil beams with and without range shifter. Finally the accuracy of dose simulated at multiple positions in water was verified using 20 spread out Bragg peaks (SOBPs) of uniform physical or RBE_{LEMI} weighted dose to cubic targets of field size of 3-17 cm, modulation length of 3-10 cm and center at depths of 4-24 cm.

2 Materials and Methods

2.1 MedAustron Ion Therapy Center

Synchrotron-based Beam Delivery System at MedAustron Ion Therapy Center can provide clinical carbon ion beams between 120 and 400 MeV/u (range in water from 3 to 27 cm).

2.2 GATE-RTion v1.0

GATE-RTion v1.0, a validated and long-term stable branch of GATE v8.1 intended for clinical purposes[3], is the dose engine of IDEAL v1.0 [4] and v1.1.

2.3 Beam model and configurations of IDEAL v1.1

The beam models of the fixed horizontal beam line (HBL) and the fixed vertical beam line (VBL) of carbon ion were constructed using the method described in [5, 6]. The source-axis-distances (SADs) were defined according to baseline commissioning measurements for HBL and VBL, respectively, corresponding to the distance from the central position

of scanning magnets to the isocenter or the equivalent virtual distance[1, 6].

Several combinations of Geant4 reference physics lists and electron production cuts were tested in order to find an optimal configuration for IDEAL. The test simulation setup was defined as a cubic target of 6 cm side centered at 6 cm depth in water. IDEAL recorded the simulation speed, defined as the number of simulated primaries per second. The simulation speed of three configurations were compared: (A) with or without secondary electrons, (B) using physics builder of Shielding or QGSP_BIC, (C) using physics builder of QGSP_BIC_HP or QGSP_INCLXX_HP. Simulations with secondary electrons used electron production cut of 0.5 mm; simulations without secondary electrons set electron production cut of 100 m.

2.4 1D/2D/3D dosimetric verification of IDEAL v1.1

Integrated Radial Profiles as function of Depth (IRPDs) were measured in water for 16 energies between 120 and 400 MeV/u using the PeakFinder (PTW, Freiburg), positioned at isocenter (i.e. isocenter to detector or phantom surface equals 0 (ISD0)). Additional IRPDs were measured for 4 energies with and without Range Shifter (RaShi) at the Non-Isocentric Reference Point (NIRP)[1] with 50 cm (HBL) or 40 cm (VBL) reduced air gap. Pencil beams of the aforementioned energies were simulated in water. The scoring diameter of the measuring chamber was considered for scoring the IRPDs. Except otherwise specified, simulations were run with 1 mm dose grid and 5E+4 primaries. The absolute difference of the range at 80% of Bragg peak dose (R_{80}) of the measured and simulated IRPD was calculated.

Lynx (IBA-dosimetry, Schwarzenbruck) 2D scintillation detector was used to acquire the spot maps of major energies at several positions: isocenter (ISD0) and with 20, 30, 40 and 50 cm reduced airgaps (ISD20, ISD30, ISD40, ISD50). Aforementioned measurement was repeated with RaShi. Monoenergetic layer of 25 (HBL) or 9 (VBL) spots were simulated in air. At each measured position, the spot full width half maximum (FWHM) was extracted from the Gaussian fit of the dose distribution. The relative difference of the spot size at all ISDs were calculated for the measured and simulated spot maps.

Absolute dose was measured with plane-parallel ionisation chamber (ROOS chamber type TM34001, PTW) placed at 14 mm depth in water. The relative difference of the measured and simulated absolute dose at reference depth in water was calculated.

In 3D dosimetry, the 3D-block (PTW, Freiburg) equipped with 24 PinPoint ionisation chamber (IC) type 31015 (PTW, Freiburg) measured the point-like absolute dose at multiple positions in water. Twenty SOBP plans were optimized in the TPS to deliver a uniform physical or RBE_{LEMI} weighted dose [7] to cubic targets with field size of 3-17 cm, modulation length of 3-10 cm, centered at depths of 4-24 cm in water. Four reference plans highlighted in Table 1 were used to derive an energy-dependent MeterSet Weight Scaling Factor (w_{msw}). The w_{msw} was subsequently applied to the carbon ion beam models and verified with all the plans listed in Table 1. Dose grid of 2 mm and statistical uncertainty goal of 1% were used in simulation and the absolute dose at the positions of the 2352 PinPoint ICs was interpolated linearly. The global dose difference were calculated for 2143 valid PinPoint ICs whose dose gradient was lower than 0.04 Gy/mm and planned dose was higher than 0.05 Gy according to the experimental PSQA procedure. The statistics (signed and unsigned mean) of the global differences per plan were computed by Equation 1 and 2, respectively.

$$\overline{\Delta} = \overline{\frac{(D_i - D_i^{\text{meas}})}{max\,(D_i^{\text{meas}})}}, i \in \text{valid} \tag{1}$$

$$\overline{|\Delta|} = \overline{\frac{|D_i - D_i^{\text{meas}}|}{max\,(D_i^{\text{meas}})}}, i \in \text{valid} \tag{2}$$

Table 1: The 20 SOBPs with field size of 3-17 cm, modulation length of 3-10 cm and center at depths of 4-24 cm are presented. The four highlighted reference plans were used to derive the MeterSetWeight Scaling Factor w_{msw}.

Field size [cm]	Modulation length [cm]	Center depth [cm]	Measurement position	RaShi	Prescribed dose
3	3	4	NIRP	Yes	RBE_{LEMI}
4	4	5	ISD0	No	RBE_{LEMI}
4	4	5	NIRP	No	RBE_{LEMI}
4	4	19	NIRP	No	RBE_{LEMI}
4	4	25	ISD0	No	RBE_{LEMI}
6	6	6	ISD0	No	Physical
		6	ISD0	No	RBE_{LEMI}
		10	ISD0	No	RBE_{LEMI}
		5	NIRP	Yes	RBE_{LEMI}
		6	NIRP	No	RBE_{LEMI}
6	10	7	NIRP	Yes	RBE_{LEMI}
8	8	13	ISD0	No	Physical
			ISD0	No	RBE_{LEMI}
			NIRP	No	RBE_{LEMI}
8	8	15	NIRP	Yes	RBE_{LEMI}
10	10	21	ISD0	No	Physical
			ISD0	No	RBE_{LEMI}
			NIRP	No	RBE_{LEMI}
16	5	6	ISD0	No	RBE_{LEMI}
16	5	6	NIRP	No	RBE_{LEMI}

3 Results

3.1 Simulation speed of tested configurations

The five simulation configurations are summarized in Table 2 and the simulation speed of each tested simulation is listed. Omitting secondary electrons by using high electron production cut improved the simulation speed by approximately a factor of 4 (configuration A1 and A2). Using physics builder of QGSP_BIC instead of Shielding further improved the simulation speed by approximately a factor of 2 (configuration A2 and B1). The physics builder QGSP_BIC and

QGSP_INCLXX had comparable simulation speed, however the option HP (for high precision neutrons) reduced the simulation speed (configuration B1, C1 and C2).

3.2 Dosimetric accuracy of monoenergetic pencil beam

The R_{80} of simulated IRPDs agreed with measurement with an absolute difference of 0.2 mm at ISD0. The differnce of R_{80} of pencil beam with RaShi measured at ISD50 fell within the tolerance of 0.3 mm [1]. The absolute dose at reference depth in water agreed with measurement better than 2%.
The relative difference of simulated spot FWHM agreed with measurement better than 5% at all the non-isocentric measurement positions. This met the tolerance requirement of 10% [1].

3.3 Dosimetric accuracy of SOBP

The empirical w_{msw} derived from the dose difference of reference plans in HBL can be expressed as a third order polynomial equation of the nominal energy, with coefficients in descending order (1.82E-9, 1.70E-6, 3.62E-3, 0.983). The application scaled up the meterset weight of low-energy energy layers by around 0.5%, and scaled down the weight of high-energy energy layers by up to 2.5%. The w_{msw} of VBL was derived similarly.
Figure 1 and 2 demonstrated the global dose difference of valid PinPoint ICs in target region for SOBP plans of HBL and VBL, respectively. For the SOBP plans of HBL, signed and unsigned mean global dose difference had maximum values of 1.4% and 1.9%, respectively. For the SOBP plans of VBL, signed and unsigned mean global dose difference had maximum values of -1.4% and 1.5%, respectively. The global dose difference of all positions in target region was better than -5.1% for HBL and -7.0% for VBL, reaching agreement within the uncertainty of the measurement data.

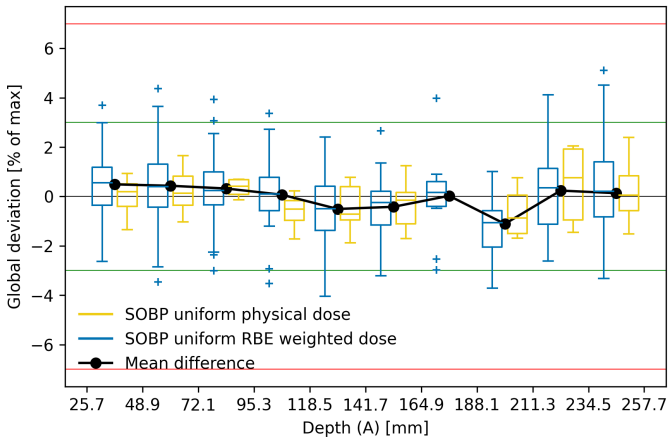


Figure 1: The histogram of the global dose difference of all the valid PinPoint ICs in the target region of all the SOBP plans of HBL.

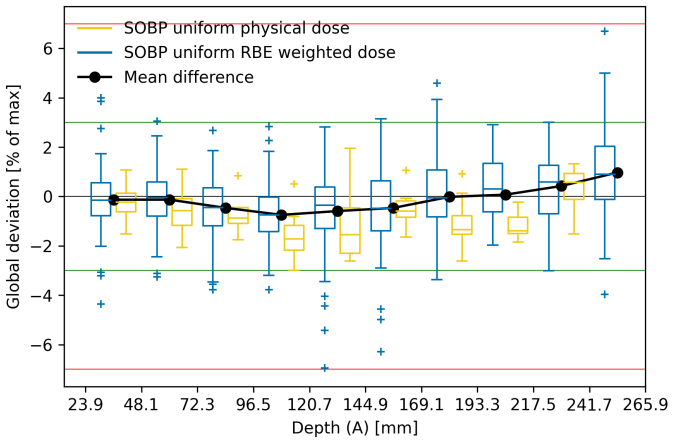


Figure 2: The histogram of the global dose difference of all the valid PinPoint ICs in the target region of all the SOBP plans of VBL.

4 Discussion

It was recommended by [3] to use the physics builder of Shiedling_EMZ, which for nuclear processes uses Quantum Molecular Dynamics (QMD) model, when simulating carbon ion with GATE-RTion v1.0/Geant4.10.3. In this work, the recommended configurations were tested against other nuclear model and electromagnetic (EM) model choices. Shielding physics builder including QMD model was slower than BIC and INCLXX by a factor of 2∼3, same finding was also reported in [8]. BIC and QMD models reported comparable agreement to experimental data with regard to the fragmentation of 400 MeV/u C-12 ion beam in water [9]. Therefore, the QGSP_BIC_EMZ physic builder was chosen in place of Shielding physics builder for the IDC application.
Configurations A2, B1, C1, C2 without secondary electrons reduced simulation time significantly by a factor of 4∼5.
Supplementary simulation of IRPD in water was conducted in order to investigate the dosimetric accuracy with high electron production cut. It was found that R_{80} of simulated pristine pencil beam had excellent agreement with measurement reference when simulated without secondary electrons. The mean range difference against commissioning measurements of 16 energies was 0.0 mm when simulated without secondary electrons, and -0.2 mm when simulated with secondary electrons. It was observed in [10] that the IRPD of the same pristine pencil beam simulated with 0.001 mm production cut and 0.01 mm step limit had shorter R_{80} than IRPD simulated with 1 mm production cut and no step limit, with the R_{80} difference of 2 mm. The finding that simulation with smaller electron production cut had shorter R_{80} was consistent between this work and [10].
The good accuracy of R_{80} with/without RaShi and absolute dose in water verified the modeling of pencil beam energy and RaShi. The agreement of FWHM in air at positions from isocenter to NIRP 50 cm off isocenter validated the optics parameters of the pencil beam and the relative position of

Table 2: The simulation speed of five configurations using different physics builders and secondary electron settings.

Configuration	Physics builder	EM option	Simulation of secondary electrons	Simulation speed [primaries/second]
A1	Shielding	EMZ	Yes	16.1
A2	Shielding	EMZ	No	72.5
B1	QGSP_BIC	EMZ	No	171.2
C1	QGSP_BIC_HP	EMZ	No	122.9
C2	QGSP_INCLXX_HP	EMZ	No	122.0

the source and the scanning magnets to isocenter. With the correction of the w_{msw}, the absolute dose in SOBP showed good agreement with measurement in spite of the positioning uncertainty of the PinPoint ICs, with no dependency on depth or optimization method (uniform physical or RBE_{LEMI} weighted dose). Similarly for proton, scaling factors of the order of 2% were required between the reference doses of monoenergetic pencil beams and the 3D doses of SOBP for TPS and IDC systems used at MedAustron[4].

5 Conclusion

IDEAL v1.1 (GATE-RTion v1.0/Geant4.10.3) settings have been evaluated with regard to simulation speed and accuracy. The physics builder of QGSP_BIC_EMZ with high electron production cut significantly improved the simulation speed, while maintaining good dosimetric accuracy compared with measurements over the clinical energy range of carbon ion radiotherapy. IDEAL in its current optimized configuration was commissioned from a dosimetric standpoint for carbon ion IDC. Further clinical commissioning comparing TPS against IDEAL in patient geometry will follow as a next step.

6 Acknowledgements

The authors acknowledge the MedAustron Medical Physics team for providing the dosimetric commissioning measurements used in this work.
This work was funded by the "Niederösterreichischen Wirtschafts- und Tourismusfonds" and European union in form of the EFRE funds.

References

[1] L. Grevillot, J. Osorio Moreno, V. Letellier, et al. "Clinical implementation and commissioning of the MedAustron Particle Therapy Accelerator for non-isocentric scanned proton beam treatments". *Medical Physics* 47.2 (2020), pp. 380–392.

[2] L. Grevillot, J. O. Moreno, H. Fuchs, et al. "Implementation of Sphinx/Lynx as daily QA equipment for scanned proton and carbon ion beams". *Journal of Applied Clinical Medical Physics* 24.4 (2023), e13896.

[3] L. Grevillot, D. Boersma, H. Fuchs, et al. "GATE-RTion: a GATE/Geant4 release for clinical applications in scanned ion beam therapy". *Medical Physics* 47.8 (2020), pp. 3675–3681.

[4] L. Grevillot, D. Boersma, H. Fuchs, et al. "The GATE-RTion/IDEAL independent dose calculation system for light ion beam therapy". *Frontiers in Physics* 9 (2021), p. 704760.

[5] H. Fuchs, A. Elia, A. F. Resch, et al. "Computer-assisted beam modeling for particle therapy". *Medical Physics* 48.2 (2021), pp. 841–851.

[6] A. Elia, A. F. Resch, A. Carlino, et al. "A GATE/Geant4 beam model for the MedAustron non-isocentric proton treatment plans quality assurance". *Physica Medica* 71 (2020), pp. 115–123.

[7] M. Krämer and M. Scholz. "Treatment planning for heavy-ion radiotherapy: calculation and optimization of biologically effective dose". *Physics in Medicine & Biology* 45.11 (2000), p. 3319.

[8] D. Bolst, G. A. Cirrone, G. Cuttone, et al. "Validation of Geant4 fragmentation for heavy ion therapy". *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 869 (2017), pp. 68–75.

[9] P. Arce, D. Bolst, M.-C. Bordage, et al. "Report on G4-Med, a Geant4 benchmarking system for medical physics applications developed by the Geant4 Medical Simulation Benchmarking Group". *Medical physics* 48.1 (2021), pp. 19–56.

[10] N. Zahra, T. Frisson, L. Grevillot, et al. "Influence of Geant4 parameters on dose distribution and computation time for carbon ion therapy simulation". *Physica Medica* 26.4 (2010), pp. 202–208.

New global reconstruction for charged particles with the FOOT experiment

Ch. Finck on behalf of the FOOT collaboration[1]

[1]Université de Strasbourg, CNRS, IPHC UMR 7871, F-67000 Strasbourg, France

Abstract In the framework of measuring double differential cross-section for hadrontherapy purposes, a new algorithm for global track reconstruction (Tracking Of Ejectile) was implemented in the analysis package for the FOOT experiment data set. The algorithm is based on a unscented Kalman filter approach. The cluster recognition efficiency was found to be around 99% with purity above 94%. The momentum resolution of the reconstructed track for particle with a momentum $\geq$ 2 GeV/c was found to be lower than 5 % what corresponds to the desired resolution for the FOOT experiment in order to separate the different isotopes.

1 Introduction

Hadrontherapy treats deep-seated tumours with beams of charged particles (protons and heavier ions) to treat. These particles have a favourable depth-dose distribution in tissue characterized by a maximum energy deposition (Bragg peak) near the end range where the tumour is located. Moreover, Carbon and Oxygen ions have an enhanced biological effect allowing to successfully treat radio-resistant tumours. However, the nuclear interactions of the beam nuclei with the patient tissues can induce fragmentation of projectiles and/or target nuclei and needs to be carefully taken into account when planning the treatment.

At present, there is a lack of experimental data about nuclear fragmentation for light fragments (Z<10) in the energy range between 100 MeV/A and 800 MeV/A, typical energies used in hadrontherapy and relevant for space radio-protection. They would be fundamental in order to further optimize treatment planning system in hadrontherapy [1] and risk models for space missions [2]. To fill this gap, the FOOT (FragmentatiOn Of Target) experiment has been conceived with the aim of measuring differential nuclear cross sections of both target and projectile fragments with an uncertainty lower than 5%, taking advantage of an inverse kinematic approach for the measurement of target fragments.

The FOOT experiment was designed to detect, track and identify nuclear fragments and aims to measure double differential cross sections both in angle and kinetic energy. To identify fragment tracks two methods are used, one based on the GenFit [3] package and another based on a Kalman filter approach named TOE (Tracking Of Ejectile)

The performance of the TOE algorithm will be studied through different simulated data sets.

2 Materials and Methods

2.1 Experimental setup

The FOOT magnetic spectrometer setup, is optimized for the identification of fragments heavier than or equal to α particles at low angle $< 10^o$, while the emulsion chamber setup is an apparatus covering large angle up to 70° and it is optimized for the detection of fragments within Z $\leq$ 3 [4]. The FOOT electronic setup [5], depicted in Figure 1, consists of an upstream region composed of the pre-target detectors, composed by a start counter (SC) and a beam monitor (BM) to characterize the incoming beam, including the target. After the target a tracking region, including a vertex detector (VTX), the magnet, a inner tracker (ITR) and a micro-strip detector (MSD), is located. The downstream region, as for it, consists of a time of flight wall (TW) and a calorimeter (CAL).

2.1.1 Start counter

The SC is made of a thin squared foil (5 cm side, 250 μm thickness) of EJ-228 plastic scintillator. It provides the trigger signal to the whole experiment and the measurement of incoming ion flux to be used for the cross section measurement. It provides the reference time for all the other detectors and allows the time of flight measurement in combination with the TW.

2.1.2 Beam monitor

The BM is a drift chamber composed of twelve wire layers, with three drift cells per layer. The cell shape is rectangular (1.6$\times$1.0 cm^2) with the long side orthogonal to the beam. The BM was operated at atmospheric pressure, at the working voltage of 1.8 kV while using an Ar/CO2, 80/20% gas mixture. The BM detector is placed between the SC and the target and is used to measure the direction and incident of the ion beam track. The BM efficiency was measured to be $\sim$ 90% and the mean track spatial resolution was measured to be of the order of $\leq$ 100 μm [6].

2.1.3 Vertex detector and Inner tracker

The VTX is a 2$\times$2 cm^2 pixelised silicon detector based on the CMOS technology made of 4 consecutive planes of MIMOSA- 28 (M28) chip [7]. The VTX is placed downstream the target. The M28 chip is composed of 928 rows $\times$

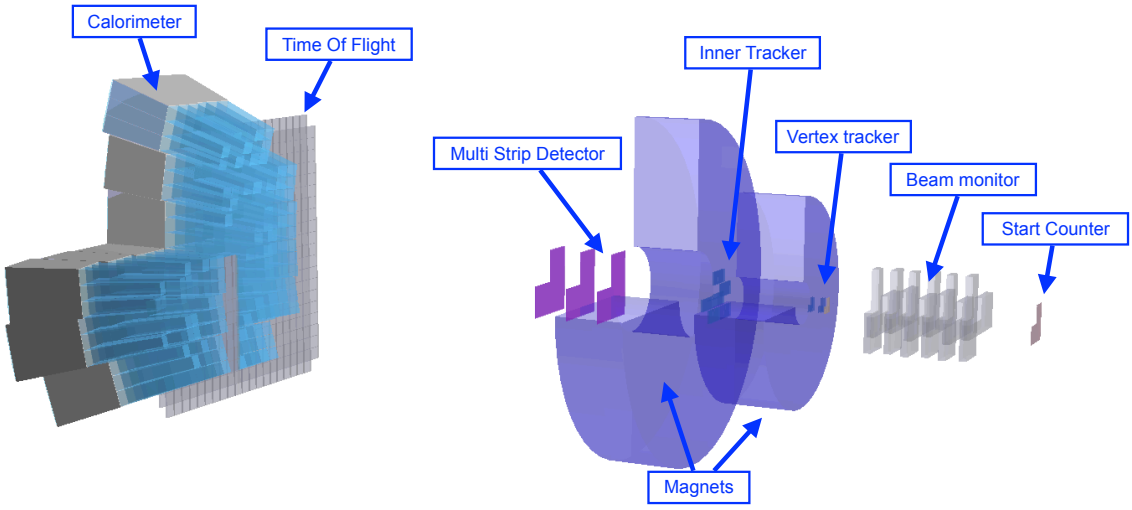


Figure 1: Schematic view of the FOOT experimental setup

960 columns with a pixels size of of 20.7 μm. The sensor has a total thickness of 50 μm with an epitaxial layer of 14 μm.
The ITR is made of two planes of 16 M28 chip each, covering a area of 4×4 cm^2. The ITR is placed in between two asymmetrical permanent magnets in a Halbach configuration with a maximum intensity of 1.4 T and 0.9 T respectively.

2.1.4 Time of flight wall

The Time of flight (TOF) wall detector (TW) is composed of two orthogonal layers of 20 plastic scintillator bars (EJ-200). Each bar is 0.3 cm thick, 2 cm wide and 44 cm long. The two x-y layers form a 40×40 cm^2 active area detector that provides the deposited energy (ΔE) measurements. The time of flight is computing by the average tine of the two planes using as reference the time given by the SC. An energy loss resolution $\sigma(\Delta E) \sim 5\%$ was obtained for different beam/energy combinations. The resolution of the time of flight was measured to be $\sim$ 75 ps [8].

2.1.5 Calorimeter

The CAL is composed of 320 $Bi_4Ge_3O_{12}$ (BGO) crystals, with a truncated pyramid shape: a front side of 2×2 cm^2, a rear side varying from 2.7×2.7 cm^2 to 2.9×2.9 cm^2 and a length of 24 cm, which are positioned with an approximately circular arrangement of about 22 cm radius. The energy resolution is expected to be less than $\sigma \leq 3\%$.

2.2 Local reconstruction

The track of a charged particle passing through the BM is reconstructed from the hits in each cell thanks to an algorithm based on the Legendre polynomials. The hit efficiency was found to be 0.929 $\pm$ 0.008 [6].

The cluster position is computed from the fired pixels of the VTX and ITR or fired strips of the MSD or fired crystal in the calorimeter, by a next neighboring algorithm [9]. The local track reconstruction of the trackers is based on a cluster position combinatorial algorithm. The tracking efficiency is higher than 98 % The vertices are computed using a less tack distance algorithm for the VTX. The efficiency was found to be higher than 98 % [9]. The alignment of the VTX detector is performed using a global χ^2 cost function [10]. The sensors are aligned better than 5 μm in x-y position and better than 0.1 ^{o} in rotation angle around the z-axis.
The space position of the track intercepted by the TW detector, the front and rear TW hits that provide respectively the (x,z) and (y,z) separate information, have to be matched. A clustering algorithm is used to pair the hits detected in the two layers and to find the proper match. The Z identification (ZID) algorithm of the TW was developed to exploits the correlation between ΔE and TOF, for a particle of charge Z, that could be expressed by means of the Bethe-Bloch formula.

2.3 Global reconstruction

The momentum (p) of the fragment can be deduced thanks to the ZID and the curvature of the track in the magnetic field. The used algorithm (TOE) is based on a unscented Kalman filter approach. The mass (m) of the fragment can be computed from the total kinetic energy (E_k) measured with the CAL using the following formula:

$$p = m\beta\gamma\, E_k = mc^2(\gamma - 1) \qquad (1)$$

where $\beta = \frac{l}{TOF}$ and $\gamma = \frac{1}{\sqrt{1-\beta^2}}$, l is the distance of the particle between the SC and the TW.
The role of the global reconstruction is dual: first, the selection of the clusters belonging to a same track, representative of a fragment trajectory inside the experimental setup; second, the extraction of the parameters of interest from the

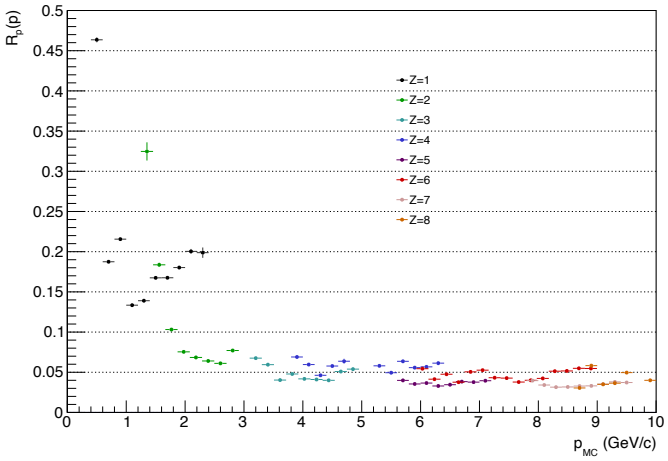


Figure 2: Resolution on momentum as a function of simulated momentum, for the data set labelled A. Each fragment charge distribution was computed individually.

corresponding reconstructed track. Those two distinct roles are respectively called track recognition and track fitting.

2.3.1 Track recognition

The goal of track recognition is to find the reconstructed clusters, formed by local reconstruction of the detectors, that belong to a same given trajectory. The method used to perform this selection must be developed and designed in accordance with the needs and characteristics of the experiment. The track recognition procedure in the TOE algorithm was designed around three intertwined components that form the basis of the algorithm: the propagator, which is dedicated to the track state propagation from one detection layer to the next; the filter, which is used to compare the extrapolation of the track parameters to the available measurements and correct the track state accordingly; and finally the arborescence, or the memory of the procedure, which holds the possible trajectories for the track under reconstruction.

2.3.2 Track fitting

Once a track has been reconstructed, i.e. the clusters belonging to it have been identified and grouped together, the extraction of the parameter of interests for the experiment can be performed; in our case, both the length of the path and the momentum of the particle.

3 Results

The results are based on simulations including the geometry and response function of the FOOT setup. Four different combinations of beam, target and energy were simulated: ^{16}O beam on C_2H_4 target at 200 MeV/u (labelled A), ^{16}O beam on carbon target at 200 MeV/u (labelled B), ^{16}O beam on carbon target at 400 MeV/u (labelled C) and ^{12}C beam on carbon target at 200 MeV/u (labelled D).

	A	B	C	D
Eff.	98.7 ± 0.4	98.7 ± 0.4	99.8 ± 0.5	99.0 ± 0.7
Purity	95.1 ± 0.1	94.5 ± 0.1	99.6 ± 0.1	94.2 ± 0.2

Table 1: Efficiency and purity for four different campaign data sets

3.1 Track recognition

The efficiency is defined as the ratio of the number of reconstructed tracks over the simulated tracks in the acceptance of the FOOT setup. The purity is defined as the number of cluster belong to the same particle over the number of clusters belonging to the same track.
The efficiency and purity results are reported in the table 1 for the given setup.

3.2 Track fitting

The resolution of the momentum $R_p(p)$ is given as the standard deviation of the difference between the simulated momentum (p_{MC}) and the reconstructed one (p_{rec}) normalized by the simulated one (p_{MC}).
Figure 2 shows the momentum resolution versus the simulation momentum for the campaign ^{16}O beam on C_2H_4 target at 200 MeV/u for each fragment in Z charge.

4 Discussion

There is no real distinction in term of efficiency separates the data sets: regardless the configuration, the overall recognition efficiency value is close to 99% for all charged particles. Efficiency for Z=1 and Z=2 (results not shown here) is not far from this value: taking into account the associated error, it fluctuates around 98% for all data sets. However, the association of clusters to the track is wrong in a non-negligible proportion of reconstructions: the purity for charge Z=1 and

Z=2 (not shown here) is respectively around 90% and 94% across data sets.

5 Conclusion

A new global tracking (TOE) algorithm was studied with four different sets of simulated data. Four data sets studied, the overall recognition efficiency of the algorithm was found to be around 99%, for a cluster purity above 94%. In order to achieve its goals, the FOOT experiment need to achieve a momentum resolution of the reconstructed track lower than 5%. For the heavier incident particles ($Z > 2$), this objective is achieved by the TOE algorithm. For lighter fragments, the performances are more nuanced: the resolution on the momentum ranges between 5% to 40%.

References

[1] G. Battistoni, I. Mattei, and S. Muraro. "Nuclear physics and particle therapy". *Advances in Physics* 4 (2016), p. 661.

[2] M. Durante and F. A. Cucinotta. "Physical basis of radiation protection in space travel". *Rev. Mod. Phys.* 83 (2011), p. 1245.

[3] J. Rauch and T. Schlute. "GENFIT—a generic track-fitting toolkit". *Journal of Physics: Conference Series* 608 (2015), p. 012042.

[4] G. Galati, V. Boccia, A. Alexandrov, et al. "Charge identification of fragments produced in ^{16}O beam interactions at 200 MeV/n and 400 MeV/n on C and C_2H_4 targets". *Frontiers in Physics* 11 (2024).

[5] G. Battistoni, M. Toppi, and V. Patera. "Measuring the impact of nuclear interaction in particle therapy and in radio protection in space: the foot experiment". *Frontiers in Physics* 8 (2021), p. 2020.

[6] Y. Dong, S. Gianluigi, A. A. C. Sofia and, et al. "The drift chamber detector of the foot experiment: Performance analysis and external calibration". *Nucl. Instrum. Methods Phys. Res. A* 986 (2021), p. 164756.

[7] C.-A. Reidel, C. Schuy, C. Finck, et al. "Response of the Mimosa-28 pixel sensor to a wide range of ion species and energies". *Nucl. Instrum. Methods Phys. Res. A* 1017 (2021), p. 165807.

[8] A. Kraan, R. Zarrella, A. Alexandrov, et al. "Charge identification of nuclear fragments with the foot time-of-flight system". *Nucl. Instrum. Methods Phys. Res. A* 1001 (2021), p. 165206.

[9] R. Rescigno, C. Finck, D. Juliani, et al. "Performance of the reconstruction algorithms of the first experiment pixel sensors vertex detector". *Nucl. Instrum. Methods Phys. Res. A* 767 (2014), p. 34.

[10] C.-A. Reidel, C. Finck, C. Schuy, et al. "Alignment procedure of silicon pixel detectors for ion-beam therapy applications". *Nucl. Instrum. Methods Phys. Res. A* 931 (2019), p. 142.

Deep-learning powered denoising of Monte Carlo dose distributions

Hannes A Loebner[1], Raphael Joost[1], Jenny Bertholet[1], Stavroula Mougiakakou[2], Michael K Fix[1], and Peter Manser[1]

[1]Division of Medical Radiation Physics and Department of Radiation Oncology, Inselspital, Bern University Hospital, and University of Bern, 3010 Bern, Switzerland
[2]ARTORG Center, University of Bern, Murtenstrasse 50, 3008 Bern, Switzerland

Abstract This work demonstrates the development and application of fast deep-learning models designed for the mitigation of noise in Monte Carlo dose distributions (MC-DDs) with high statistical uncertainty (SU) of radiotherapy treatment plans. Five 3D U-net models were developed, varying in input, input size and batch normalization. The models were trained on pairs of high/low SU MC-DD pairs of randomly generated clinically motivated treatment plans calculated on open source available computed tomography (CT) scans. Depending on the model, the CT was included as input. The accuracy of the models is evaluated on the test set using gamma passing rate (2% global, 2 mm, 10% threshold) comparing denoised and low SU MC-DD. The best performing model included the CT input and no batch normalization. It was retrained for detailed evaluation on inhouse data, consisting of high/low-SU MC-DD pairs generated from 106 clinically-motivated volumetric modulated arc therapy (VMAT) arcs for 29 CT scans. The dataset was augmented to encompass a total of 3074 pairs. On the test set, the denoised MC-DDs agree with low-SU MC-DDs, by an average (standard deviation) gamma passing rate of 82.9% (4.7%). Applied to 12 previously unobserved clinically-motivated cases originating from different treatment sites yields an average gamma passing rate of 91.0%. Denoised DDs are computed, on average, in 35.1 seconds, a 340-fold efficiency gain compared to the calculation of low-SU MC-DDs.

1 Introduction

Monte Carlo (MC) dose calculation is widely acknowledged as the gold-standard for radiation therapy dose calculation, mostly due to its ability to accurately simulate the particle transport in heterogeneous media [1]. However, the necessity for a substantial number of particle histories in MC simulations to achieve reasonable statistical uncertainty (SU) leads to prolonged calculation times [2] and poses challenges for clinical implementation. Current clinical practice often resorts to analytical or simplified MC dose calculation algorithms, balancing accuracy with calculation times typically in the range of several minutes [3].

Deep learning (DL) presents a promising avenue to circumvent long calculation times while maintaining promising accuracy. There exists several approaches, incorporating U-nets [4, 5] and general adversarial networks [6–8] to perform dose predictions, relying only on CT and/or structure set as input, for specific treatment techniques. However, these dose predictions approaches have inherent drawbacks. They may produce results substantially deviating from the ground truth, introducing errors such as predicting doses far away from the tumor [9]. Additionally, their employment is often constrained by training dataset specificity, rendering them dependent on particular cancer types, treatment sites, or techniques [5].

To address these limitations, a promising approach is to give an "informed" input: This involves first the computationally cheap and fast calculation of MC dose distributions (MC-DDs) with high SU using only a limited number of particle histories. This informed input potentially reduces the risk of introducing errors such as mentioned before. In a second step, these DDs are denoised using DL to infer DDs similar to MC-DD with low SU [10, 11]. This approach is expected of being applicable to different treatment sites and techniques, as the primary task is denoising instead of full dose prediction. Moreover, using this approach, the problem of limited training data can be adressed, as random MC-DDs with low/high SU can be generated if needed. In our research group, the Swiss Monte Carlo Plan (SMCP) [12] has been developed and presents a flexible framework to calculate MC-DDs of various treatment techniques and is able to generate the necessary MC-DDs for training such DL models.

The objective of this study is therefore to develop and test different flavors of these denoising models and to apply the best performing model to 12 unseen clinically motivated cases.

2 Materials and Methods

In this work five different DL-powered denoising models are developed and tested (table 1). The main architecture consists of a standard 3D U-net [13] with four layers, integrating skip connections between encoding and decoding layers. Convolutional kernel sizes are 3*3*3 voxels, and max-pooling dimensions are 3*3*1 for the first and 2*2*2 voxels for the subsequent layers. The input dimension is 192*192 voxels in x*y direction and varying size in z direction. Geometries exceeding 192 voxels in x- and y-direction are cropped centrally and geometries exceeding 64 voxels in z-direction are handled using a patched-based approach [14]. Resulting overlapping voxels are averaged using TorchIO [15]. Geometries with less voxels are padded with zero voxels to fit the model input size. The model has 1-2 input channels, depending whether the CT is given as an additional input to the MC-DD with high SU or not.

Model	z-size	CT	Batch normalization
1	32	yes	no
2	64	yes	no
3	96	yes	no
4	64	no	no
5	64	yes	yes

Table 1: 3D U-net model flavours 1-5. The models differ by with varying input, input size and batch normalization.

During training, random patches with size 192*192*z are selected from DDs (and CTs, if specified), with additional data augmentation involving random mirroring and/or rotations in multiples of 90° in the xy-plane. The training employs a summed-squared-error loss function and a batch size of two. The ADAM optimizer [16] is used, starting with an initial learning rate of 10^{-4}. The learning rate decreases by a factor of 5, if the validation loss remains unchanged for four consecutive epochs. Training concludes when there is no improvement in the loss function over the last 30 epochs. The model is trained on a GeForce RTX 3090 GPU.

The dataset for training, validation and testing consisted of 5364 randomly generated volumetric modulated arc treatment (VMAT) plans: random multi-leaf collimator shapes are generated for a full gantry rotation, the collimator angle is randomly assigned and the monitor unit weights for each control point are assigned by randomly sampling numbers between 0 and 1 and sorting them in ascending order. Additionally, a random isocenter (within +- 5 cm in x, y and z of the center of the CT) is assigned. The beam energy is set to 6 MV. Always 12 plans were applied to 447 open source clinically motivated CTs to obtain a total of 5364 high/low SU MC-DD pairs. The high/low SU MC-DDs were calculated using SMCP [11], with $1.5*10^6$ particle histories (high SU >60%) and $3*10^8$ particle histories (low SU <2%), respectively. The dataset is divided into training, validation, and test set with 80%/10%/10% split.

The model accuracy is evaluated by comparing the low-SU MC-DD (reference) with the denoised DD by means of root-mean-squared error (RMSE) over all voxels and Gamma passing rate on the test set. Gamma evaluation with 2%/2 mm [17] and 3%/3 mm [18], 10% threshold is denoted as gamma-2 and gamma-3, respectively.

The best performing model out of those considered was re-trained on 106 clinically motivated VMAT applied to 29 CTs from our own data base. The plans were augmented by randomly changing the beam energy (6, 10, 15 MV), the isocenter (within +-5 cm in x, y, z direction), and the collimator angle, to a total of 3074 high/low SU MC-DD pairs. Training was completed within 28 h 19 min (149 epochs).

The re-trained model is applied to 12 previously unseen clinically motivated cases, including 4 head and neck, 4 non-small-cell lung, and 4 prostate cases. RMSE, gamma-2, and gamma-3 are evaluated. The noise reduction from high SU MC-DD to denoised DD, is quantified using the Improved Signal-to-Noise Ratio (ISNR) [11].

The computation time to calculate a high-SU MC-DD, the subsequent preprocessing, denoising and storing is recorded and compared to the calculation time of the low-SU MC-DD.

3 Results

The accuracy of the five models is shown in table 2. Including the CT information improved the model accuracy. Moreover, batch normalization had a negative impact on the model accuracy.

Model	Average gamma-2 [%]	Average gamma-3 [%]	RMSE [10^{-4} Gy/MU]
1	89.16±3.99	97.30±1.69	1.21±0.22
2	**89.74±4.00**	**97.50±1.66**	**1.16±0.20**
3	89.72±3.92	97.48±1.61	1.15±0.20
4	89.23±4.16	97.30±1.78	1.16±0.20
5	83.54±4.77	94.50±2.55	1.26±0.27

Table 2: Model accuracy including standard variation. The best performing model is marked in bold.

The model 2 had a mean (standard deviation) calculation time of 32.7 s (7.5 s) to obtain a high-SU MC-DD on a single CPU. Subsequent data preprocessing, denoising, and storage were completed on average (standard deviation) within 2.6 s (0.3 s). This contrasts an average (standard deviation) calculation time of 3 hours and 19 minutes (42 minutes) for low-SU MC dose calculation, including storage, on one CPU: an efficiency gain of factor 340. The application of the retrained model to 12 unseen cases is demonstrated for one head and neck and one lung case in figure 1.

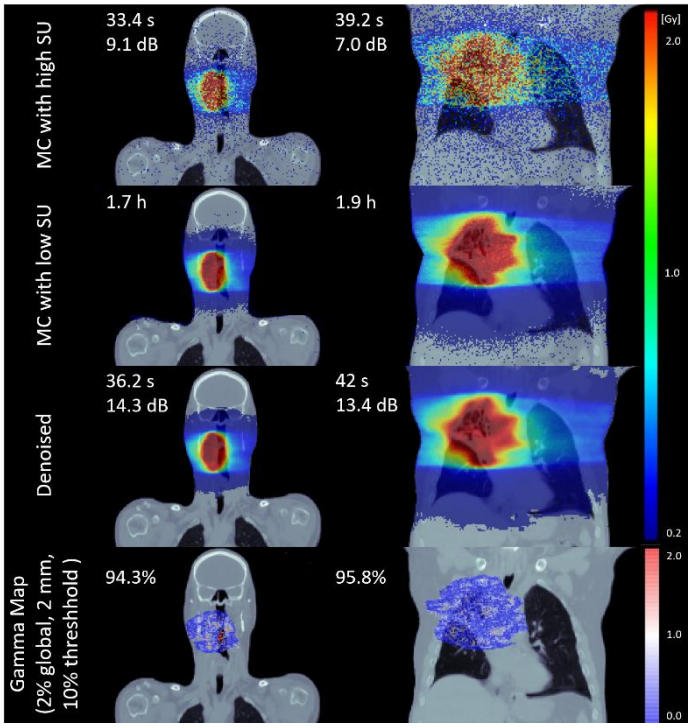


Figure 1: High SU, low SU and denoised DDs are visualized with 10% dose threshold for a head and neck (left) and a lung (right) case. Their computation times and signal to noise ratio as compared to the MC-DD with low SU are shown on the respective upper left corner. The gamma evaluation (low SU DD as reference and denoised DD as evaluation) is shown on the bottom. In the gamma map, voxels failing the gamma criteria (> 1.0) are shown in red tones with the gamma passing rate shown in the top-left corner of the gamma map.

For the 12 cases, the average [range] RMSE is $1.1*10^{-4}$ Gy/MU [$0.5*10^{-4}$, $1.7*10^{-4}$]. The average [range] gamma-3 and gamma-2 is 97.5% [93.4, 98.7] and 91.0% [79.6, 95.9], respectively. The average [range] ISNR is 9.9 dB [4.2, 15.2].

4 Discussion

In this work, we have developed and tested DL-powered models to denoise MC-DDs.

Generally, the differences in accuracy between the models are small: the different z input sizes had little impact on the gamma-2, gamma-3 and RMSE. However, the patched borders could be identified visually due to "steps" in the dose distribution. Keeping in mind, that the GPU has a limited memory [19], the results suggest choosing a z-size conforming to the availability and specifications of the available GPU model. Similarly, no substantial improvement in accuracy due to the additional CT input was observed. However, in the context of explainable results and minimizing the risk of false dose predictions, the CT input is potentially beneficial. Including batch normalization substantially degraded the accuracy of the denoising. Although batch normalization is used in many segmentation tasks and similar U-nets [20], it was not beneficial for the denoising task. This might be due to the fact that batch normalization tends to lose the information of the absolute dose values, which is critical for predicting accurate dose distributions. This effect can be magnified when using patching, as the information of the previous patch is not available for the next one.

Denoised DDs generated with the best performing model considered exhibit an average RMSE of $1.1*10^{-4}$ Gy/MU when compared to low-SU MC-DDs, aligning with findings in literature using similar methodologies [10, 19]. For a representative VMAT arc with 300 MU and a prescription of 50 Gy, delivered in 2 Gy fractions, this leads to an average RMSE of 0.9 Gy for the whole treatment. Although these results are quite promising, they do not conform to the dose calculation accuracy levels recommended by AAPM [21].

For the 12 unseen clinically motivated cases slightly lower gamma passing rates are observed when compared to previous works using DL-powered denoising or dose prediction methods [10, 18]. This can be partly explained by the slightly different task of denoising the DD of entire VMAT arcs rather than beams from single directions, resulting in a 9.9 dB average improvement in SNR, a factor 2 smaller than in other studies [11]. Moreover, the other studies used less/more particle histories for the high/low SU MC-DDs.

The denoising computation time is comparable to the findings in literature [10]. The entire workflow, encompassing the generation and denoising of high-SU MC-DDs, is completed within less than a minute, surpassing the speed of the majority of clinically applied dose calculation algorithms [3, 22].

5 Conclusion

We developed and tested DL-powered denoising model for MC-DD. The most accurate model showed an impressive efficiency gain (factor 340), substantial noise reduction and promising accuracy.

Acknowledgements

This work was supported by Varian, a Siemens Healthineers Company. Dose calculations were performed on UBELIX (www.id.unibe.ch/hpc), the high-performance cluster at the University of Bern.

References

[1] Fauzi A, Zahra F, Miftahul Munir M, Krieger T, Sauer OA. Monte Carlo- versus pencil-beam-/collapsed-cone-dose calculation in a heterogeneous multi-layer phantom. Phys Med Biol 2005;50:859. https://doi.org/10.1088/0031-9155/50/5/010.

[2] Chetty IJ, Curran B, Cygler JE, DeMarco JJ, Ezzell G, Faddegon BA, et al. Report of the AAPM Task Group No. 105: Issues associated with clinical implementation of Monte Carlo-based photon and electron external beam treatment planning. Med Phys 2007;34:4818–53. https://doi.org/10.1118/1.2795842.

[3] Das IJ, Moskvin V, Johnstone PA. Analysis of Treatment Planning Time Among Systems and Planners for Intensity-Modulated Radiation Therapy. Journal of the American College of Radiology 2009;6. https://doi.org/10.1016/j.jacr.2008.12.013.

[4] Dong P, Xing L. DoseNet: A Deep Neural Network for Accurate Dosimetric Transformation between Different Spatial Resolutions and/or Different Dose Calculation Algorithms for Precision Radiation Therapy. International Journal of Radiation Oncology*Biology*Physics 2019;105:S255–6. https://doi.org/10.1016/J.IJROBP.2019.06.2471.

[5] Gronberg MP, Gay SS, Netherton TJ, Rhee DJ, Court LE, Cardenas CE. Technical Note: Dose prediction for head and neck radiotherapy using a three-dimensional dense dilated U-net architecture. Med Phys 2021;48:5567–73. https://doi.org/10.1002/MP.14827.

[6] Kearney V, Chan JW, Wang T, Perry A, Descovich M, Morin O, et al. DoseGAN: a generative adversarial network for synthetic dose prediction using attention-gated discrimination and generation. Sci Rep 2020;10. https://doi.org/10.1038/s41598-020-68062-7.

[7] Babier A, Mahmood R, McNiven AL, Diamant A, Chan TCY. Knowledge-based automated planning with three-dimensional generative adversarial networks. Med Phys 2020;47. https://doi.org/10.1002/mp.13896.

[8] Zhan B, Xiao J, Cao C, Peng X, Zu C, Zhou J, et al. Multi-constraint generative adversarial network for dose prediction in radiotherapy. Med Image Anal 2022;77:102339. https://doi.org/10.1016/J.MEDIA.2021.102339.

[9] Teng L, Wang B, Xu X, et al.. Beam-wise dose composition learning for head and neck cancer dose prediction in radiotherapy (2024). Medical Image Analysis. https://doi.org/10.1016/j.media.2023.103045.

[10] Fornander H. Denoising Monte Carlo Dose Calculations Using a Deep Neural Network (2019). https://WwwDiva-PortalOrg/Smash/Get/Diva2:1366439/FULLTEXT01Pdf 2019 (accessed August 16, 2023).

[11] Javaid U, Souris K, Dasnoy D, Huang S, Lee JA. Mitigating inherent noise in Monte Carlo dose distributions using dilated U-Net. Med Phys 2019;46:5790–8. https://doi.org/10.1002/MP.13856.

[12] Fix MK, Manser P, Frei D, Volken W, Mini R, Born EJ. An efficient framework for photon Monte Carlo treatment planning*. Phys Med Biol 2007;52:N425. https://doi.org/10.1088/0031-9155/52/19/N01.

[13] Ronneberger, O., Fischer, P., Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Navab, N., Hornegger, J., Wells, W., Frangi, A. (eds) Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. MICCAI 2015. Lecture Notes in Computer Science(), vol 9351. Springer, Cham. https://doi.org/10.1007/978-3-319-24574-4_28

[14] Ghimire K, Chen Q, Feng X. Patch-Based 3D UNet for Head and Neck Tumor Segmentation with an Ensemble of Conventional and Dilated Convolutions. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 12603 LNCS, 2021. https://doi.org/10.1007/978-3-030-67194-5_9.

[15] Pérez-García F, Sparks R, Ourselin S. TorchIO: A Python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning. Comput Methods Programs Biomed 2021;208. https://doi.org/10.1016/j.cmpb.2021.106236.

[16] Kingma DP, Ba JL. Adam: A Method for Stochastic Optimization. 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings 2014.

[17] Venselaar J, Welleweerd H, Mijnheer B. Tolerances for the accuracy of photon beam dose calculations of treatment planning systems. Radiotherapy and Oncology 2001;60. https://doi.org/10.1016/S0167-8140(01)00377-2.

[18] Tsekas G, Bol GH, Raaymakers BW. Robust deep learning-based forward dose calculations for VMAT on the 1.5T MR-linac. Phys Med Biol 2022;67:225020. https://doi.org/10.1088/1361-6560/AC97D8.

[19] Wang L, Ye J, Zhao Y, et al. SuperNeurons: Dynamic GPU Memory Management for Training Deep Neural Networks. Arxive 2018. https://doi.org/10.48550/arXiv.1801.04380

[20] C. Szegedy et al., "Going deeper with convolutions," 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 2015, pp. 1-9, doi: 10.1109/CVPR.2015.7298594.

[21] Smilowitz, J.B., Das, I.J., Feygelman, et al. (2015), AAPM Medical Physics Practice Guideline 5.a.: Commissioning and QA of Treatment Planning Dose Calculations — Megavoltage Photon and Electron Beams. Journal of Applied Clinical Medical Physics, 16: 14-34. https://doi.org/10.1120/jacmp.v16i5.5768

[22] Failla GA, Wareing T, Archambault Y, Thompson S. Acuros® XB Advanced Dose Calculation for the EclipseTM Treatment Planning System 2015. https://varian.widen.net/s/n2hmf2pshr/acuros_clinicalperspectives_rad10156a_april2015 (accessed August 16, 2023).

Development of a Range Probe to Improve Dose Delivery Accuracy in Proton Beam Therapy

Josephine Jones[1,2], Adam Aitkenhead[1,2], Samuel Manger[1,2], and Michael Taylor[1,2]

[1]University of Manchester, Division of Cancer Sciences, Manchester, UK
[2]The Christie NHS Foundation Trust, Manchester, UK

Abstract Proton therapy requires accurate calculation of proton range in the patient. Several factors are potential sources of range error, and in-vivo range verification could be used to improve range accuracy. However, no methods for in-vivo range verification are currently in routine clinical use. A proton range probe is a possible solution, which we define as a small number of pencil beams that traverse the patient to measure water equivalent thickness (WET) and verify calculations from X-ray CT. The proposed range probe is based on time-of-flight (TOF) measurements. A pair of sensors upstream of the patient measure initial proton energy and a pair of sensors downstream measure final proton energy. The mean energy change can be used to calculate the WET along the proton path using a calibration curve. Monte Carlo modelling of the TOF range probe was done using Geant4 to investigate how sensor parameters affect the accuracy and precision of WET measurement and quantify sensor requirements. Simulations with homogeneous materials demonstrated WET uncertainty below 1% with sensor timing resolutions ≤60ps. WET projections of an inhomogeneous phantom were also simulated and used to assess the performance of the probe near boundaries of different media.

1 Introduction

Range error is a significant challenge in proton beam therapy treatment planning. It arises due to several factors, including dose calculation error and patient set-up uncertainty. Most significantly, the treatment is planned using an X-ray CT and CT Hounsfield units must be converted into proton stopping power. The stoichiometric method is the most common approach to generating a CT-proton calibration curve but was found to generate up to 3.4% error for some treatment sites[1].

There is currently no method of in-vivo proton range verification in clinical use. Several methods have been proposed including PET, prompt gamma imaging and proton-CT. These require complicated and expensive technology, and it could be several years before they are implemented in the clinic.

Proton range probes have been suggested as an alternative method of range verification. In this study, a range probe is defined by a small number of proton pencil beams that fully traverse the patient to measure the water equivalent thickness (WET) along their path. The purpose of the range probe is to make measurements that complement data from X-ray CT.

The use of multi-layer ionization chambers for range probing has previously been investigated [2][3]. These devices are limited by their large size, low sensitivity and poor spatial resolution. This study will explore an alternative range probe design, based on time-of-flight (TOF) measurements.

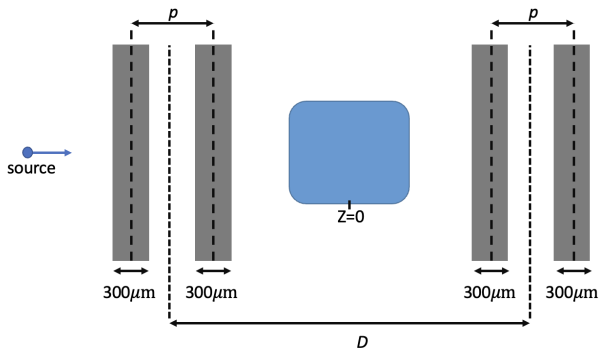


Figure 1: Schematic of range probe detector set-up (not to scale). Four identical 300μm silicon sensors, with two upstream and two downstream of the object to be measured, which is at the beam isocentre.

2 Materials and Methods

2.1 Geant4 Simulation Toolkit

Geant4 is a Monte Carlo toolkit for the simulation of particle transport through matter (https://geant4.web.cern.ch/). Version 11.0.3 was used to simulate a TOF detector for use as a proton range probe. These simulations used the '*QGSP_BIC*' physics list with EM option 3, as recommended for medical applications.

2.2 TOF Detector Design

The TOF range probe consists of two sensors upstream and two sensors downstream of the object to be imaged. In simulations, the sensors were modelled by 300μm of silicon. Each pair of sensors had a plane separation p, and the two pairs were separated by a flight distance, D. This setup can be seen in Figure 1. The object was placed at the beam isocenter.

2.3 Kinetic Energy Calculation

Analysis was performed on a proton-by-proton basis. The proton's hit time, t_i, and position, (x_i,y_i) on each sensor ($i = 1,2,3,4$) were recorded. The hit data were filtered to include only those protons that recorded exactly one hit on each sensor. The distance a single proton travelled between the

planes is described by Equation 1 and its TOF by Equation 2.

$$\Delta d_{initial} = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + p^2} \quad (1)$$

$$\text{TOF}_{initial} = \Delta t_{initial} = t_2 - t_1 \quad (2)$$

The initial proton energy was then calculated using the relativistic energy equation:

$$E_{initial} = m_p c^2 \left(\frac{1}{\sqrt{1 - \frac{v_{initial}^2}{c^2}}} - 1 \right) \quad (3)$$

Where c is the speed of light, m_p is proton mass and $v_{initial} = \frac{\Delta d_{initial}}{TOF_{initial}}$. The final proton energy, E_{final}, was calculated in the same way, using the hit data for sensors 3 and 4. A Gaussian distribution was then fitted to find the mean energy change ΔE.

2.4 Calibration Curves

Calibration curves were generated that related the average change in proton energy, ΔE, to the WET of the path along which it travelled. Simulations were performed with water blocks of thicknesses between 5mm and 340mm. For each thickness, 10^7 protons were simulated and ΔE was found, as described in section 2.3. Calibration curves with varying sensor timing resolution are shown in Figure 2.

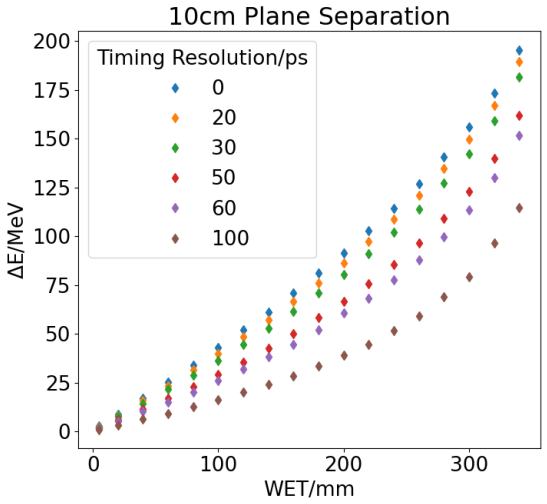


Figure 2: Calibration curves with 1m flight distance for 10cm plane separation and varying timing resolution.

2.5 Investigation of Detector Parameters

The aim of this investigation was to quantify the detector requirements for achieving accurate and precise WET measurements. Intrinsic sensor specifications (timing resolution, σ_t, and spatial resolution, σ_{xy}), as well as set-up geometry (plane separation, p, and flight distance, D), were investigated. The number of protons simulated, n, was also varied. The values used for each parameter are shown in Table 1.

Timing and spatial resolution were modelled by Gaussian blurring of hit times and positions with distributions of width σ_t and σ_{xy}. Measurements were simulated in blocks of water, PMP and Teflon with WET values between 10mm and 300mm. The simulation was run 100 times for every measurement. For each run, n protons were simulated and ΔE fitted. An estimate of the WET was then obtained from the calibration curve and the mean and standard deviation of the estimates were calculated across 100 runs.

Parameter	Values
σ_t/ps	0,20,30,50,60,100
σ_{xy}/mm	0.1,0.5,1,5
p/cm	5,10,15,20
D/m	0.5,1,1.5
n	10^5,10^6,10^7

Table 1: Detector parameters and the values that were used in an investigation of WET accuracy and precision.

2.5.1 True WET Calculation

The WET of a block with physical thickness T can be calculated using Equation 4. The values R_{water} and $R_{material}$ are the CSDA (constant slowing down approximation) ranges for protons in the specified material. The function '*GetCSDARange*' in the Geant4 '*G4EMCalculator*' module was used to obtain the CSDA ranges for water, PMP and Teflon. For each WET value investigated, the required physical thicknesses of PMP and Teflon were then calculated.

$$\text{WET} = T \frac{\text{R}_{water}}{\text{R}_{material}} \quad (4)$$

2.6 Measurements in an Inhomogeneous Phantom

Measurements were also simulated in an inhomogeneous object. The phantom consists of an ethylene cylinder with two cylindrical inserts of tissue-equivalent materials; dense bone and lung. Pencil beams were simulated through a CT of the phantom on a two-dimensional grid, at 2mm increments along x and y. The phantom was oriented with its axis along the z-direction, such that the beam was perpendicular to the circular face. These simulations were run with 10^6 protons, p=10cm and D =1m. For each probe position, ΔE was fitted and the WET value was obtained from the appropriate calibration curve.

2.6.1 True WET projection

A WET projection of the phantom along the beam path was generated from the X-ray CT. For each voxel, the material was determined from the HU value, and the corresponding CSDA range was used to calculate the WET with Equation 4. The WET values were then summed along the z-axis to generate a two-dimensional map (Figure 4(a)).

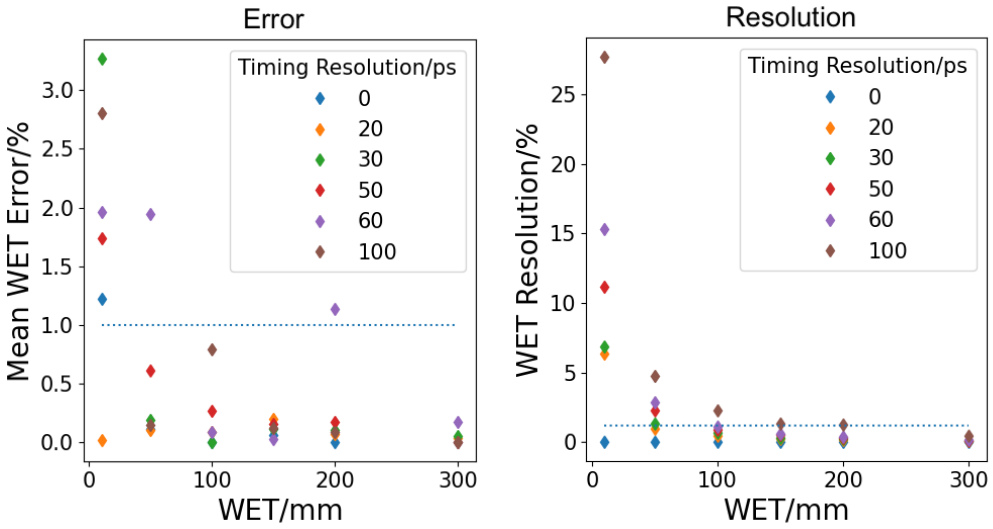


Figure 3: Results of WET measurements in water, with 10^6 protons, 10cm plane separation and 1m flight distance (a) Mean WET percentage error (b) WET resolution.

3 Results

3.1 Parameter Investigation

Results of WET measurements in water are shown in Figure 3. Generally, measurement accuracy and resolution are better at higher WET values and the resolution improves as timing resolution decreases. There is no clear relationship between accuracy and timing resolution. Greater plane separation and larger proton number improve measurement resolution. There was no difference in results between measurements in water, PMP and teflon. Spatial resolution and flight distance were not found to affect results. Overall, measurements made with a timing resolution of 60ps or better and at least 10^6 protons had accuracy and resolution better than 1%.

3.2 WET Image of Phantom

The phantom WET projection from the CT image is shown in Figure 4(a), the phantom range probe measurements in Figure 4(b) and the absolute difference between the two in Figure 4(c). Qualitatively, it can be seen that greater errors occur near boundaries of different media, for example at the edges of the phantom and the inserts. Error in the homogeneous regions is generally below 1mm for measurements with 0ps timing resolution. Near the edge of the phantom, however, the range error is as high as 50mm. This is caused by lateral proton scattering, meaning individual protons pass along paths of different WET. The introduction of an angular cut, to include only those protons that travelled along a near straight path could be used to improve the calculation[4]. It was also found that the average magnitude and spread in the WET error increase as the timing resolution increases.

4 Discussion

This detector is intended for use as a range probe with a small number of pencil beams. A method must be developed for choosing the optimal positions of these beams, which should take into account both measurement accuracy and clinical application.

The method proposed relies on proton-by-proton tracking. Therapeutic proton beams deliver a particle flux of up to 10^{10} protons per second. The data acquisition system must, therefore, be able to operate at MHz frequencies to avoid pile-up effects[5]. Methods have also been proposed to reduce the beam intensity. For example, the OPTIma project has explored the use of a collimator to reduce beam currents to 10pA[6]. Further work will explore the compatibility of the TOF range probe with a clinical beam. It will consider alternatives to proton-by-proton tracking using statistical methods.

The intrinsic timing resolution of the sensors will also affect the ability to make single proton measurements. As uncertainty in the arrival time increases, it becomes harder to identify which hits correspond to the same proton. Furthermore, as timing resolution increases, the resolution of the WET measurement worsens. The results suggest a timing resolution of $\leq$60ps would be required, which has been achieved with some detectors such as low-gain avalanche diodes (LGADs)[7].

A major advantage of this TOF technique is that a large range of WET values can be measured. This is because protons exiting the patient only need enough energy to travel as far as the final sensor. The largest WET that can be measured with the probe, however, is still limited by the maximum proton energy. A further limitation of range probing is that it is a projection of the WET along the whole path, including tissue outside of the treatment area.

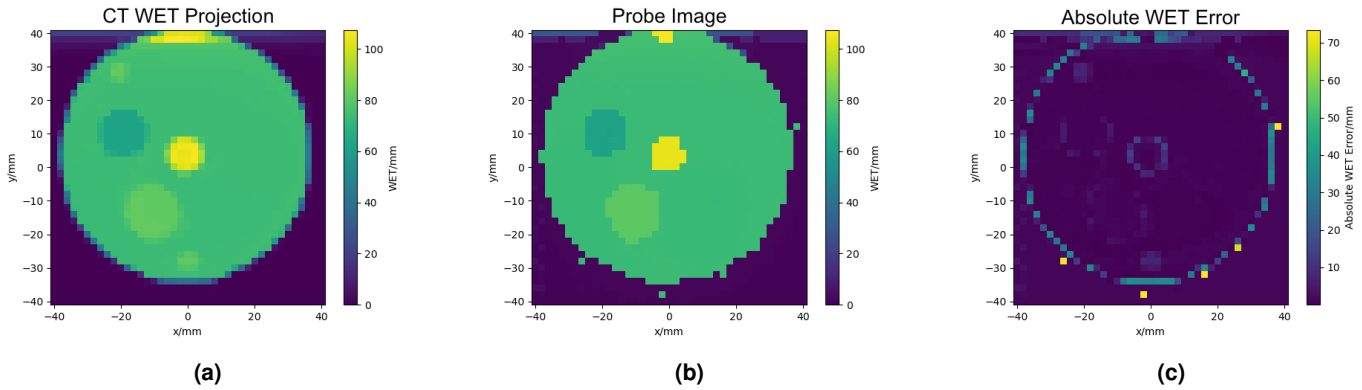


Figure 4: WET map of inhomogeneous phantom (a) WET projection from x-ray CT (b) Simulated range probe WET measurements (c) Absolute range error map

Experimental validation of these results is required. From a clinical perspective, there is an additional need to assess the best use of the probe. Among the questions to be addressed are the number of pencil beams required and at which point in the treatment process measurements will be made. The answers will depend on measurement dose, speed and accuracy.

5 Conclusion

It has been shown that a TOF range probe can be used to measure WET with an accuracy better than 1% and a resolution close to the range straggling limit. This is achievable with sensor timing resolutions better than 60ps. Experimental validation of this range probe is required, as well as an evaluation of the dose delivered during measurement.

References

[1] M. Yang, X. R. Zhu, P. C. Park, et al. "Comprehensive analysis of proton range uncertainties related to patient stopping-power-ratio estimation using the stoichiometric calibration". *Physics in Medicine & Biology* 57.13 (2012), pp. 4095–4115. DOI: 10.1088/0031-9155/57/13/4095.

[2] M. Mumot, C. Algranati, M. Hartmann, et al. "Proton range verification using a range probe: definition of concept and initial analysis". *Physics in Medicine & Biology* 55 (2010), pp. 4771–4782. DOI: 10.1088/0031-9155/55/16/010.

[3] A. Hammi, L. Placidi, D. C. Weber, et al. "Positioning of head and neck patients for proton therapy using proton range probes: a proof of concept study". *Physics in Medicine & Biology* 63 (2017). DOI: 10.1088/1361-6560/aa9cff.

[4] B. Schultze, P. Karbasi, C. Sarosiek, et al. "Particle-Tracking Proton Computed Tomography-Data Acquisition, Preprocessing, and Preconditioning". *IEEE access : practical innovations, open solutions* 9 (2021), pp. 25946–25958. DOI: 10.1109/ACCESS.2021.3057760.

[5] K. Parodi and J. C. Polf. "In vivo range verification in particle therapy." *Medical physics* 45 (2018), e1036–e1050. DOI: 10.1002/mp.12960.

[6] A. Winter, A. H. Aitkenhead, N. M. Allinson, et al. "OPTIma: a tracking solution for proton computed tomography in high proton flux environments". *Journal of Instrumentation* 18 (2023). DOI: 10.1088/1748-0221/18/04/P04026.

[7] R. Heller, C. Madrid, A. Apresyan, et al. "Characterization of BNL and HPK AC-LGAD sensors with a 120 GeV proton beam". *Journal of Instrumentation* 17 (2022). DOI: 10.1088/1748-0221/17/05/P05001.

Implementation of the Mayo Clinic Florida Microdosimetric Kinetic Model and AMDM into GATE/Geant4: Investigation of the impact of nozzle configurations

Hermann Fuchs[1,2], Alessio Parisi[2], Keith Furutani[2], Dietmar Georg[1], and Chris Beltran[2]

[1]Department of Radiation Oncology, Medical University of Vienna, Vienna, Austria
[2]Department of Radiation Oncology, Mayo Clinic Florida, Jacksonville, FL, United States of America

Abstract In ion beam therapy a higher relative biological effectiveness (RBE) with respect to X-rays is well established. The RBE is known to be changing dynamically depending on multiple factors requiring complex biological models. A posteriori recalculation of most of these biological models using e.g. different cell-line data are challenging, requiring a complete recalculation of the whole treatment plan.

The abridged microdosimetric distribution methodology (AMDM) attempting to solve this issue, was integrated into the open-source Monte Carlo toolkit OpenGATE10. Furthermore, the calculation of RBE based on the Mayo Clinic Florida microdosimetric kinetic model (MCF MKM) was implemented and used to compare the RBE of two nozzle designs of the MedAustron and Mayo Clinic Florida. First investigations indicate that the nozzle design does not markedly affect the RBE. However, for mono energetic carbon ion beams changes in the initial energy spread due to different ripple filter designs may have a non-negligible influence.

1 Introduction

Radiation therapy uses ionizing radiation to treat malignant tissues. Due to their physical characteristics, ion beam therapy allows for improved dose conformity compared to conventional photon-based irradiation. In addition to the physical dose characteristics, ion beam therapy is general characterized by higher the relative biological effectiveness (RBE) with respect to the X-rays used in conventional radiation therapy.

For protons, the RBE is attributed to be 1.1, while for carbon ions, the second clinically most used ion, the RBE is known to be changing dynamically depending on multiple factors such as energy or penetration depth. So far two biological models to calculate the RBE are employed clinically, the local effect model (LEM1) [1] and the microdosimetric kinetic model (MKM) [2]. However, both models were found to have limitations, leading to the development of improved models. The Mayo Clinic Florida microdosimetric kinetic model (MCF MKM) [3] aims to predict the in-vitro clonogenic survival in the low as well as the high linear energy transfer (LET) range using computationally efficient single-event microdosimetric spectra.

The multitude of existing biological models indicates the need of model comparisons. Due to the implementational details of most of these models a posteriori calculations using different cell-line data or even different model formulations are challenging, requiring a complete recalculation of the whole treatment plan. Our group recently introduced a methodology potentially enabling a much faster and efficient comparison of biological modelling aspects [4].

The Mayo Clinic Florida (MCF) announced the construction of a synchrotron-based ion-therapy center being able to provide proton and carbon ion for treatment [5]. Ion therapy centers from different vendors are in operation all over the world. Although they all provide nominally comparable carbon ion beams in terms of energy or spot size, due to different design approaches especially of the beam delivery system and respective nozzles, it is not fully clear whether the different materials involved might alter the radiation quality, potentially affecting the RBE.

We have modelled the treatment nozzles of the MedAustron ion therapy center (MedAustron iontherapy center, Wr. Neustadt, Austria) and the Mayo Clinic Florida (Mayo Clinic Florida, Jacksonville, FL, USA) using the GATE Monte Carlo code. After implementing a method to calculate the RBE, we compared the biological impact of both nozzle designs.

2 Materials and Methods

Monte Carlo tool kit

The development version of OpenGATE10, based upon the Monte Carlo toolkit Geant4 was employed for this study. OpenGATE10 is designed to lower the user threshold for Monte Carlo simulations as much as possible. It uses a python wrapper, allowing user interaction directly from python scripts, without having to refer to C++ or macro commands. Installation can be performed using a simple pip install command without requiring the user to manually download and compile multiple software packages. OpenGATE10 is an open-source toolkit for Monte Carlo simulations initially developed for nuclear imaging applications, but later extended towards other modalities such as ion beam therapy [6–9].

Biological actor implementation

The abridged microdosimetric distribution methodology (AMDM) [4] was implemented as an actor into OpenGATE10. An actor allows to extract data from an OpenGATE simulation. AMDM creates spatially resolved multi-dimensional output which can then be used in post processing to calculate the RBE. This approach enables

the use different biological models or input parameters with the same simulation output, without having to redo time and resource intense Monte Carlo calculations. For our study post processing was performed using data from the HSG cell line and the MCF MKM biological model [10].

Nozzle modelling

For the MedAustron ion therapy center a pre-existing model of the treatment nozzle was employed [11,12]. The particles leaving the treatment nozzle were stored into separate phase space files for each treatment energy.

For the MCF treatment nozzle, confidential manufacturer data sheets were used to model the nozzle components in OpenGATE10 directly. It features an extensible nozzle allowing to adjust the distance between the nozzle exit and the isocenter in the range from 20 to 120 cm. For simplicity only the extended nozzle was used in this manuscript.

As at the facility is currently under construction, no existing beam data could be used. Instead, the beam parameters were modelled using design specifications. To facilitate comparison, the ranges of the MCF system was matched with the MedAustron ranges in terms of the range at 80% dose fall off (R80).

Simulations

Simulations were performed using a development version of OpenGATE10 alongside Geant4 v11.1.1. After exiting the nozzle (in case of the MCF nozzle) or the corresponding phase space (for the MedAustron nozzle) the beam impinged a water phantom with dimensions of 50x40x40 cm³. The physics list QGSP_BIC_EMZ with global cuts of 10 cm and smaller cuts of 0.1 mm in the water phantom were employed.

A total of 10^6 carbon ions were simulated for each set-up.

Mono energetic carbon ion beams with a range in water of 3 and 27 cm, corresponding to approximately 120 and 400 MeV/u were used, termed low and high energy in this manuscript, respectively.

Range shifters consisting of water were inserted for the 27 cm carbon ion beams. The thickness of the range shifters was chosen such, that the R80 of the higher energy was reduced to match the 3 cm range in R80 of the lower energetic beams. The carbon ion energy spectra at the beginning of the water phantom were evaluated.

The laterally integrated dose deposited in each voxel, using a resolution of 0.1 mm was stored. In addition, the multidimensional AMDM output, consisting of 23 additional parameters was stored in three files (restricted energy deposition, dose-mean lineal energy histogram, fraction of absorbed dose deposited by lineal energy events histogram).

Using a python script, the RBE for each voxel was calculated based on AMDM data using the MCF MKM after the simulation was finished. For RBE calculations the HSG cell line was used with RBE = $RBE_\alpha = \alpha_{AMDM}/\alpha_{ref}$.

3 Results

The R80 between MedAustron and MCF beamlines were matched for the extended MCF nozzle to be within 0.1 mm. Differences between the peak heights were observed within 5 to 14%.

In Figure 1 a comparison of laterally integrated depth-dose curves is displayed. The range at 80% dose fall-off was matched to be within 0.1 mm between beams with and without range shifter, respectively. As expected, the Bragg-peak width is increased for the initially higher energetic particle beams.

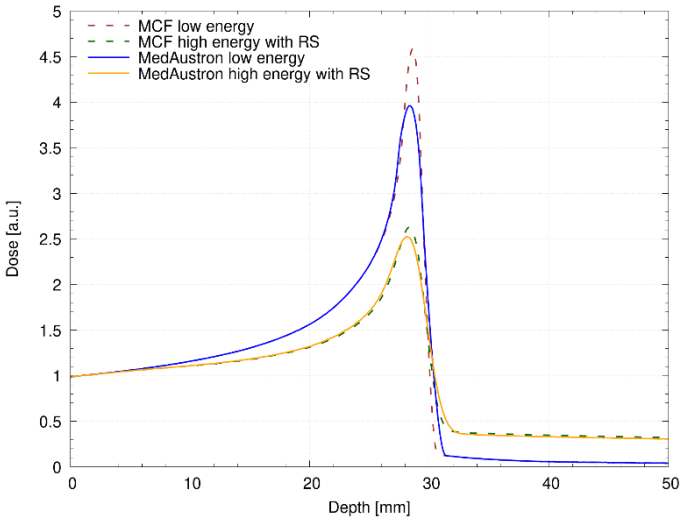


Figure 1 Laterally integrated depth-dose curves of carbon ions for the MCF and MedAustron nozzles. Range shifters were introduced to match the ranges to 3 cm for all initial energies and beam lines. Profiles were normalized at a depth of 1 mm.

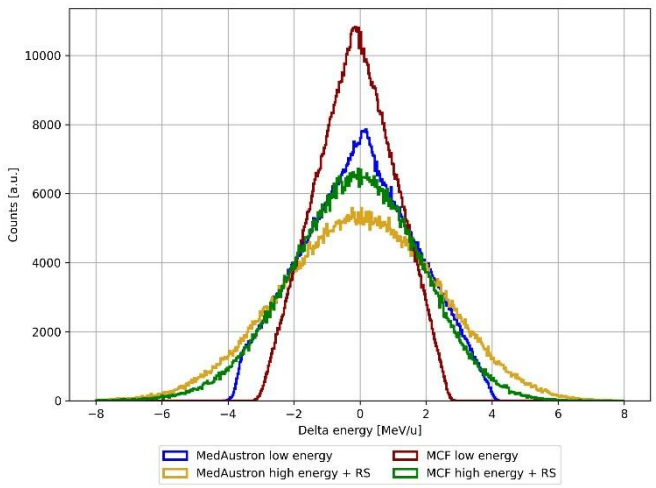


Figure 2: Energy spectra at the water phantom start for all investigated configurations. Data was normalized to the mean energy of the MedAustron low energy beam.

A comparison of the energy spectra (see Figure 2) showed that the overall energies were similar, with only minor deviations of the mean due to the range matching accuracy.
As expected, the width of the energy spread increased for the initially higher energetic particle beams, compared to the low energetic beams. The energy spread of the MCF nozzle was found to be narrower compared to the MedAustron nozzle.

In Figure 2Figure 3 the corresponding RBE of the mono energetic particle beams in water is shown. As expected, the maximum RBE of the broader Bragg peaks of the high energy carbon ions including a range shifter was up to 14% lower compared to the low energy beams. The average RBE of both nozzles agreed within 0.1%.

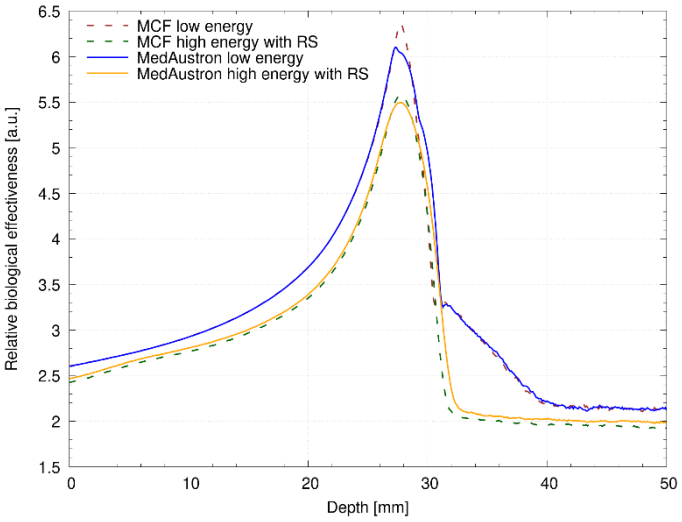


Figure 3. Relative biological effectiveness (RBE) of carbon ion beams for the MCF and MedAustron nozzle. Range shifters were introduced to match the ranges to 3 cm for all initial energies and beam lines.

4 Discussion

The overall shapes of the Bragg peaks for the MCF and MedAustron nozzle were found to be very similar. Differences between the peak heights are most likely related to the different specifications of the ripple filters. This was substantiated by the energy spectra. For the dose contribution a sensitivity study indicated that a variation in the initial energy spread from 0.01% to 0.1% did not have a considerable effect.

Multiple coulomb scattering resulted in larger Bragg peak width for the carbon ion beams transversing the water range shifters compared to the lower energetic carbon ions. This was also apparent in the energy spectra.

The RBE curves were similar with some notable changes. A change in RBE after the Bragg peak was apparent for all particles traversing the range shifter, indicating a small change in beam quality.

Changes in RBE between the MCF and MedAustron nozzle were most likely related to the slightly different employed ripple filters and correspondingly different energy spread. Mono energetic carbon ion beams might be the most sensitive to changes in this regard. A more realistically comparison would be to using spread out Bragg peaks instead of mono energetic ones. Such a comparison is currently ongoing.

5 Conclusion

The abridged microdosimetric distribution methodology (AMDM) was integrated into the open-source Monte Carlo toolkit OpenGATE10. Furthermore, the calculation of RBE based on the Mayo Clinic Florida microdosimetric kinetic model (MCF MKM) was successfully tested.
First investigations indicate that the nozzle itself does not markedly affect the radiation quality and the RBE. The average RBE of both nozzles agreed within 0.1%. However, for mono energetic carbon ion beams changes in the initial energy spread due to different ripple filter designs may have non-negligible influences.

References

[1] Elsässer, Krämer, Scholz. Int J Radiat Oncol Biol Phys 2008;71:866–72. https://doi.org/10.1016/j.ijrobp.2008.02.037.
[2] Inaniwa, Furukawa, Kase, Matsufuji, Toshito, Matsumoto, et al. Phys Med Biol 2010;55:6721–37. https://doi.org/10.1088/0031-9155/55/22/008.
[3] Parisi, Beltran, Furutani. Phys Med Biol 2022;67. https://doi.org/10.1088/1361-6560/ac7375.
[4] Parisi, Beltran, Furutani. J Appl Clin Med Phys 2023;24:1–9. https://doi.org/10.1002/acm2.14049.

[5] Beltran, Amos, Rong. J Appl Clin Med Phys 2020;21:6–9. https://doi.org/10.1002/acm2.13133.
[6] Jan, Santin, Strul, Staelens, Assié, Autret, et al. Phys Med Biol 2004;49:4543–61. https://doi.org/10.1088/0031-9155/49/19/007.
[7] Grevillot, Boersma, Fuchs, Bolsa-Ferruz, Scheuchenpflug, Georg, et al. Front Phys 2021;9:1–14. https://doi.org/10.3389/fphy.2021.704760.
[8] Padilla- Cabal, Alejandro Fragoso, Franz Resch, Georg, Fuchs. Med Phys 2020;47:223–33. https://doi.org/10.1002/mp.13883.
[9] Jan, Benoit, Becheva, Carlier, Cassol, Descourt, et al. Phys Med Biol 2011;56:881–901. https://doi.org/10.1088/0031-9155/56/4/001.
[10] Parisi, Beltran, Furutani. Phys Med Biol 2022;67:185013. https://doi.org/10.1088/1361-6560/ac7375.
[11] Fuchs, Elia, Resch, Kuess, Lühr, Vidal, et al. Med Phys 2021;48:841–51. https://doi.org/10.1002/mp.14647.
[12] Elia, Resch, Carlino, Böhlen, Fuchs, Palmans, et al. Physica Medica 2020;71:115–23. https://doi.org/10.1016/j.ejmp.2020.02.006.

Diffusion Schrödinger Bridge Models for High-Quality MR-to-CT Synthesis for Head and Neck Proton Treatment Planning

Muheng Li[1,2], Xia Li[1,3], Sairos Safai[1], Damien Weber[1,4,5], Antony Lomax[1,2], and Ye Zhang[1]

[1]Center for Proton Therapy, Paul Scherrer Institut, Villigen
[2]Department of Physics, ETH Zurich, Zurich
[3]Department of Computer Science, ETH Zurich, Zurich
[4]Department of Radiation Oncology, University Hospital of Zurich, Zurich
[5]Department of Radiation Oncology, Inselspital, Bern University Hospital, University of Bern, Bern

Abstract In recent advancements in proton therapy, MR-based treatment planning is gaining momentum to minimize additional radiation exposure compared to traditional CT-based methods. This transition highlights the critical need for accurate MR-to-CT image synthesis, which is essential for precise proton dose calculations. Our research introduces the Diffusion Schrödinger Bridge Models (DSBM), an innovative approach for high-quality MR-to-CT synthesis. DSBM learns the nonlinear diffusion processes between MR and CT data distributions. This method improves upon traditional diffusion models by initiating synthesis from the prior distribution rather than the Gaussian distribution, enhancing both generation quality and efficiency. We validated the effectiveness of DSBM on a head and neck cancer dataset, demonstrating its superiority over traditional image synthesis methods through both image-level and dosimetric-level evaluations. The effectiveness of DSBM in MR-based proton treatment planning highlights its potential as a valuable tool in various clinical scenarios.

1 Introduction

In radiation therapy, proton therapy stands out for its exceptional precision and ability to minimize radiation exposure to non-target tissues. In traditional proton therapy treatment planning, reliance on Computed Tomography (CT) scans is standard practice. However, CT imaging for pre-treatment planning, as well as daily verification, often results in extra radiation doses to the patient. Recognizing this challenge, there's a growing interest in exploring Magnetic Resonance (MR) imaging as a potential alternative, given its ability to provide detailed anatomical information without the associated radiation exposure [1]. This paradigm shift towards MR-based proton treatment planning underscores the need for high-quality MR-to-CT image synthesis, where attaining both anatomical precision and accurate density information within synthetic CTs is essential for proton dose calculations.

Recent advancements in AI-based generative imaging techniques, particularly the Diffusion Models (DMs) [2], have shown promising results for natural image generation tasks [3]. DMs have also been proven for their capacity to produce detailed and realistic synthetic medical images [4]. However, traditional DMs often involve lengthy sampling chains, which can significantly reduce sampling efficiency. Furthermore, given the higher image generation variability compared to conventional approaches, DMs may have limitations in clinical MR-to-CT synthesis, where both image generation precision and stability are crucial.

To tackle these challenges, we propose the Diffusion Schrödinger Bridge Model (DSBM) for MR-to-CT synthesis. Unlike conventional diffusion methods that sample from a Gaussian distribution, this approach directly learns the nonlinear diffusion processes between MR and CT data distributions. This method can establish a more direct transition model between MR and CT data, enhancing image generation quality and stability while increasing efficiency through fewer sampling steps, thereby boosting its potential application in proton therapy clinical practices.

2 Materials and Methods

2.1 Diffusion Schrödinger Bridge Models (DSBM)

Schrödinger Bridge (SB), as introduced by Schrödinger in 1932 [5], represents an entropy-regularized optimal transport model that incorporates the following forward-backward Stochastic Differential Equations (SDEs):

$$\mathrm{d}X_t = \left[f_t + g_t^2 \nabla \log \Psi_t\left(X_t\right)\right]\mathrm{d}t + g_t\, \mathrm{d}W_t, \tag{1a}$$

$$\mathrm{d}X_t = \left[f_t - g_t^2 \nabla \log \widehat{\Psi}_t\left(X_t\right)\right]\mathrm{d}t + g_t\, \mathrm{d}\bar{W}_t, \tag{1b}$$

where $X_t \in \mathbb{R}^d$ is a d-dimensional stochastic process, indexed by discrete time steps $t \in [0, T]$. The Wiener process is represented as W_t, while its reverse-time process is denoted as $\bar{W}_t$. Here, X_0 is sampled from the boundary distribution p_{CT}, representing CT image data, while X_T is drawn from p_{MR}, representing the paired MR image data. The functions $\Psi_t, \widehat{\Psi}_t$ represent time-dependent energy potentials that satisfy the following coupled Partial Differential Equations (PDEs):

$$\begin{cases} \frac{\partial \Psi_t(x)}{\partial t} = -\nabla \Psi^\top f - \frac{1}{2} g^2 \nabla^2 \Psi \\ \frac{\partial \widehat{\Psi}_t(x)}{\partial t} = -\nabla \cdot (\widehat{\Psi} f) + \frac{1}{2} g^2 \nabla^2 \widehat{\Psi} \end{cases} \tag{2}$$
$$\text{s.t. } \Psi_0(x)\widehat{\Psi}_0(x) = p_{\mathrm{CT}}(x), \Psi_T(x)\widehat{\Psi}_T(x) = p_{\mathrm{MR}}(x)$$

However, addressing the SB models involves simulating stochastic processes and requires resource-intensive iterative procedures [6, 7]. Consequently, this approach faces challenges in terms of tractability and practical applicability.

To tackle this problem, we implement the concepts from [8], transforming the SB model into the following SDEs:

$$\mathrm{d}X_t = f_t(X_t)\,\mathrm{d}t + g_t\,\mathrm{d}W_t, X_0 \sim \widehat{\Psi}_0, \tag{3a}$$

$$\mathrm{d}X_t = f_t(X_t)\,\mathrm{d}t + g_t\,\mathrm{d}\bar{W}_t, X_T \sim \Psi_T. \tag{3b}$$

This form adheres to the fundamental structure of the score-based generative models (SGM) [9], where $\nabla \log \widehat{\Psi}_t(X_t)$ and $\nabla \log \Psi_t(X_t)$ are the *score functions* of the above SDEs. Then, we delve into practical strategies for applying the above model to MR-to-CT synthesis. Given the fact that the availability of paired information during training, where $p(X_0, X_T) = p_{\mathrm{CT}}(X_0)\, p_{\mathrm{MR}}(X_T \mid X_0)$, with p_{CT} and p_{MR} corresponding to the CT and MR data distributions respectively. This framework allows us to construct tractable Schrödinger Bridges between X_0 and $p_{\mathrm{MR}}(X_T \mid X_0)$. According to [8], we can adopt an analytic posterior given the boundary pair MR-CT data (X_0, X_T):

$$\begin{gathered} q(X_t \mid X_0, X_T) = \mathcal{N}(X_t; \mu_t(X_0, X_T), \Sigma_t), \\ \mu_t = \frac{\bar{\sigma}_t^2}{\bar{\sigma}_t^2 + \sigma_t^2} X_0 + \frac{\sigma_t^2}{\bar{\sigma}_t^2 + \sigma_t^2} X_T, \quad \Sigma_t = \frac{\sigma_t^2 \bar{\sigma}_t^2}{\bar{\sigma}_t^2 + \sigma_t^2} \cdot I. \end{gathered} \tag{4}$$

Here, we define $\sigma_t^2 = \int_0^t g_\tau^2 \mathrm{d}\tau$, $\bar{\sigma}_t^2 = \int_t^1 g_\tau^2 \mathrm{d}\tau$. Then, based on the network parameterization settings in [3] with a 3D U-Net [10], the training and sampling procedures can be done under the same framework as [8].

2.2 Dataset and Model Implementations

This study used a dataset of head-and-neck images from 58 patients treated at PSI between 2017 and 2022, divided into training (46 patients) and testing (12 patients) sets. CT scans were performed using a Siemens Sensation Open CT scanner (resolution: 1 × 1 × 2 mm), and MR scans with a 1.5T Siemens Aera MR scanner (voxel size: 1 × 1 × 1 mm), primarily using the T1 vibe Dixon sequence. Data pre-processing, including a neural network-based data registration step, followed protocols outlined in [11]. We strictly adhered to ethical standards in human data research, obtaining informed consent for using anonymized patient data and ensuring confidentiality and ethical compliance. DSBM was comprehensively compared with two other established medical image synthesis approaches, namely the pix2pix model [12] and the nnU-Net model [13]. While the nnU-Net model utilized the same 3D data inputs as DSBM, the pix2pix model was trained on 2D image slices.

2.3 Image-level Evaluation

We evaluated the accuracy and anatomical fidelity of synthetic CT (sCT) and planning CT (pCT) using various image-related metrics, including: Mean Absolute Error (MAE) on Hounsfield Units (HU), Peak Signal-to-Noise Ratio (PSNR), and Structural Similarity Index (SSIM) for intensity value mapping, focusing on different regions of interest (RoIs). For anatomical conformity, Dice Coefficients (Dice) were calculated to assess structural overlaps, especially in bone regions (HU greater than 250).

2.4 Dosimetric-level Evaluation

Dosimetric evaluations were performed to validate the clinical utility of the generated sCT images in proton therapy. Treatment plans were optimized on pCT using in-house developed Juliana software [14], and doses were recalculated for sCTs as direct comparison. We assessed Mean Dose Error at various dose thresholds (10%, 50%, and 90%), and used gamma pass rates (1%/1mm and 2%/1mm) for gamma analysis between planned and recalculated dose distributions. Dose-Volume Histograms (DVHs) and Differential Dose-Volume Histograms (DDVHs) were analyzed, providing insights into the dosimetric impact of using sCTs. We further investigated the variations in several typical dose indices, specifically V95, D95, D2, Dmean, and Vdosediff>3% (the percentage of voxels in the same structure with an absolute point dose difference larger than 3% of the prescribed dose), for both PTVs and OARs, to comprehensively understand the accuracy of the new MR-to-CT synthesis approach for calculating proton therapy treatment plans.

3 Results

At the image level, DSBM outperformed conventional nnU-Net and pix2pix models in the testing set. It achieved a lower Mean Absolute Error (MAE) of 72.30 ± 8.31 HU and a higher Dice score of 83.10 ± 3.90%, indicating enhanced anatomical accuracy, especially in bone structures (Table 1). Visual comparisons in Figure 1 further corroborate this, with error histograms showing DSBM's reduced errors and improved fidelity in synthetic CT images (sCTs).

At the dosimetric level, DSBM also demonstrated improved agreement for calculating doses when compared to those calculated on the corresponding pCT. It showed mean voxel-wise differences in the body region of less than 0.01% of the prescribed dose (Table 2), and the 1%/1mm gamma pass rate of 95.53 ± 2.89% in the RoI (Table 3). Figure 2 offers a visual comparison between synthetic (sDoses) and planned doses (pDose) in terms of voxel-wise dose differences, DVHs and DDVHs of a single case. Figure 3 compares the differences of dosimetric indices for PTVs and OARs across all 12 cases in the testing set. These results underscore DSBM's dosimetric accuracy, confirming its efficacy and potential utility in clinical proton therapy settings.

4 Discussion

The evaluations in our study, both at the image and dosimetric levels, have demonstrated the superiority of the Diffusion Schrödinger Bridge Model (DSBM) approach over

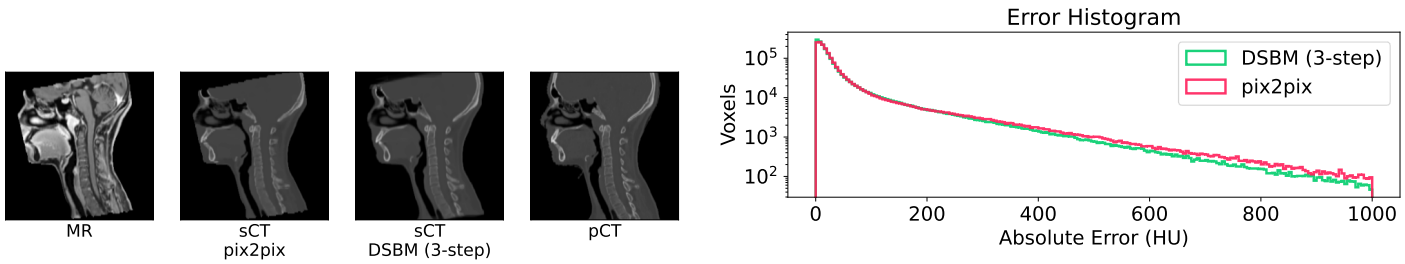


Figure 1: Visual comparisons of sCTs and error histograms from pix2pix and DSBM. Errors are calculated as $|sCT - pCT|$.

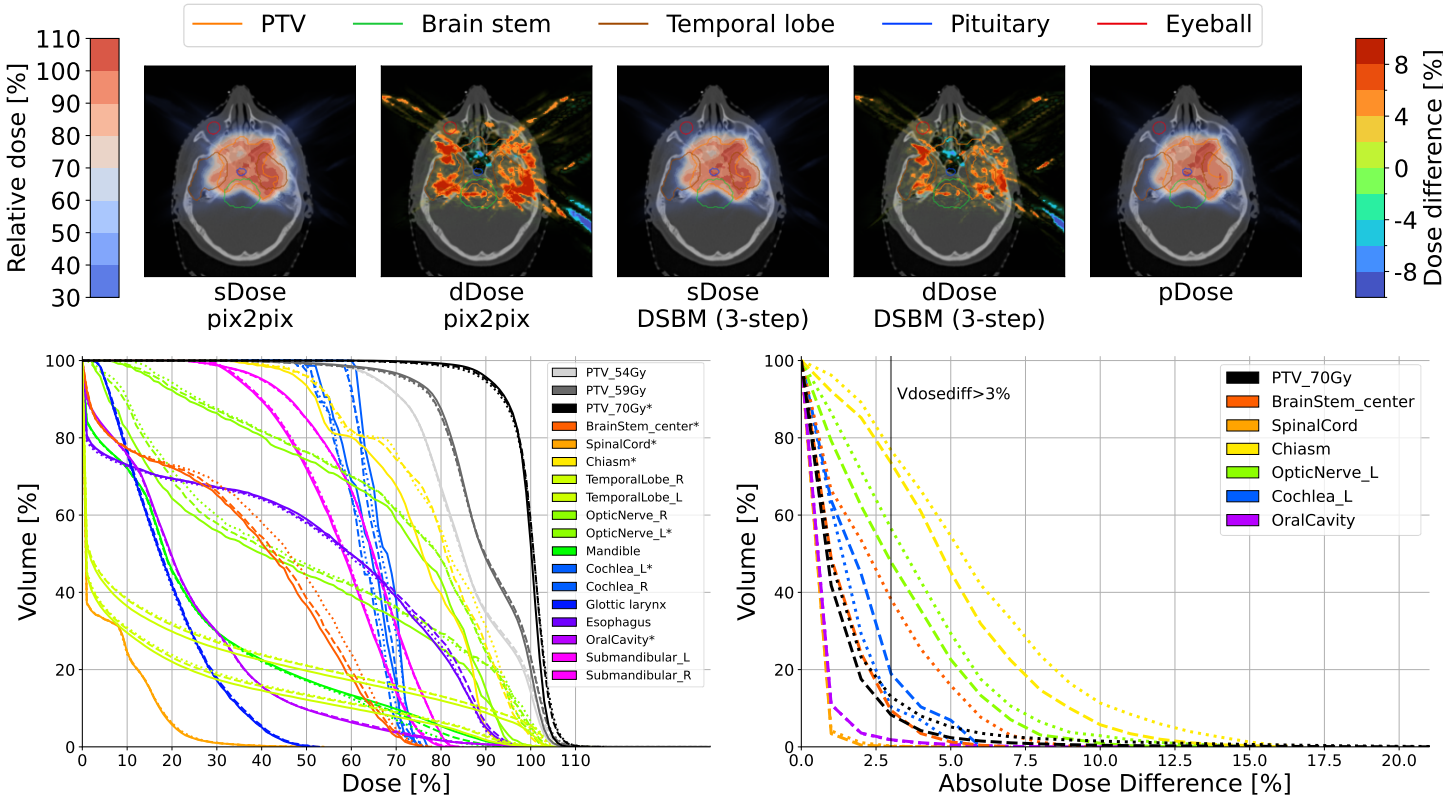


Figure 2: Dosimetric study of a single case. **(Top)** Comparison of dose distribution. dDose was calculated as sDose - pDose. **(Bottom Left)** DVHs, with structures highlighted in the DDVH graph denoted by (*). **(Bottom Right)** DDVHs. Representation for DVHs and DDVHs: solid lines (—) for pDose, dotted lines (...) for pix2pix model, dashdot (- - -) lines for DSBM (50-step).

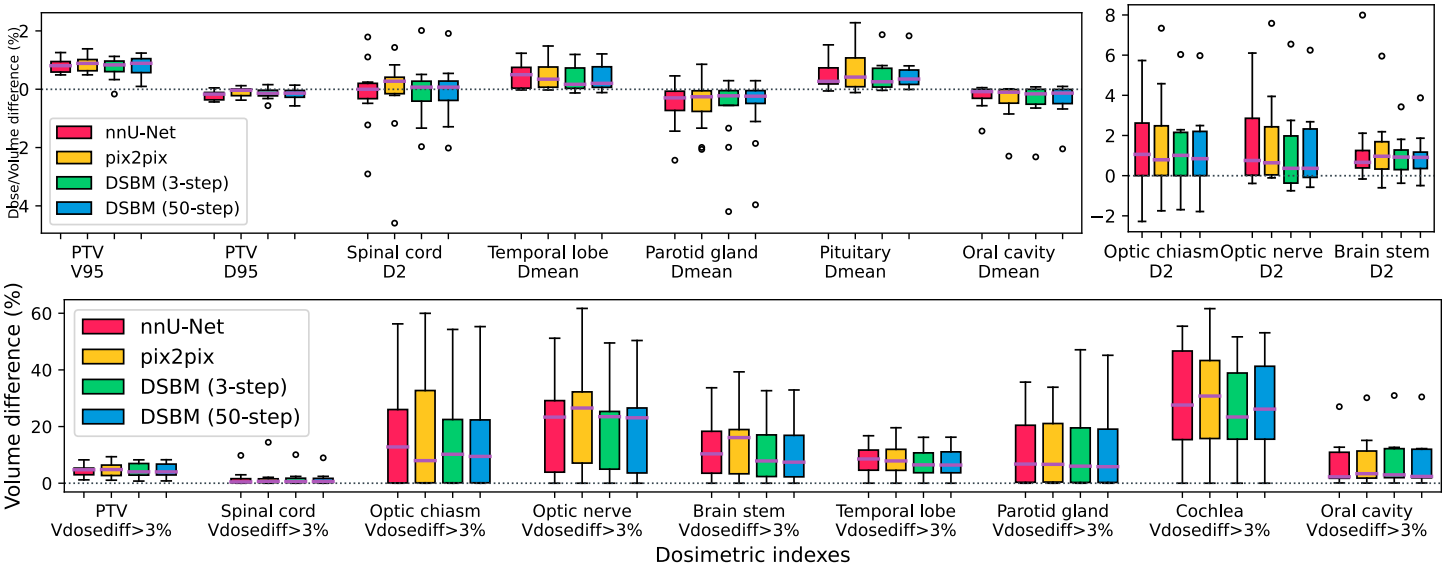


Figure 3: Comparison of the dosimetric accuracy of the sCTs (N=12), in terms of (**upper**) differences in typical dose indexes from corresponding DVHs, and (**bottom**) differences in Vdosediff>3% from corresponding DDVH.

Table 1: Quantitative comparisons between results from DSBM (using different number of neural function evaluation steps) and conventional approaches.

Method	MAE (HU)	PSNR (dB)	SSIM (%)	Dice_bone (%)
nnU-Net	80.44±9.02	27.12±1.09	80.52±2.72	81.55±4.00
pix2pix	77.41±7.81	27.50±0.89	81.95±2.68	80.72±3.74
DSBM (1-step)	73.00±8.41	28.16±1.07	83.09±2.72	82.95±3.92
DSBM (3-step)	**72.30±8.31**	**28.20±1.07**	**83.26±2.69**	**83.10±3.90**
DSBM (10-step)	72.64±8.25	28.19±1.05	83.21±2.71	82.99±3.90
DSBM (50-step)	73.29±8.28	28.13±1.05	83.13±2.72	82.83±3.89

Table 2: Mean Dose Error (%) from different methods by different thresholds.

Method	Dose_body	Dose>10%	Dose>50%	Dose>90%
nnU-Net	0.04±0.05	0.13±0.09	0.39±0.22	0.40±0.25
pix2pix	0.02±0.05	0.08±0.10	0.32±0.23	0.37±0.25
DSBM (1-step)	-0.01±0.07	**0.06±0.13**	**0.26±0.32**	**0.31±0.32**
DSBM (3-step)	**0.00±0.07**	0.08±0.12	0.29±0.31	0.33±0.30
DSBM (10-step)	0.01±0.06	0.09±0.11	0.31±0.29	0.35±0.29
DSBM (50-step)	0.01±0.06	0.09±0.11	0.32±0.29	0.36±0.28

conventional classic MR-CT synthesis methods. The DSBM-generated CT images not only exhibit remarkable accuracy and anatomical fidelity but also demonstrate dosimetric accuracy, vital for applications like adaptive or even MR-only proton treatment planning. A notable aspect of the DSBM method is its efficiency in generating high-quality results even with a limited number of neural function evaluation steps (NFEs). Our findings reveal that the DSBM can still outperform conventional methods when operating in very few sampling steps, even in a single step. In practical terms, generating a volume batch of size $128 \times 128 \times 4$ pixels takes only about 0.2 seconds per sampling step with DSBM. It should be noticed that conventional diffusion models require over 100 sampling steps for comparable results, which can lead to a generation time of more than 20 seconds per volume batch. Consequently, the DSBM method offers a speed enhancement of 10 to 50 times, marking a substantial leap in efficiency. Moreover, increasing the number of sampling steps does not necessarily correlate with further outcome improvement. We postulate that this phenomenon could be attributed to the introduction of noise at each step due to the inherent randomness in the sampling process.

5 Conclusion

In this study, we propose the Diffusion Schrödinger Bridge Models (DSBM) which enhances the performance of MR-to-CT image synthesis, both in quality and efficiency. By learning the nonlinear diffusion processes between MR and CT data distributions and initiating synthesis from a prior distribution, DSBM outperforms traditional methods, as validated in our head and neck cancer case studies. This approach has demonstrated its value through comprehensive evaluations at both the image and dosimetric levels. DSBM can improve the accuracy and efficiency of MR-based treatment planning in proton therapy, which could potentially reduce imaging-related dose and improve patient outcomes in radiotherapy.

Table 3: Gamma pass rate (%) of sDoses over different thresholds.

Method	1%/1mm	2%/1mm
nnU-Net	95.28±3.19	96.01±2.96
pix2pix	95.42±2.95	96.15±2.72
DSBM (1-step)	95.47±2.99	96.18±2.77
DSBM (3-step)	95.51±2.93	96.22±2.70
DSBM (10-step)	95.53±2.91	**96.23±2.68**
DSBM (50-step)	**95.53±2.89**	96.22±2.70

References

[1] A. Hoffmann, B. Oborn, M. Moteabbed, et al. "MR-guided proton therapy: a review and a preview". *Radiation Oncology* 15.1 (2020), pp. 1–13.

[2] J. Ho, A. Jain, and P. Abbeel. "Denoising diffusion probabilistic models". Vol. 33. 2020, pp. 6840–6851.

[3] P. Dhariwal and A. Nichol. "Diffusion models beat gans on image synthesis". *NeurIPS*. Vol. 34. 2021, pp. 8780–8794.

[4] F. Khader, G. Müller-Franzes, S. Tayebi Arasteh, et al. "Denoising diffusion probabilistic models for 3D medical image generation". *Scientific Reports* 13.1 (2023), p. 7303.

[5] E. Schrödinger. "Sur la théorie relativiste de l'électron et l'interprétation de la mécanique quantique". 2.4 (1932), pp. 269–310.

[6] V. De Bortoli, J. Thornton, J. Heng, et al. "Diffusion Schrödinger bridge with applications to score-based generative modeling". *NeurIPS*. Vol. 34. 2021, pp. 17695–17709.

[7] T. Chen, G. Liu, and E. A. Theodorou. "Likelihood training of Schrödinger bridge using forward-backward SDEs theory". *ICLR*. 2022.

[8] G. Liu, A. Vahdat, D. Huang, et al. "I^2SB: Image-to-Image Schrödinger Bridge". *ICML*. Vol. 202. 2023, pp. 22042–22062.

[9] Y. Song, J. Sohl-Dickstein, D. P. Kingma, et al. "Score-based generative modeling through stochastic differential equations". *ICLR*. 2021.

[10] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, et al. "3D U-Net: learning dense volumetric segmentation from sparse annotation". *MICCAI*. Springer. 2016, pp. 424–432.

[11] X. Li, R. Bellotti, B. Bachtiary, et al. "A unified generation-registration framework for improved MR-based CT synthesis in proton therapy". *arXiv preprint arXiv:2401.12878* (2024).

[12] P. Isola, J.-Y. Zhu, T. Zhou, et al. "Image-to-image translation with conditional adversarial networks". *CVPR*. 2017, pp. 1125–1134.

[13] F. Isensee, J. Petersen, A. Klein, et al. "nnU-Net: Self-adapting framework for u-net-based medical image segmentation". *arXiv preprint arXiv:1809.10486* (2018).

[14] R. Bellotti, J. Willmann, A. J. Lomax, et al. "JulianA: An automatic treatment planning platform for intensity-modulated proton therapy". *arXiv preprint arXiv:2305.10211* (2023).

Self super-resolution of anisotropic volumes in prostate MRI with normalized edge priors

Vincent Jaouen[1,2], Pierre-Henri Conze[1,2], Julien Bert[2,3], and Dimitris Visvikis[2]

[1]IMT Atlantique, Brest, France
[2]UMR 1101 INSERM LaTIM, IMT Atlantique, UBO, Brest, France
[3]CHRU Brest, France

Abstract Due to involuntary organ motion, prostate magnetic resonance images are often acquired in 2D with thick through plane spacings and reconstructed into anisotropic voxelized volumes. In prostate cancer (PCa) radiotherapy, annotations such as zonal or lesion segmentation are then performed on the few axial slices in which the prostate appears, leading to coarse and often inconsistent delineations across slices, which are nevertheless used as ground truth for deep segmentation model training and performance reports.
Recently, deep self-supervised super resolution (SSR) was proposed in brain MRI to remove the need for calibrated training data, by learning the low-resolution to high-resolution mapping from the anisotropic image itself. In this work, we propose a new training objective (loss function) for SSR based on the alignment of normalized image gradient vectors, a technique normally used for image registration. Preliminary experiments on the public Prostate158 dataset suggest that automatic PCa detection could be improved using SSR with the proposed training objective. However, the low quality of lesion annotations obtained on a few slices is a major bottleneck to confirm these findings quantitatively.

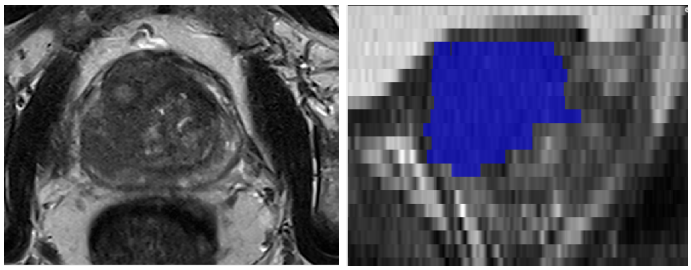
1 Introduction

1.1 Context

Prostate cancer (PCa) is a major form of cancer that affects about one man in six during his lifetime [1]. Approximately 75% of PCa originate in the peripheral zone of the gland (PZ), as opposed to the more central transition zone (TZ) [2]. Radiotherapy treatment planning of PCa relies on multiparametric magnetic resonance imaging (mpMRI) comprising T2-weighted and diffusion-weighted sequences for lesion detection and characterization [3].

1.2 Image anisotropy

Due to involontary movements, recommendations are to perform fast 2D T2-weighted pulse sequences, preferably along the three anatomical planes but sometimes along the axial direction only [3]. However, such acquisitions lead to high resolution (HR) in-plane images (e.g. from 0.5×0.5 mm^2 down to 0.25×0.25 mm^2, Fig. 1a) but to low resolution (LR) through-plane images due to large spacings (typically of around 3 mm, Fig. 1b). Such anisotropy in the data representation may hamper image analysis downstream. In particular, the detection and characterization of the PCa lesions from such degraded acquisitions is difficult and subjected to high inter reader variability. In this context, the automatization of these tasks of high clinical relevance.



(a) Axial view **(b)** Sagittal + lesion "ground truth"

Figure 1: Example 2D axial prostate MRI acquisition with high in-plane resolution (0.27 mm^2) and thick through-plane spacing (3 mm). The lesion mask is also imprecise with strong "staircasing" effect.

1.3 Objective

In this work, we study whether enhancing image resolution through Single Image Super-Resolution (SISR) techniques can improve PCa lesion deep learning-based segmentation. More specifically, we leverage a self-supervised super resolution (SSR) method, SMORE, recently proposed for brain MR SISR [4]. SMORE alleviates the need for any training data outside the anisotropic image itself. In this context, we propose a new reconstruction objective (loss function) to improve SMORE results based on the alignment of normalized gradient fields (NGF) between the LR and HR planes, a technique normally used for image registration [5]. Preliminary results on the Prostate158 public dataset suggest that SSR is indeed beneficial to PCa detection, and that the addition of the proposed NGF loss term may further improve the quality of the detection.

2 Related works

The majority of modern computerized medical image analysis approaches rely on (deep) machine learning techniques. A canonical example is the U-Net convolutional neural network (CNN) model [6] for segmentation [7] and detection [8]. When dealing with highly anisotropic data, little to no gain is achieved using 3D models, and 2D models are generally preferred [7]. However, 2D models inherently discards the consistency of the signal along the (degraded) third dimension, which is suboptimal.

Among major research directions, careful curation and preprocessing of input data is generally considered a critical aspect for the improvement of learning-based model performances [7]. In this work, we hypothesize that improving the resolution of highly anisotropic images through self-supervised SISR techniques can be beneficial to a variety of downstream image analysis tasks.

2.1 Single image super resolution

Current state of the art SISR methods generally require curated external supervised training data with predetermined downsampling factors to learn a mapping between the LR and HR spaces using e.g. deep residual networks [9]. However, such training data may be difficult to obtain in practice, as data collections are typically gathered from multiple centers with heterogeneous practices. Moreover, these methods are subject to the domain shift between training and deployment conditions.

2.2 Self-supervised super-resolution with SMORE

Recently, Zhao et al. proposed Synthetic Multi-Orientation Resolution Enhancement (SMORE), a self-supervised super resolution (SSR) approach for anisotropic volumes [4]. SMORE removes the need for external training data as the LR to HR mapping is learned directly on the 3D volume itself (i.e. between the two LR planes and the HR plane). First, SMORE proposes a forward degradation model on the HR 2D slices to mimick the degradation in the LR planes due to the 2D acquisition protocol, taking into consideration both aliasing due to sampling in the slice direction and blur. In practice, the degradation is achieved through downsampling followed by upsampling and blurring. A supervised deep-learning super-resolution model, such as EDSR [10] is then used to learn the mapping between synthetically degraded LR patches and HR patches.
SMORE operates under a dual objective that combines a pixel-level reconstruction loss (using the ℓ_1 norm) with an edge loss ensuring similarity between image gradients using a Sobel filter.
Once trained, the model is applied to LR orientations (here sagittal and coronal, without loss of generality). Finally, a unique HR isotropic volume is reconstructed by merging volumes upsampled by SSR in the frequency domain through Fourier burst accumulation [11].

3 Method

3.1 Problem statement

Let $\boldsymbol{x}_{\mathrm{LR}}$ be a LR T2 image having thick through-plane spacing along the last dimension, i.e. a resolution of $r_x \times r_y \times r_z$ mm^3, with voxel spacing $r_z > r_x$ and where $r_y = r_x$ for simplicity. We hypothesize that $\boldsymbol{x}_{\mathrm{LR}}$ is a downsampled version of an unobserved HR image $\boldsymbol{x}_{\mathrm{HR}}$ of superior resolution $r_x \times r_x \times r_x$ mm^3. Our objective is to recover an estimate $\hat{\boldsymbol{x}}_{\mathrm{HR}} = \mathscr{G}(\boldsymbol{x}_{\mathrm{LR}})$ from $\boldsymbol{x}_{\mathrm{LR}}$ using a neural network $\mathscr{G}$, which is a typical single image super resolution (SiSR) problem [12].

3.2 Normalized Gradient Field loss

The original SMORE is trained under a joint ℓ_1 reconstruction and Sobel gradient objective, the latter promoting edge similarity. However, gradient similarity prioritizes high-contrast edges over low-contrast ones (i.e. of weaker gradient magnitude). As a result, some of the finer aspects of the image might be lost or blurred in the enhancement process. To counter this effect, we propose to give equal weight to weak and strong edges through gradient normalization. The idea of normalizing gradients is popular in image registration to obtain contrast-independent similarity metrics and known as the normalized gradient field (NGF) [5].
The NGF loss if defined as :

$$\mathscr{L}_{NGF} := 1 - \frac{1}{2}\langle \boldsymbol{n}_\varepsilon(\boldsymbol{x}_{\mathrm{HR}}), \boldsymbol{n}_\varepsilon(\mathscr{G}(\boldsymbol{x}_{\mathrm{LR}}))\rangle^2, \tag{1}$$

where $\langle\cdot,\cdot\rangle$ is the dot product and $n_\varepsilon(\boldsymbol{x})$ is the NGF :

$$\boldsymbol{n}_\varepsilon(\boldsymbol{x}) = \frac{\nabla \boldsymbol{x}}{\|\nabla \boldsymbol{x}\| + \varepsilon}, \tag{2}$$

with ε a noise threshold scalar parameter controlling the influence of noise, with higher values corresponding to higher suppression. The reason for normalization is to encourage gradient vector alignment between generated and target images while being agnostic to contrast.
In this work, we thus propose to use the NGF in the context of image super resolution by training the SSR model under a joint ℓ_1 and a NGF objective (instead of the Sobel loss).

4 Experiments and Results

4.1 Experiments

The evaluation of super resolution can be done either using reconstruction metrics when ground truth HR images are available, or task-based considering a downstream objective (e.g. segmentation, detection, prediction, ...). In this work, we studied to what extent the proposed SSR approach can be beneficial to the training of deep learning-based PCa lesion detection models. In our experiments, we used the public Prostate158 dataset [13] constituted of 158 patient image data, of which we selected 63 images with annotated PCa lesions on axial T2-weighted MR images, showing voxel anisotropy ranging from $0.27 \times 0.27 \times 3$ mm^3 to $0.47 \times 0.47 \times 3$ mm^3. Images were clipped to the 1%-99% percentile and rescaled in the range [0-1] following common practice. The noise parameter ε was set to 0.1. Note that the annotations were performed on the LR axial images and are thus imperfect ground truth.
The SSR method was implemented using PyTorch and Kornia and applied individually to each image of the training and

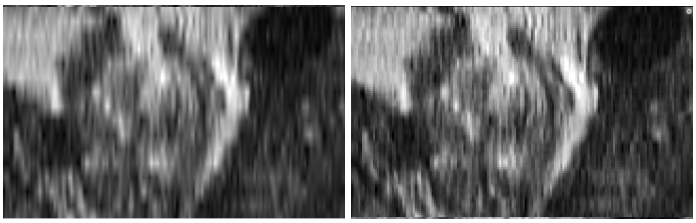
testing set. The SSR model required approximately one hour per image on a NVIDIA RTX A6000 GPU. Once processed by the various upsampling methods, images were fed to a 3D full resolution nn-UNet model [7] for 1000 epochs using 5-fold cross validation and ensembling. Our baseline comparisons comprised results achieved the original SMORE model (using a joint ℓ_1 and Sobel loss), or by letting the nnU-Net chose the resampling space dynamically using its inner heuristics (resampling to the median resolution of the training set). We used the first 47 images in lexicographic order patients for training and the remaining 16 for testing. Detection evaluation was performed using volumetric overlap between predicted and real lesion masks. A detection was considered positive if the Dice score exceeded 0.2.

4.2 Results

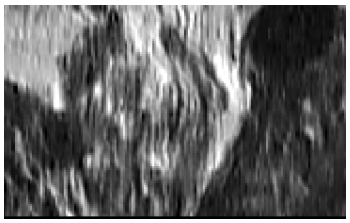
Fig. 2 shows representative upsampling results obtained in one patient of the Prostate158 set, with much better recovery of high frequency details using SSR, and especially when considering the NGF loss with a clearly sharper output. Out of the 16 images tested, 8 were deemed unsuccessful as they recorded a Dice score of less than 0.2 across all methodologies. This can be attributed to the inherent challenges associated with prostate cancer (PCa) segmentation on complex images. Achieving accurate segmentation in these cases often requires the support of additional modalities, typically Apparent Diffusion Coefficient (ADC) MRI, which were not utilized in our study. However, using SMORE allowed for a better detection on average, which translated into better quantitative scores (Table 1). Note one failure case for nnU-Net interpolation and one failure case for baseline SMORE. Our proposed NGF training objective for SMORE allowed better results on average, with a Dice score of around 50. However, these quantitative results are to be taken with caution due to the very limited quality of the lesion segmentation in such anisotropic images. For instance, patient 118 has only two annotated axial slices.

Patient ID	nnU-Net interp.	SMORE	SMORE+NGF
112	18.1	**29.0**	13.7
118	59.2	**71.2**	62.2
125	**64.5**	62.8	34.1
131	**72.7**	69.2	70.6
133	79.4	**80.3**	79.9
138	31.7	46.9	**59.8**
140	30.8	0.0	**53.6**
148	0.0	9.3	**60.0**
Mean	44.55	46.09	**54.24**
Std Dev	26.59	28.25	**19.69**

Table 1: Dice scores for prostate cancer detection using three different upsampling methods.



(a) Resampled, 3rd order B-Spline **(b)** SMORE ($\ell_1 + \text{Sobel}$)



(c) SMORE ($\ell_1 + \text{NGF}$)

Figure 2: Upsampling result in the Prostate158 dataset. Self Super-Resolution with the proposed NGF objective lead to better detail recovery and sharper outputs.

5 Conclusion

In this preliminary work, we have investigated whether self supervised resolution could be used to improve the detection of PCa images from T2-weighted MR images. Leveraging the SMORE model, we have proposed a new loss function that aims at promoting the alignment of normalized gradient vectors between high resolution and low-resolution patches during training. Preliminary results tend to suggest a positive impact of such an approach, both visually and quantitatively. However, further experiments need to be conducted in a more controlled environment to ensure the interest of the approach, for instance using better annotations or synthetically degraded pseudo low resolution images. Other downstream tasks will also be investigated, such as e.g. segmentation of vascular structure in liver CT images for surgical planning.

References

[1] O. W. Brawley, D. P. Ankerst, and I. M. Thompson. “Screening for prostate cancer”. *CA: a cancer journal for clinicians* 59.4 (2009), pp. 264–273.

[2] R. J. Cohen, B. A. Shannon, M. Phillips, et al. “Central zone carcinoma of the prostate gland: a distinct tumor type with poor prognostic features”. *The Journal of urology* 179.5 (2008), pp. 1762–1767.

[3] T. Barrett, B. Turkbey, and P. L. Choyke. “PI-RADS version 2: what you need to know”. *Clinical radiology* 70.11 (2015), pp. 1165–1176.

[4] C. Zhao, B. E. Dewey, D. L. Pham, et al. “SMORE: a self-supervised anti-aliasing and super-resolution algorithm for MRI using deep learning”. *IEEE transactions on medical imaging* 40.3 (2020), pp. 805–817.

[5] E. Haber and J. Modersitzki. “Intensity gradient based registration and fusion of multi-modal images”. *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2006: 9th International Conference, Copenhagen, Denmark, October 1-6, 2006. Proceedings, Part II 9*. Springer. 2006, pp. 726–733.

[6] O. Ronneberger, P. Fischer, and T. Brox. "U-net: Convolutional networks for biomedical image segmentation". *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer. 2015, pp. 234–241.

[7] F. Isensee, P. F. Jaeger, S. A. Kohl, et al. "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation". *Nature methods* 18.2 (2021), pp. 203–211.

[8] P. F. Jaeger, S. A. Kohl, S. Bickelhaupt, et al. "Retina U-Net: Embarrassingly simple exploitation of segmentation supervision for medical object detection". *Machine Learning for Health Workshop*. PMLR. 2020, pp. 171–183.

[9] R. Sood and M. Rusu. "Anisotropic super resolution in prostate MRI using super resolution generative adversarial networks". *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. IEEE. 2019, pp. 1688–1691.

[10] B. Lim, S. Son, H. Kim, et al. "Enhanced deep residual networks for single image super-resolution". *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 2017, pp. 136–144.

[11] M. Delbracio and G. Sapiro. "Burst deblurring: Removing camera shake through fourier burst accumulation". *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015, pp. 2385–2393.

[12] W. Yang, X. Zhang, Y. Tian, et al. "Deep learning for single image super-resolution: A brief review". *IEEE Transactions on Multimedia* 21.12 (2019), pp. 3106–3121.

[13] L. C. Adams, M. R. Makowski, G. Engel, et al. "Prostate158-An expert-annotated 3T MRI dataset and algorithm for prostate cancer detection". *Computers in Biology and Medicine* 148 (2022), p. 105817.

Patient-specific simultaneous CT and MR reconstruction from sparse projection measurements using implicit neural representation

Oscar Pastor-Serrano, Liang Qiu, Siqi Ye, and Lei Xing

Department of Radiation Oncology, Stanford University, Stanford CA, United States

Abstract Radiotherapy workflows can significantly benefit from the simultaneous acquisition of paired computed tomography (CT) and magnetic resonance (MR) images, particularly in MR-guided treatments. In this study, we introduce a patient-specific algorithm designed to precisely reconstruct CT scans and their corresponding MR intensities using sparse measurements (i.e., as few as 10 projections). We propose a novel framework based on a generalizable implicit neural representation (GINR) with two channels, a method that incorporates prior patient information into the weights of a neural network mapping spatial coordinates to CT or MR values. With only a few projections, GINR can reconstruct the new CT anatomy and the corresponding MR within a few minutes. GINR achieves a peak signal-to-noise ratio (PSNR) of 40.13±1.05 dB when reconstructing brain CTs from 10 projections, and 26.64±0.57 dB for their corresponding MR counterparts, establishing itself as a fast and accurate alternative to traditional algorithms.

1 Introduction

Imaging plays a pivotal role in radiation therapy (RT) treatments, from patient diagnosis to treatment simulation and patient positioning and monitoring during treatment delivery. Computed tomography (CT) stands as the reference image modality in radiotherapy, providing both the patient's geometry and the electron densities needed for the calculation of the physical radiation dose [1]. Alternatively, the utilization of magnetic resonance (MR) images, such as for tumor localization or treatment delivery guidance, is swiftly gaining prominence. This surge is attributed to the heightened soft tissue contrast and diminished radiation exposure that MR provides [2].

The success of an RT treatment depends, among other factors, on the speed and quality of critical steps, such as image acquisition and the subsequent organ delineation and dose calculation based on the acquired images. These speed requirements become notably stringent in adaptive or MR-guided treatments. In these approaches, aimed at minimizing treatment uncertainties and tumor margins, the entire workflow, possibly involving re-planning, is compressed into just a few seconds before or during the delivery of each treatment fraction. Hence, fast and accurate imaging emerges as a pivotal factor for the success of such treatments.

Currently, CT and MR image acquisition is formulated as an inverse problem aimed at reconstructing the final image from measured sensor data. Generally, obtaining low-noise images necessitates dense sampling in measurement space, posing challenges for both image modalities, e.g., increasing the dose delivered by CT projections and the MR reconstruction times. Furthermore, while ideally employed in tandem to leverage their respective strengths and weaknesses, pairing CT and MR images presents its own set of challenges. Firstly, both image modalities must be independently acquired within the shortest time possible to minimize anatomical changes, thereby increasing the total treatment time. Secondly, the registration of both images to the same coordinate system introduces additional uncertainties. Only after both steps have been completed can one, in principle, calculate the outcome of the treatment, such as the dose distribution, which is also impacted by the quality of the images used as input.

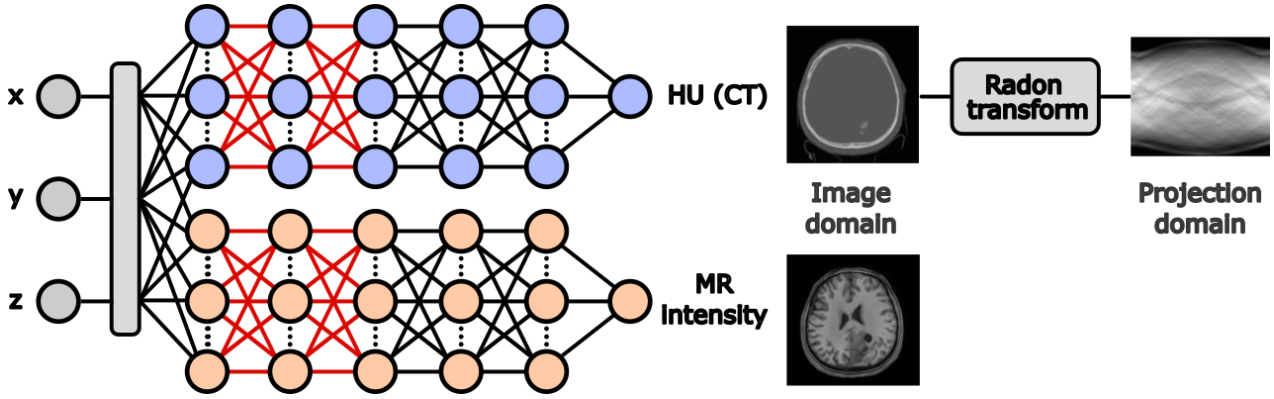


Figure 1: ***Overall generalizable implicit representation workflow.*** *A two-branch multi-layer perceptron maps spatial coordinates to CT, and MR or dose values. First, any prior information (few available scans from a single patient) can be directly embedded into the weights of each branch by minimizing the square difference between the network outputs and the prior images. The weights of the first layers (in red), containing transformations specific to each geometry, are shared between branches. Once obtaining new projections, these shared weights can be adjusted via backpropagation through the Radon transform operation.*

To minimize treatment time, image reconstruction must leverage sparse measurement data. Likewise, simultaneous CT/MR image acquisition is highly preferable. Recent deep learning approaches that learn a mapping function between synthetic CT/MR [3] face the challenge of choosing between speed and accuracy. This dilemma arises either by opting to wait until the source image is obtained from dense measurements or by compromising performance when using a sub-optimal sparsely reconstructed source image.

Recent advancements in implicit neural representation methods such as NeRP [4] represent the culmination of a series of works leveraging prior knowledge to address the challenges of sparse sampling image reconstruction [4]. In this study, we introduce an INR-based algorithm capable of reconstructing CT images from sparse projection data while simultaneously providing the corresponding MR image. As a patient-specific approach, the resulting INR model embeds prior patient information (e.g., the planning CT or MR scan) into the weights of a multi-layer perceptron (MLP) neural network. This design eliminates the need for extensive training or large-scale datasets. With only a few newly acquired projections, the MLP weights can be adjusted to accurately represent the new CT anatomy and its corresponding MR counterpart.

2 Materials and Methods

Figure 1 illustrates the proposed framework. We employ a generalizable implicit neural representation (GINR) using two parallel multi-layer perceptron (MLP) branches that map input spatial coordinates to their respective CT/MR values. The first MLP, represented in blue, embeds CT images by mapping spatial coordinates to Hounsfield Unit (HU) values. The second MLP (in orange) learns the paired MR image by mapping spatial coordinates to MR intensities. The initial layers' weights (in red) are shared between both networks, capturing instance-specific variations, while the remaining weights are consistent across patient samples and encapsulate modality-specific characteristics. When new projection measurements are recorded, we exclusively adjust the shared weights using the CT branch. Subsequently, these adjusted weights are utilized to obtain the corresponding MR image from the secondary branch.

Problem formulation. We frame the reconstruction problem as starting from a forward process $y = Ax + e$, where $x \in \mathbb{R}^M$ is the CT image of the unknown subject with M voxels, $y \in \mathbb{R}^{N_p \times D}$ are a finite number N_p of sampled projections, A represents the forward model (i.e., the Radon transform), and e is acquisition noise. The corresponding inverse process aims at recovering x from sparse y measurements, and is generally formulated as an optimization problem with regularization as

$$\min_{x} E(Ax, y) + \rho(x), \quad (1)$$

with an error term $E(Ax, y)$ such as the L2 norm, and a regularization term $\rho(x)$ characterizing generic image prior information, which can be defined in multiple ways.

Implicit neural representation. Following previous work [4], we use a MLP $M_{\theta,\phi}$ with parameters θ and ϕ to represent images CT images, mapping spatial coordinates $c \in [0,1)^3$ to the corresponding HU values. Likewise, we define a secondary MLP $M_{\theta,\eta}$ mapping coordinates to a secondary modality such as MR intensity. The parameters θ from the first layers are shared across both MLPs and capture sample specific variations, while the parameters ϕ and η embed modality-specific information.

Prior embedding. For a given subject, we can embed a small set of prior CT images x_p and a paired prior MR images or dose distributions z_p of the same patient within the weights of the MLPs. Examples of these datasets can be the images recorded during the first few days of treatment delivery, or even a single planning image with simulated deformations. Embedding this prior data into a GINR consists of obtaining a set of optimum weights $\{\phi^*, \eta^*, \ \{\theta_n\}_{n=1}^N\}$ via an iterative optimization problem:

a) For each $\{x_p^n, z_p^n\}_{n=1}^N$, keeping ϕ^* and η^* fixed:

$$\min_{\theta_n} \frac{1}{M} \sum_{i=1}^{M} \left\| M_{\theta_n,\phi^*}(c_i) - x_{p,i}^n \right\|_2^2 + \left\| M_{\theta_n,\eta^*}(c_i) - z_{p,i}^n \right\|_2^2. \quad (2)$$

b) For all $\{x_p^n, z_p^n, \theta_n\}_{n=1}^N$, freezing θ_n:

$$\phi^*, \eta^* = \underset{\phi,\eta}{\operatorname{argmin}} \frac{1}{M} \sum_{i=1}^{M} \left(\left\| M_{\theta_n,\phi}(c_i) - x_{p,i}^n \right\|_2^2 + \left\| M_{\theta_n,\eta}(c_i) - z_{p,i}^n \right\|_2^2 \right). \quad (3)$$

where Equation 2 (inner loop) finds the optimum geometry-specific weights θ_n for each image pair, and Equation 3 (outer loop) finds the best ϕ and η weights over all available prior images. This process is repeated over T iterations.

Image reconstruction. Starting from the trained weights (ϕ^* and one of the θ_n), and once new projections are obtained, we can effectively formulate the search in image space as an optimization over the θ weight space as

$$\min_{\theta_r} \left\| A M_{\theta_r,\phi^*} - y \right\|_2^2, \quad (4)$$

aiming at obtaining the shared weights θ_r of the new geometry. Since the Radon transform A is differentiable, the gradients can be backpropagated to obtain the new INR minimizing Equation 4. Note that we disregard the explicit regularization term $\rho(x)$ in Equation 1, since all prior information is contained in the starting network's weights.

Technical details. In our implementation, we construct 2 MLPs with 9 layers, 256 neurons per layer, and a periodic activation function, where the weights of the first layer are shared for both image modalities. Based on previous works [5], we use a Fourier feature mapping of size 256 that encodes the input coordinates before the first layer and helps the network learn high-frequency functions. The MLPs are implemented in Pytorch [6] and trained for $T = 1000$ iterations using the Adam optimizer [7], with a starting learning rate of $1e^{-4}$ that is halved each time the loss plateaus for 20 epochs.

3 Experiments and Results

To demonstrate the capabilities of the proposed GINR method, we conducted experiments to reconstruct brain CTs and their corresponding MR images. We compare the performance of reconstructing CTs (Equation 4) from 10, 20 or 30 projections, using the peak signal-to-noise ratio (PSNR) as a metric representing reconstruction quality for MR and CT images. As a baseline, we report the results of performing CT image reconstruction using the filtered back-projection (FBP) algorithm.

Table 1: CT/MR reconstruction accuracy. *Peak signal-to-noise ratio (PSNR) between model outputs and the original image pairs, for both the proposed GINR model and the filtered back-projection (FBP) baseline. All values are in decibels (dB).*

Image	Methods	Pat 1	Pat 2	Pat 3
10 projections				
CT	FBP	14.78	13.71	14.77
	GINR	39.57	39.22	41.60
MR	GINR	26.27	26.20	27.45
20 projections				
CT	FBP	15.52	14.63	15.47
	GINR	39.81	41.24	42.87
MR	GINR	26.28	26.59	27.68
30 projections				
CT	FBP	15.73	14.75	15.77
	GINR	40.93	41.66	43.19
MR	GINR	26.48	26.70	27.73

Dataset. We obtained CT/MR pairs from brain patients undergoing radiation therapy from the publicly available dataset of the SynthRad challenge [8]. All pairs were rigidly aligned and anonymized via de-facing both images. We pre-process CT scans, cropping them and rescaling all HU values to the interval [0,1], using the maximum and minimum of each image. Likewise, after performing a bias field correction, all MR images are standardized by subtracting the mean and dividing by the standard deviation. All volumes are cropped to a cube of size $200 \times 200 \times 32$ and a voxel resolution of 1 mm.

Reconstruction accuracy. For each patient, we train the GINR using 10 artificial image pairs, obtained from applying random elastic deformations to the original images, with random displacements large enough to cause significant differences. The average PSNR for the embedded CT and MR images is 44.2 dB and 33.5 dB, respectively. After embedding these 10 prior pairs, and starting from a random prior representation (i.e., a random θ_n weight matrix), we reconstruct the original CT/MR from sparse measurements pair via Equation 4. Table 1 shows the PSNR values obtained by each method when reconstructing images with 10, 20 and 30 projections. With a higher PSNR for CT scan reconstruction, GINR significantly outperforms the FBP baseline. Although also being able to reconstruct MR images from the shared weights as seen in Figure 2, GINR cannot match the CT reconstruction accuracy for its MR reconstruction task, producing images that are in between the ground truth and the prior.

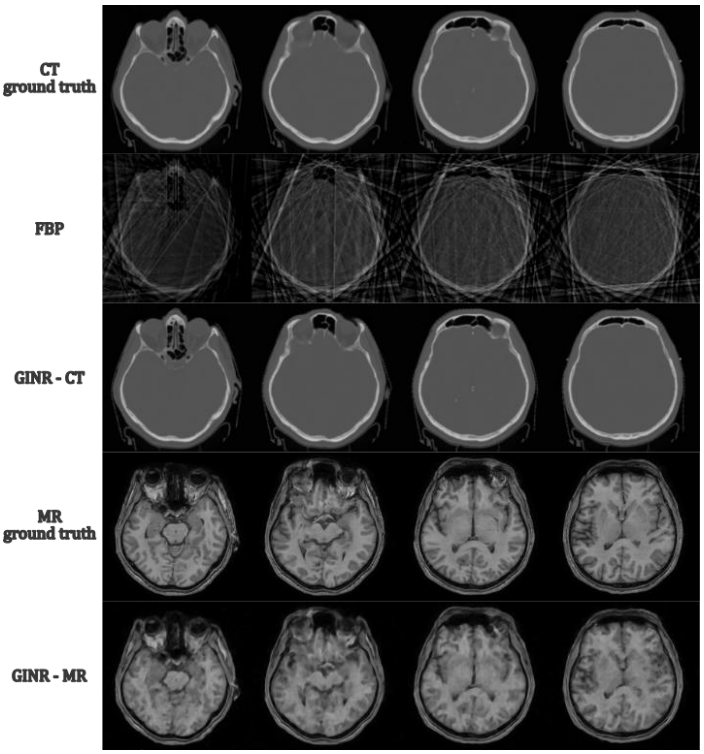


Figure 2: Qualitative results, using 10 projections. *From top to bottom: the CT ground truth image for patient 1, the CT reconstructed via filtered back-projection (FBP), the CT reconstructed by the proposed GINR, the MR ground truth image, and the MR reconstruction from GINR. All rows show 4 2D slices obtained from the 3D volumes.*

4 Discussion

The challenge of acquiring paired CT/MR images has remained a focal point within the realm of radiation therapy. Recent advancements in deep learning have predominantly focused on training models to establish mappings between various image modalities, relying on extensive datasets and substantial computational resources. In contrast, the proposed approach is patient-specific and obviates both problems. GINR maps input coordinates to image intensities and can be trained with minimal data samples in just a few hours with a single graphics processing unit.

The patient-specific nature of our methodology also confers distinct advantages in terms of reconstruction performance. Population-based deep learning methods, which must contend with the challenge of generalizing to unseen samples beyond the training distribution, especially when confronted with variations in machinery settings across diverse clinical environments. Instead, our proposed GINR method deliberately over-fits to prior information derived solely from a single subject, effectively capturing patient-specific intricacies often eluded by other models.

The expeditious image reconstruction capabilities of GINR render it well-suited for real-world clinical applications within the domain of radiotherapy, especially those demanding swift processing times. Notably, it proves instrumental in online adaptive workflows that necessitate the rapid acquisition of CT/MR pairs, concurrently facilitating contour delineation and physical dose distribution reconstruction.

While GINR demonstrates notable capabilities, there exists room for improvement. This is particularly evident in its MR outputs, which exhibit lower quality compared to its CT counterparts, and retain discernible traits from the prior images. This discrepancy is attributed to the exclusive reliance on CT projection data for the reconstruction of both images, leading to a dominance of shared weights that play a more significant role in CT output calculations. The resulting limitation in accurately capturing MR image characteristics underscores the need for refinements in GINR's methodology, with considerations for a more balanced incorporation of MR-specific information to enhance the fidelity of MR image reconstructions.

Looking ahead, future research endeavors could be directed towards extending the applicability of GINR to reconstruct paired images directly from MR sensor data. This extension holds significant promise, as it would enable the utilization of GINR in calculating doses within MR-guided treatments. By addressing the reconstruction challenge inherent in MR sensor data, GINR could play a pivotal role in advancing the capabilities of radiotherapy systems, offering enhanced precision and efficiency in treatment planning and delivery within the MR-guided paradigm.

5 Conclusion

In this work, we propose a framework for concurrent reconstruction of CT and MR images, based on a generalizable implicit neural representation (GINR) algorithm that embeds few prior CT/MR pairs within the weights of a neural network. By comparing the reconstructed images with the recorded data from a few brain cancer patients, we demonstrate that GINR can correctly reconstruct CT/MR scans even with sparsely sampled measurements, providing high-quality CT images. Compared to other deep learning approaches, the proposed GINR is patient-specific and does not require large datasets extensive training, or broad generalization capabilities across treatment sites and machines.

References

[1] Chernak, E. S., Rodriguez-Antunez, A., Jelden, G. L., Dhaliwal, R. S., & Lavik, P. S. (1975). The use of computed tomography for radiation therapy treatment planning. Radiology, 117(3 Pt 1), 613–614. https://doi.org/10.1148/117.3.613.

[2] Schmidt, M. A., & Payne, G. S. (2015). Radiotherapy planning using MRI. Physics in medicine and biology, 60(22), R323–R361. https://doi.org/10.1088/0031-9155/60/22/R323.

[3] Li, X., Bellotti, R., Meier, G., Bachtiary, B., Weber, D., Lomax, A., Buhmann, J., & Zhang, Y. (2023). Uncertainty-aware MR-based CT synthesis for robust proton therapy planning of brain tumor. Radiotherapy and oncology: journal of the European Society for Therapeutic Radiology and Oncology, 191, 110056. Advance online publication. https://doi.org/10.1016/j.radonc.2023.110056.

[4] Shen, L., Pauly, J., & Xing, L. (2022). NeRP: Implicit Neural Representation Learning with Prior Embedding for Sparsely Sampled Image Reconstruction. IEEE transactions on neural networks and learning systems, PP, 10.1109/TNNLS.2022.3177134. Advance online publication. https://doi.org/10.1109/TNNLS.2022.3177134.

[5] Tancik, M., Srinivasan, P., Mildenhall, B., Fridovich-Keil, S., Raghavan, N., Singhal, U., ... & Ng, R. (2020). Fourier features let networks learn high frequency functions in low dimensional domains. Advances in Neural Information Processing Systems, 33, 7537-7547.

[6] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Chintala, S, et. al. (2019). Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems, 32.

[7] Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

[8] Thummerer, A., van der Bijl, E., Galapon Jr, A., Verhoeff, J. J., Langendijk, J. A., Both, S., ... & Maspero, M. (2023). SynthRAD2023 Grand Challenge dataset: Generating synthetic CT for radiotherapy. Medical Physics.

FDDM: Unsupervised MR-to-CT Translation with a Frequency-Decoupled Diffusion Model

Yunxiang Li[1], Hua-Chieh Shao[1], Xiaoxue Qian[1], and You Zhang[1]

[1]Department of Radiation Oncology, University of Texas Southwestern Medical Center, Dallas, TX, USA

Abstract Diffusion models have shown great promise in unsupervised medical image translation applications. However, diffusion models are often challenged by inaccurate anatomical structure translations especially for unpaired training data, while anatomical structures are crucial for reliable disease diagnosis and treatment planning. We introduced a Frequency-Decoupled Diffusion Model (FDDM) to achieve unsupervised, accurate MR-to-CT translations for unpaired data. FDDM separates the frequency components of medical images in the Fourier domain to enable structure-preserving image conversion. Specifically, it combines an unsupervised frequency conversion module with a diffusion model that is initialized and conditioned on the converted frequency information. We evaluated FDDM on a brain MR-to-CT dataset, comparing it with GAN-, VAE-, and diffusion-based models. FDDM excelled in all metrics including Fréchet inception distance (FID), peak signal-to-noise ratio (PSNR), and structural similarity index measure (SSIM). The substantial improvement offered by FDDM demonstrated its ability to generate realistic CT images without compromising the underlying anatomical accuracy.

1 Introduction

Magnetic Resonance (MR) and Computed Tomography (CT) imaging, pivotal in medical diagnosis and treatment planning, offer complementary information: MR provides superior soft-tissue contrast while CT shows detailed bony anatomy. Specifically for radiation therapy, MR is often used for tumor localization, while CT is used for treatment dose calculations with its accurate electron density information. In current clinical settings, MR and CT are acquired separately in different machines, resulting in increased imaging time, radiation dose (from CT), and healthcare costs. Moreover, this sequential process leads to potential mis-registrations which impact anatomical alignment and localization. To address these challenges, researchers are exploring deep learning-based Generative Adversarial Networks (GANs) or Variational Autoencoders (VAEs) to convert MR images to synthetic CT scans. Notable developments in GAN-based methods include CycleGAN, which is based on cyclic consistency; and GcGAN, which is built on geometrically consistent mappings. Despite some success, GANs often suffer from low-quality MR-to-CT translations (low realism), due to the stark differences in anatomical presentations between MR and CT. In addition to GANs, VAE-based approaches such as UNIT [1] and MUNIT [2] share similar problems of low-quality translations.

Recently, diffusion models have emerged as powerful tools for generating high-quality and realistic images, surpassing GANs and VAEs. An example of the diffusion model is SDEdit [3] , which translates images through iterative denoising via stochastic differential equations. Another diffusion model, EGSDE [4], guides unpaired image translations with pre-trained energy features. In medical image translation, SynDiff [5] represents a significant advancement with its adversarial diffusion design, although its dependence on a separate CycleGAN model for paired data synthesis has limited its performance. In general, diffusion models work by a forward diffusion process and a reverse diffusion process. Forward diffusion adds Gaussian noise via a predefined scheme to gradually corrupt an image, while the diffusion model applies a reverse diffusion process to de-noise and recover the high-quality image by learning the underlying data distribution. In medical image translation, the forward diffusion process tends to eliminate critical structural details that are difficult to recover in the reverse diffusion. This issue is further compounded in unpaired medical datasets, where a one-to-one anatomical correspondence does not exist between source and target domains. However, unpaired datasets are far more prevalent and accessible than paired datasets. Thus, it is of practical importance to develop an unsupervised diffusion model on unpaired training datasets to both achieve high-quality image translations (as compared to GANs and VAEs) and preserve the anatomical details (maintaining the faithfulness of the translated anatomical structures).

To address this unmet need, we proposed a novel diffusion model-based approach for unsupervised MR-to-CT conversion, named Frequency-Decoupled Diffusion Model (FDDM). FDDM utilizes the decoupled frequency information to guide the anatomical structure-preserving image conversion. Specifically, FDDM is composed of an unsupervised frequency conversion module and a diffusion model initialized and conditioned on the converted frequency information. The unsupervised frequency conversion module, which can be GAN- or VAE-based, performs an initial MR-to-CT conversion. Based on the pros/cons of the GAN- or VAE-based methods, such conversion can produce CT images with high anatomical translation accuracy (faithfulness), but suffers from low image quality (realism). By FDDM, the frequency conversion module was designed to output two channels, one with high-frequency CT information to extract the faithfully-translated anatomical structure boundaries, and the other with full-frequency CT information for further diffusion model-based fine-tuning. In detail, the high-frequency CT output, containing the anatomical information, can be used to condition the subsequent diffusion model for structure-preserving MR-to-CT translation. The full-frequency CT, on the other

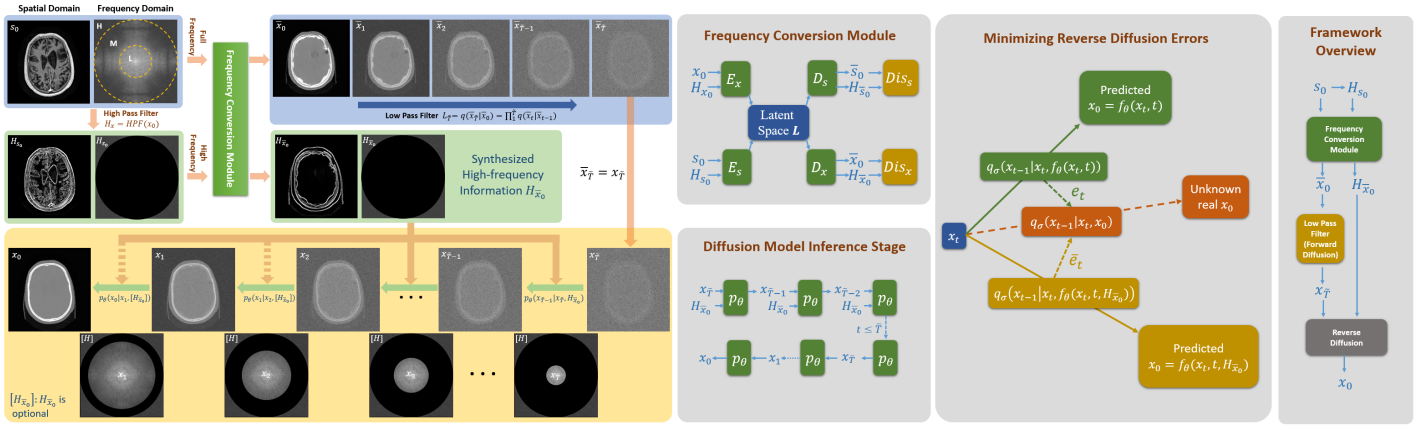


Figure 1: FDDM first uses an unsupervised frequency conversion module to generate two outputs: a high-frequency CT image and a full-frequency CT image. The full-frequency CT image undergoes a forward diffusion process that acts as a low-pass filter to extract the low-frequency information, which initiates a reverse diffusion process conditioned by the high-frequency CT information from the frequency conversion module, to produce the final synthetic CT image.

hand, is fed into a forward diffusion process (acting as a low-pass filter) to capture the overall CT intensities and semantic contents. The output is then used by the diffusion model as a starting step for the reverse diffusion process to gradually recover the missing frequency information of the CT image (especially that at the middle-frequency ranges), conditioned by the aforementioned high-frequency CT information. Compared with the original full-frequency CT generated by the unsupervised frequency conversion module, the new CT generated by the conditioned diffusion model sees boosted realism and image quality, while retaining accurate anatomical structures. In summary, our key contributions are:

1. We constructed a frequency-decoupled framework that utilizes high-frequency information to achieve anatomy-preserved image translations.

2. We proposed a flexible high-frequency conditioning and low-frequency initialization scheme in the diffusion model (based on a reverse diffusion error minimization scheme) to reduce/minimize the errors from the frequency conversion module to achieve more accurate MR-to-CT conversion.

3. FDDM demonstrated superior performance in MR-to-CT translations compared to existing state-of-the-art methods, including those based on GANs, VAEs, and other diffusion models.

2 Methods

Given a source MR image s_0 and a target CT image x_0, we used a VAE-based unsupervised frequency conversion module to convert the high-frequency and full-frequency information of the MR image into its counterpart in the CT image. The synthesized full-frequency CT images are then fed into a forward diffusion process (as a low-pass filter), and the output low-frequency information is used to initialize the reverse diffusion process for further fine-tuning. During reverse diffusion, the synthesized high-frequency CT information serves as a condition/guide to generate the final translated CT image. In the following sections, we first introduced the diffusion model, frequency filters, and the frequency conversion module. Following those, we introduced an unique and flexible high-frequency conditioning and low-frequency initialization scheme employed in FDDM to minimize reverse diffusion process errors, shown in Figure 1.

2.1 Diffusion Model

Diffusion models, as discussed in [6], use a forward diffusion procedure $q(\cdot)$, which progressively introduces white Gaussian noise with a constant variance I to the initial image x_0 through T steps:

$$q(x_{1:T}|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1}) \tag{1}$$

$$q(x_t|x_{t-1}) = N\left(x_t; \sqrt{\frac{\alpha_t}{\alpha_{t-1}}}x_{t-1}, (1-\frac{\alpha_t}{\alpha_{t-1}})I\right) \tag{2}$$

x_t denotes the noise-corrupted image at step t, and N is a normal distribution. The sequence $\alpha_{1:t}$ decreases monotonically and stays within $(0,1]^T$. With θ denoting model parameters, the reverse diffusion process, $p_\theta(\cdot)$, is then described as:

$$p_\theta(x_{0:T}) = p(x_T)\prod_{t=1}^{T} p_\theta(x_{t-1}|x_t) \tag{3}$$

Based on the study of Ho et al. [7], the forward diffusion process can be viewed as a Markov chain with Gaussian transitions, regulated by the descending sequence $\alpha_{1:t}$. Derived from Eq. 2, a specific attribute of the forward process can be expressed as:

$$q(x_t|x_0) = N(x_t; \sqrt{\alpha_t}x_0, (1-\alpha_t)I) \tag{4}$$

2.2 Frequency Decoupling and Conversion

On the basis of our previous work [8], it is known that when noise is introduced to an image, the power spectral density of the noise gets combined with that of the image. Since α_t decreases monotonically with diffusion steps, when the diffusion step t increases, the signal-to-noise ratio (SNR) decreases. Given an SNR threshold ϕ (below which the signal is viewed as totally corrupted), beyond a specific step t, only low-frequency information remains:

$$SNR_t = \frac{\sqrt{\alpha_t}k}{\sqrt{1-\alpha_t}\|(u,v)\|^a I}, \quad \begin{cases} < \phi, & \|(u,v)\| > \psi \\ \geq \phi, & \|(u,v)\| \leq \psi \end{cases} \quad (5)$$

where u and v represent the horizontal and vertical components of spatial frequencies. For detailed derivation of Eq. 5, please see our previous work [8]. If we denote this step as $\tilde{T}$, then the forward diffusion mirrors the effect of a low-pass filter:

$$x_{\tilde{T}} = LPF_{\tilde{T}}(x_0) = q(x_{\tilde{T}}|x_0) \quad (6)$$

In addition to the low-pass filtering naturally offered by the forward diffusion, FDDM employs a high-pass filter, the Sobel operator, to extract high-frequency information, which is based on the calculation of the image intensity gradient.

$$H_{x_0} = HPF(x_0) = Sobel(x_0) \quad (7)$$

In our previous work on CBCT-to-CT translation [8], we can directly extract the high-frequency information from CBCT via the Sobel filter to condition the diffusion model; and can directly use the forward-diffused (low-pass filtered) CBCT to initiate the reverse diffusion process for CT generation, since CBCT and CT share similar high-frequency and low-frequency information, and the diffusion model mainly works to correct the middle frequency differences. However, since MR and CT differ substantially in all frequency ranges, aligning MR's frequency traits with those of CT's requires an additional frequency conversion before applying our diffusion model. For FDDM, the frequency conversion module, based on a VAE framework with cycle consistency and a shared latent space L, is illustrated in Figure 1.

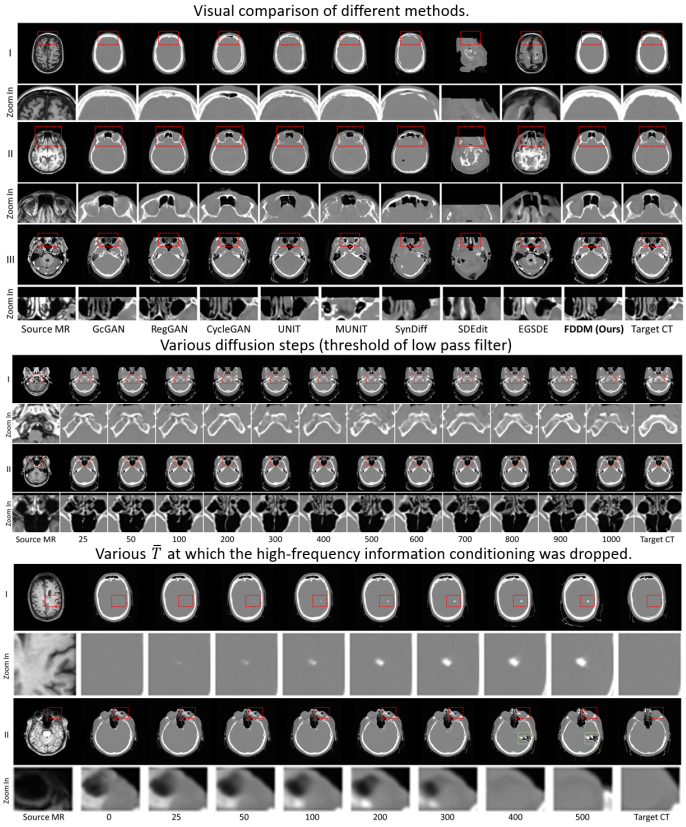


Figure 2: Visual results on the brain MR-to-CT translations, based on different methods and testing conditions

Comparison of different methods on the brain MR-to-CT translation dataset			
Method	FID ↓	SSIM ↑	PSNR ↑
GCGAN	100.1453	0.8012 ± 0.0584	35.5610 ± 1.5398
RegGAN	71.8493	0.7748 ± 0.0987	35.0037 ± 1.2602
CycleGAN	65.4173	0.7779 ± 0.0544	35.1246 ± 1.0129
UNIT	67.9931	0.7905 ± 0.0579	35.3229 ± 1.5297
MUNIT	82.5412	0.7516 ± 0.0678	34.9922 ± 1.2491
SynDiff	93.2486	0.8082 ± 0.0580	36.0907 ± 2.1146
SDEdit	96.7234	0.7302 ± 0.0739	34.8021 ± 2.2519
EGSDE	150.5618	0.6965 ± 0.0675	33.7938 ± 2.2106
FDDM	**29.8828**	**0.8255 ± 0.0544**	**36.4282 ± 1.7404**
Performance of using various diffusion steps during the test time (threshold of low pass filter)			
Step $\tilde{T}$	FID ↓	SSIM ↑	PSNR ↑
25	41.7128	0.8085 ± 0.0521	36.2313 ± 1.6768
50	38.6834	0.8129 ± 0.0522	36.3102 ± 1.6825
100	35.0101	0.8256 ± 0.0516	36.4354 ± 1.7440
200	31.4107	**0.8357 ± 0.0516**	**36.5114 ± 1.8025**
300	30.5433	0.8335 ± 0.0519	36.5057 ± 1.8183
400	30.0098	0.8307 ± 0.0518	36.4807 ± 1.7941
500	**29.8828**	0.8255 ± 0.0544	36.4282 ± 1.7404
600	30.2683	0.8215 ± 0.0572	36.3740 ± 1.7201
700	30.5590	0.8131 ± 0.0629	36.2645 ± 1.7021
800	31.7188	0.8020 ± 0.0727	36.1182 ± 1.7067
900	33.9832	0.7844 ± 0.0911	35.9113 ± 1.7694
1000	38.9938	0.7461 ± 0.1313	35.4433 ± 2.0954
Performance of stopping adding high-frequency information at different steps			
Step $\bar{T}$	FID ↓	SSIM ↑	PSNR ↑
0	53.0742	0.8004 ± 0.0556	36.1084 ± 1.6453
25	42.0695	0.8062 ± 0.0566	36.2401 ± 1.6389
50	37.5870	0.8101 ± 0.0564	36.2981 ± 1.6498
100	33.2123	0.8190 ± 0.0562	36.3811 ± 1.7195
200	29.8828	**0.8255 ± 0.0544**	**36.4282 ± 1.7404**
300	28.6831	0.8234 ± 0.0549	36.3753 ± 1.7666
400	**27.9087**	0.8170 ± 0.0558	36.2473 ± 1.7613
500	27.9878	0.8044 ± 0.0585	36.0336 ± 1.7653
Ablation study of FDDM			
Method	FID ↓	SSIM ↑	PSNR ↑
FDDM (w/o $H_{\bar{x}_0}$)	**27.9878**	0.8044 ± 0.0585	36.0336 ± 1.7653
FDDM (w/o Diffusion)	56.8436	0.7873 ± 0.0525	35.9617 ± 1.7167
FDDM	29.8828	**0.8255 ± 0.0544**	**36.4282 ± 1.7404**

Table 1: Comparison of different methods, parameters, and ablation experiments on brain MR-CT datasets. The arrows are pointing to the direction of improved accuracy.

2.3 Minimizing Reverse Diffusion Errors by Flexible High-Frequency Conditioning and Low-Frequency Initialization

After obtaining the full-frequency information $\bar{x}_0$ and high-frequency information $H_{\bar{x}_0}$ of the synthesized CT through the frequency conversion module, the low-frequency CT image $\bar{x}_{\tilde{T}}$ is further derived through the forward diffusion (as a low-pass filter) of $\bar{x}_0$. In reverse diffusion, x_{t-1} is obtained through a specific reverse conditional distribution $q_\sigma(x_{t-1}|x_t, x_0)$, defined in [6]. To derive x_{t-1}, we first need to predict x_0 by our model $f_\theta(x_t, t, [H_{\bar{x}_0}])$, where $[H_{\bar{x}_0}]$ indicates that the high-frequency condition $H_{\bar{x}_0}$ is optional. We define e_t as the error of $q_\sigma(x_{t-1}|x_t, f_\theta(x_t, t))$, where x_0 is predicted without the high-frequency condition; and we define $\bar{e}_t$ the error of $q_\sigma(x_{t-1}|x_t, f_\theta(x_t, t, H_{\bar{x}_0}))$, where x_0 is predicted with the high-frequency condition . If there is an optimal step $\overline{T}$ in the reverse process, beyond which $\bar{e}_t > e_t$, dropping the high-frequency condition can further minimize the error in the generated CT image. The condition-dropping process is shown in Figure 1.

In addition to the high-frequency condition dropping mechanism, the low-frequency CT information, which was low-pass filtered by the forward diffusion process from the converted full-frequency CT, can also be customized by using an optimized number of forward diffusion steps during testing.

The optimized step number determines the threshold of the low-pass filter, and the optimization of which will allow us to remove unnecessary/incorrect information from the converted full-frequency CT image while maintaining sufficient information to initialize the conditioned reverse diffusion process.

3 Results

Dataset: In our study, we employed the SynthRAD2023 dataset[1], which includes 31,142 MR and 31,142 CT slices for training and 3,417 MR and 3,417 CT slices for testing. To simulate unpaired (unsupervised) training, we re-arranged the slices and removed the pairing information from the training dataset. All slices were uniformly resized to a resolution of 256×256 pixels.

Comparison with other methods: We compared FDDM with eight other image translation methods, based on metrics including FID, SSIM, and PSNR scores. GcGAN, CycleGAN, and RegGAN (GAN-based), UNIT and MUNIT (VAE-based), and SynDiff, SDEdit, EGSDE (diffusion-based) are included in the comparison. From Table 1, GAN and VAE methods generated images with consistently worse FID scores, indicating challenges in creating realistic images. These methods however performed comparably to FDDM in SSIM, showing good anatomical preservation ability in image translations. SynDiff excelled in SSIM by using CycleGAN for paired MR-CT data synthesis, but training on CycleGAN-generated data limited its FID score. Both SDEdit and EGSDE used low-frequency MR information to synthesize the CT images, while the low-frequency information of MR and CT varies greatly, which in turn affects the translation accuracy and results in overall poor results. FDDM outperformed all other methods with the lowest FID and the highest SSIM and PSNR, demonstrating its strength in balancing image realism and structural fidelity.

Parameter Study: For FDDM, the total steps of the low-pass forward diffusion process can be adjusted. As shown in Figure 2, using too many forward diffusion steps can lead to significant information loss which cannot be fully recovered by the high-frequency condition, and result in worse FID, SSIM, and PSNR scores. In contrast, using too few diffusion steps may fail to remove incorrect information from the preceding frequency conversion module and result in low realism (FID). Based on the tests in Table 1, we chose 500 steps to enhance FID while preserving relatively high SSIM/PSNR.

Regarding high-frequency conditioning, Table 1 showed that by applying the condition up until step 200 in the reverse diffusion process yields a good FID and SSIM/PSNR balance. Beyond this point (by using a smaller $\overline{T}$), continuing applying the condition can increase the structural errors due to the inaccuracy inherent in the converted high-frequency information and lead to worse SSIM, PSNR, and FID. In contrast, dropping the high-frequency condition too early ($\overline{T}$ = 500) will diminish its guidance to the diffusion model and result in lower SSIM and PSNR scores.

Ablation Study: After acquiring the synthesized high- and low-frequency information, the final CT image is generated by reverse diffusion. An ablation study in Table 1 assessed the importance of using the diffusion model and adding the frequency guidance. The SSIM and PSNR scores for FDDM without high-frequency information conditioning were lower than those of the conditioned FDDM, demonstrating the benefit of introducing high-frequency information to preserve anatomical details. On the other hand, the removal of the diffusion model resulted in scores consistently worse than FDDM, suggesting that the diffusion model not only improves the image quality (better FID), but also corrects the erroneous information generated in the frequency conversion module, which in turn improves the SSIM and PSNR.

4 Conclusion

In this study, we introduced FDDM, a novel frequency-decoupled diffusion model for MR-to-CT image translations. FDDM uniquely combines the structure preservation ability of VAE-based models with the realistic image generation capability of diffusion models, achieving high faithfulness and realism in unsupervised MR-to-CT translations. Its frequency decoupling mechanism allows different frequency information to be flexibly applied to initiate and condition the diffusion models and maintain the integrity of translated anatomies. Extensive evaluations showed FDDM's superiority over current GAN-, VAE-, and diffusion-based models, making it a valuable tool for accurate MR-to-CT translations and downstream clinical applications.

References

[1] M.-Y. Liu, T. Breuel, and J. Kautz. "Unsupervised image-to-image translation networks". *Advances in neural information processing systems* 30 (2017).

[2] X. Huang, M.-Y. Liu, S. Belongie, et al. "Multimodal unsupervised image-to-image translation". *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 172–189.

[3] C. Meng, Y. Song, J. Song, et al. "Sdedit: Image synthesis and editing with stochastic differential equations". *arXiv preprint arXiv:2108.01073* (2021).

[4] M. Zhao, F. Bao, C. Li, et al. "Egsde: Unpaired image-to-image translation via energy-guided stochastic differential equations". *arXiv preprint arXiv:2207.06635* (2022).

[5] M. Özbey, O. Dalmaz, S. U. Dar, et al. "Unsupervised medical image translation with adversarial diffusion models". *IEEE Transactions on Medical Imaging* (2023).

[6] J. Song, C. Meng, and S. Ermon. "Denoising Diffusion Implicit Models". *International Conference on Learning Representations*. 2021.

[7] J. Ho, A. Jain, and P. Abbeel. "Denoising diffusion probabilistic models". *Advances in Neural Information Processing Systems* 33 (2020), pp. 6840–6851.

[8] Y. Li, H.-C. Shao, X. Liang, et al. "Zero-shot Medical Image Translation via Frequency-Guided Diffusion Models". *IEEE transactions on medical imaging* (2023).

[1] https://synthrad2023.grand-challenge.org/

Enhancing CBCT Image Quality Using a Novel sCT-based Shading Correction

Francesco Maria Russo[1,2], Lukas Lamminger[1], Gabriele Belotti[1], Katia Parodi[2], Heinz Deutschmann[1], and Philipp Steininger[1]

[1]medPhoton GmbH, Salzburg, Austria
[2]Department of Medical Physics in Radiation Oncology, Ludwig-Maximilians-Universität München, Am Coulombwall 1, D-85748,Garching, Germany

Abstract This study presents an innovative method for correcting shading artefacts in cone-beam computed tomography (CBCT) by employing an Artificial Intelligence (AI) generated synthetic CT (sCT) as a reference template. Our approach enhances CBCT image quality, improving both visual clarity and achieving a notable reduction in spatial non-uniformity (SNU) and mean Hounsfield Unit (HU) error between the CBCT and its paired planning CT (pCT). For fat tissue, SNU decreased from an average of 360.76 HU to 129.09 HU, and for muscle tissue it decreased from 277.69 HU to 109.44 HU. Similarly, the mean HU error between the CBCT and its pCT for fat tissue was reduced on average from 122.76 HU to 42.28 HU, and for muscle tissue from 113.56 HU to 62.59 HU. This approach represents an advancement over previous techniques by eliminating the need for a previously available pCT and increasing the correction's speed from an average of 45 seconds of previous methods to 23 seconds. Most importantly, our method demonstrates its robustness by effectively enhancing CBCT image quality even when the sCT is of very low quality, characterized by significant hallucinations and artefacts. This feature is crucial as it allows for the effective use of sCT to correct shading artefacts in CBCT while ensuring that any existing artefacts in the sCT are not transferred to the corrected image, possibly introducing a more reliable pathway for integrating sCTs into clinical practice.

1 Introduction

Cone-beam computed tomography (CBCT) is crucial for accurate patient positioning in radiation therapy and image guided surgery [1, 2]. However, the quality of CBCT images often falls short compared to fan-beam CTs, primarily due to significant scatter artefacts [3]. Additionally, other artefacts such as beam hardening and motion artefacts further challenge its image quality [4]. Among the numerous methods to address scatter radiation in CBCT, analytical modelling remains the dominant approach in commercial CBCT systems. Yet, even with the latest analytical scatter correction techniques, CBCT images occasionally still exhibit substantial shading artefacts, adversely impacting diagnostic and therapeutic accuracy. Recent advancements in deep learning have shown promise in effectively addressing CBCT artefacts via the generation of synthetic CT (sCT), notably improving image quality and Hounsfield Unit (HU) value accuracy [5]. However, networks trained on specific datasets, may not perform predictably with input images that deviate from their training conditions, such as those with unique patient anatomy or different imaging settings. Complicating this unpredictability in AI-generated sCTs is the lack of a widely recognized and established quality control (QC) tools for the sCT verification, posing a hurdle to sCTs direct and unrestricted clinical use [6]. Recent developments in shading correction methods by L. Shi et al. have shown the potential to further improve the shading artefacts correction when applied in tandem with traditional analytical scatter correction techniques [7, 8]. This study introduces a novel approach that combines the strengths of a sCT reference images, which include eliminating the need for an available pCT and omitting the registration step, with the improved shading correction techniques developed by L. Shi et al. By doing so, we aim at increasing the speed and versatility of the traditional method and facilitate the safer integration of AI-generated images into clinical practice.

2 Materials and Methods

2.1 Network implementation and training

In order to generate the sCTs for this study, we implemented a 3D UNet in PyTorch based on the library implemented by Pérez-García F. [9] consisting of 4 encoding blocks, linear upsampling and a PReLU activation function. The data used for training the network comes mainly from the SynthRAD2023 Grand Challenge [10] consisting of 360 CBCT-CT pairs of H&N and pelvis anatomies coming from different vendors. Additional 104 image pairs, covering H&N, pelvis and spine anatomical regions, come from paired CBCT-CT images acquired using medPhoton's mobile ImagingRing [11]. The data has been divided into train, validation and test using the proportion 80/10/10. During training, data augmentation was implemented via the TorchIO [12] framework, and the images were preprocessed using the Z-normalization. Batch normalization was also used to better stabilize the training. For the loss function of the network, we used a combination of the structural similarity index (SSIM) and perceptual loss that have been highlighted as yielding better results than standard MAE or MSE loss functions [5]. The network used for the perceptual loss is implemented in the paper by S. Chen et al. [13].

2.2 Shading correction implementation

The proposed method aims to correct shading artefacts in CBCT images by employing an sCT as a template. Initially, a binary mask f is created from the CBCT data using specific HU thresholds (-600 HU and 300 HU were used in our case). This mask isolates the fat and muscle tissues while excluding air and bone, thereby focusing the correction on relevant tissue areas in this study. Within this mask, the

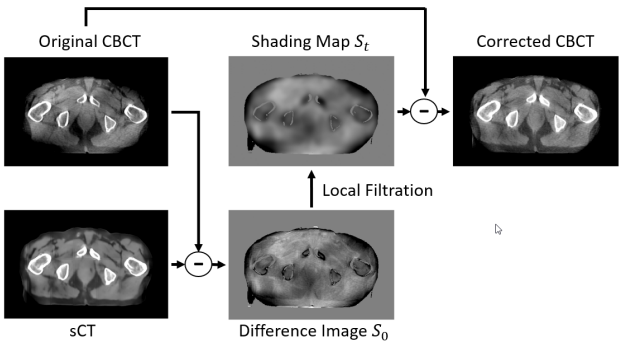


Figure 1: Shading Correction workflow using a sCT image as a template.

sCT is subtracted from the CBCT, resulting in an intermediate image that we call S_0 in continuity with [8]. S_0 is then processed yielding the final shading map S_t using a local filtration algorithm [7, 14] implemented as:

$$S_t = \frac{(S_0 \cdot f) * w}{f * w} \tag{1}$$

where w is a 3D Gaussian smoothing kernel with standard deviation σ. In our study, we selected a kernel with $\sigma = 5$ *mm*. This shading map is then subtracted from the original CBCT, yielding a corrected CBCT image free from the identified shading artefacts. A schematic diagram of the proposed workflow can be seen in Fig.1. The entire shading correction algorithm was implemented using the Python wrapping of the Insight Segmentation and Registration Toolkit (ITK)[15] on a Dell Precision laptop equipped with an Intel i7-12800H 2.40 GHz CPU and an NVIDIA RTX A2000 8GB GPU running Windows 10.

2.3 Method evaluation

In this preliminary evaluation, our method was applied to CBCTs of five different patients. Two of the CBCT images were acquired using an Elekta XVI CBCT system (Patient 1 and Patient 2) coming from the SynthRAD2023 Grand Challenge dataset [10], while the other three were acquired using medPhoton's ImagingRing (Patient 3, Patient 4 and Patient 5). Paired rigidly registered pCT images were available for two of the CBCT images from the ImagingRing system and for both CBCT images from the Elekta system. Both imaging systems incorporate an embedded proprietary analytical scatter correction algorithm to correct scatter artefacts. The two CBCT images obtained from the SynthRAD2023 Grand Challenge dataset originally had a voxel spacing of $(1 \times 1 \times 2.5)$ *mm*. To accommodate the requirements of our 3D U-Net architecture, specifically the z-dimension of the input patches, these images were resampled to a uniform spacing of $(1 \times 1 \times 1)$ *mm*. This resampling was essential because the original z-spacing of 2.5 *mm* resulted in fewer pixels in the z-direction than necessary for the U-Net input patches' specifications. For evaluating the performance of the method, as in [8], we manually selected for each image five regions of interest (ROI) with a size of 15-by-15 pixels for the two investigated tissue types (e.g. fat and muscle). For each tissue type, we then computed the Spatial Non-Uniformity (SNU) defined as:

$$SNU = \bar{\mu}_{max} - \bar{\mu}_{min} \tag{2}$$

where $\bar{\mu}_i$ represents the average CT number computed inside a single ROI, and we indicated with $\bar{\mu}_{max}$ ($\bar{\mu}_{min}$) the maximum (minimum) $\bar{\mu}_i$ among the ROIs. In the cases where the pCT was available, it was used to further evaluate the model's performance. This was done by calculating, for each tissue type, the mean CT value error $\langle\Delta\rangle$ between the CBCT and the pCT ROIs defined as:

$$\langle\Delta\rangle = \frac{1}{N}\sum_{i=1}^{N} |\Delta_i| \tag{3}$$

where $\Delta_i = \bar{\mu}_{CBCT,i} - \bar{\mu}_{pCT,i}$ represents the difference of the mean CT number between the CBCT and its paired pCT in the same ROI. Additionally, we investigated the impact of varying the σ parameter of the Gaussian filter. This parameter acts as a proxy for the level of trust in the generated sCT. A lower σ, indicating high trust in the sCT, allows the method to correct high-frequency artefacts. Conversely, when the sCT is considered less reliable or of lower quality, increasing the σ can still enhance CBCT image quality by correcting predominantly low-frequency shading artefacts.

3 Results

3.1 Paired image evaluation

Fig.3 shows the CBCTs from Patient 1 to Patient 4 with and without the applied shading correction. First of all, it is shown visually how the proposed method successfully suppresses the shading artefacts in all the CBCT images and improves the overall image uniformity. Table 1 presents the computed SNU for the five selected ROIs alongside the mean CT errors. Additionally, Fig.3 illustrates that some hallucination artefacts in the sCT are absent in the corrected CBCT, which retains the original CBCT's correct anatomical information. For our dataset, the processing times for the correction varied depending on the image size. For the largest image, with dimensions of $(500 \times 482 \times 266)$ voxels, the time required for our sCT generation was 25 seconds, and the subsequent shading correction took 6 seconds. Conversely, for the smallest image in our dataset with dimension $(396 \times 256 \times 260)$ voxels, the sCT generation and shading correction were completed in 13 and 2 seconds, respectively.

3.2 Effect of the sigma

We investigated the impact of varying the Gaussian kernel's standard deviation on the CBCT correction, particularly in challenging cases where the sCT generated by the network is

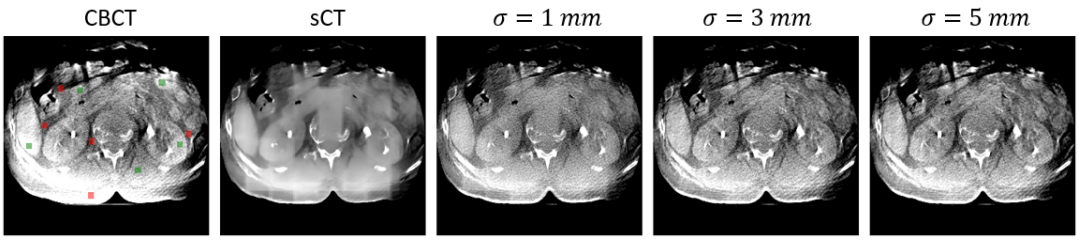


Figure 2: Shading Correction applied on Patient 5 when using different σ for the Gaussian filter of the local filtration. The red and green squares on the original CBCT represent the five selected ROIs for fat and muscle tissue respectively. Notice how the sCT displays square artifacts at the lower portion and inadequate clear separation of fat and muscle tissues. Display window: [-250 250] HU.

SNU (HU)	**On fat**			**On muscle**		
	pCT	**Uncorrected CBCT**	**Corrected CBCT**	**pCT**	**Uncorrected CBCT**	**Corrected CBCT**
Patient 1	30.77	352.37	70.43	26.63	293.34	58.87
Patient 2	23.87	350.40	75.91	17.82	249.49	51.82
Patient 3	27.81	420.23	101.65	23.19	280.83	146.28
Patient 4	22.39	342.02	157.69	35.29	395.47	190.09
Patient 5	-	338.78	239.75	-	169.31	100.16
Average	26.21	360.76	129.09	25.73	277.69	109.44
$\langle\Delta\rangle$ (HU)	**On fat**			**On muscle**		
		Uncorrected CBCT	**Corrected CBCT**		**Uncorrected CBCT**	**Corrected CBCT**
Patient 1		102.11	36.10		115.07	55.47
Patient 2		97.71	24.27		93.36	45.85
Patient 3		166.94	34.50		142.30	59.37
Patient 4		124.26	74.26		103.53	89.69
Patient 5		-	-		-	-
Average		122.76	42.28		113.56	62.59

Table 1: SNU and mean CT number error $\langle\Delta\rangle$ for the different patients for the two analyzed tissue types: fat and muscle.

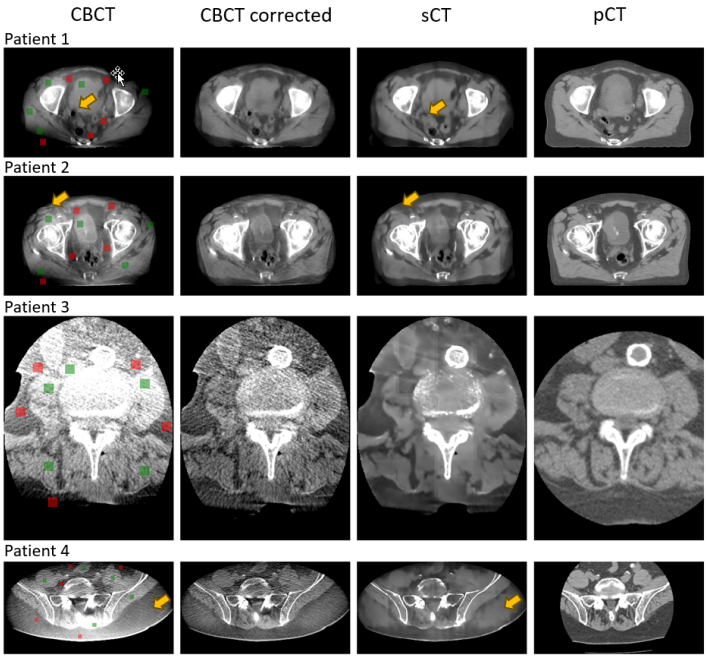


Figure 3: Shading correction results on the four patients where the pCT was available. The red and green squares on the original CBCTs represent the five selected ROIs for fat and muscle tissue respectively. The yellow arrows highlight anatomy discrepancies between the CBCT and the sCT. Display window: [-250 250] HU.

of extremely low quality and clinically unsafe due to severe hallucinations. Our findings, illustrated in Fig.2, reveal that our correction method effectively improves CBCT uniformity across a range of σ values despite the low quality of the sCT, therefore allowing for the safe use of the sCT in scenarios where its quality would typically render it unusable.

4 Discussion

Our proposed method for correcting shading artefacts in CBCT images, utilizing an sCT as a template, has shown success in improving CBCT image quality. Specifically, we observed a significant decrease in the SNU; for fat tissue, SNU reduced on average from 360.76 HU to 129.09 HU, and for muscle tissue, from 277.69 HU to 109.44 HU. Additionally, the accuracy of HU values improved, with the average CT number error for fat tissue dropping from 122.76 HU to 42.28 HU, and for muscle tissue from 113.56 HU to 62.59 HU. Patients 1 and Patient 2 exhibited smaller SNU than Patient 3, Patient 4, and Patient 5, possibly reflecting the imbalance in our training dataset, which predominantly features images from the SynthRAD2023 Grand Challenge. The smallest improvement in terms of SNU is achieved on Patient 5, where the correction was made using a low-quality sCT. Despite this, our method still improved the CBCT visualization by correcting the shading artefacts, and correctly mitigating the severe artefacts present in the sCT, showing

how our implementation could be useful for a safer implementation of sCTs in clinical scenarios. Using the sCT as a template for shading correction in CBCT presents advantages over traditional methods in terms of speed and in terms of complete independence of the availability of a pCT that would also need to be registered to the CBCT. Indeed, our processing time, achieved without an intensive focus on performance optimization, still renders our method faster overall than the previously implemented techniques by decreasing the correction time from 45 s to 23 s on average. Despite these results, the effectiveness of our method in terms of HU accuracy depends on the fidelity of the sCT's HU values. The strategy of blurring the difference between the CBCT and the sCT adds robustness against sCT artefacts, but further refinements in sCT generation could improve the accuracy of our corrections. Currently, our method results mainly in adaptive filtration at the CBCT border due to the absence of sparse sampling of the difference between the CBCT and the sCT. Future implementations will incorporate error estimation from the sCT using Monte Carlo (MC) dropout. This will allow us to identify and exclude voxels with high prediction uncertainty and use local filtration to create a more accurate shading map for a more precise correction.

5 Conclusion

This study presents a novel sCT-based shading correction method for CBCT, offering improvements over existing techniques both in terms of speed and versatility. The main impact of our method is promoting the accuracy of segmentation and navigation for image guided radiotherapy or surgery using in room CBCT. Particularly noteworthy is the method's ability to enhance CBCT image quality safely, even when the sCT is inaccurately rendered or contains severe artefacts, paving the way for a safer integration of sCT in clinical practice. Future research directions include enhancing the accuracy of the sCT HU values to further reduce the mean CT number error of the shading correction, and integrating uncertainty prediction in the network. This advancement will allow for the exclusion of voxels with higher uncertainty in the correction process, ensuring even more precise and reliable results in term of HU values accuracy. Such improvements would allow us to assess the feasibility of the method for dose recalculation.

6 Acknowledgement

This research was funded by Marie Skłodowska-Curie grant RAPTOR: Real-Time Adaptive Particle Therapy of Cancer, grant agreement No. 955956.

References

[1] D. Létourneau, J. W. Wong, M. Oldham, et al. "Cone-beam-CT guided radiation therapy: technical implementation". *Radiotherapy and Oncology* 75.3 (2005), pp. 279–286. DOI: https://doi.org/10.1016/j.radonc.2005.03.001.

[2] J. Tonetti, M. Boudissa, G. Kerschbaumer, et al. "Role of 3D intraoperative imaging in orthopedic and trauma surgery". *Orthopaedics Traumatology: Surgery Research* 106.1, Supplement (2020). 2019 Instructional Course Lectures (SoFCOT), S19–S25. DOI: https://doi.org/10.1016/j.otsr.2019.05.021.

[3] J. H. Siewerdsen and D. A. Jaffray. "Cone-beam computed tomography with a flat-panel imager: Magnitude and effects of x-ray scatter". *Medical Physics* 28.2 (2001), pp. 220–231. DOI: https://doi.org/10.1118/1.1339879.

[4] R. Schulze, U. Heil, D. Gro, et al. "Artefacts in CBCT: a review". *Dentomaxillofacial Radiology* 40.5 (2011). PMID: 21697151, pp. 265–273. DOI: 10.1259/dmfr/30642039.

[5] B. Rusanov, G. M. Hassan, M. Reynolds, et al. "Deep learning methods for enhancing cone-beam CT image quality toward adaptive radiation therapy: A systematic review". *Medical Physics* 49.9 (2022), pp. 6019–6054. DOI: https://doi.org/10.1002/mp.15840.

[6] C. Seller Oria, A. Thummerer, J. Free, et al. "Range probing as a quality control tool for CBCT-based synthetic CTs: In vivo application for head and neck cancer patients". *Medical Physics* 48.8 (2021), pp. 4498–4505. DOI: https://doi.org/10.1002/mp.15020.

[7] L. Shi, T. Tsui, J. Wei, et al. "Fast shading correction for cone beam CT in radiation therapy via sparse sampling on planning CT". *Medical physics* 44.5 (2017), pp. 1796–1808. DOI: https://doi.org/10.1002/mp.12190.

[8] L. Shi, A. Wang, J. Wei, et al. "Fast shading correction for cone-beam CT via partitioned tissue classification". *Physics in Medicine & Biology* 64.6 (2019), p. 065015. DOI: 10.1088/1361-6560/ab0475.

[9] F. Pérez-García. *fepegar/unet: PyTorch implementation of 2D and 3D U-Net*. Version v0.7.5. Mar. 2020. DOI: 10.5281/zenodo.3697931.

[10] A. Thummerer, E. van der Bijl, A. Galapon Jr, et al. "SynthRAD2023 Grand Challenge dataset: Generating synthetic CT for radiotherapy". *Medical Physics* 50.7 (2023), pp. 4664–4674. DOI: https://doi.org/10.1002/mp.16529.

[11] C. Wei, J. Albrecht, S. Rit, et al. "Reduction of cone-beam CT artifacts in a robotic CBCT device using saddle trajectories with integrated infrared tracking". *Medical Physics* (2024). DOI: https://doi.org/10.1002/mp.16943.

[12] F. Pérez-García, R. Sparks, and S. Ourselin. "TorchIO: a Python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning". *Computer Methods and Programs in Biomedicine* (2021), p. 106236. DOI: https://doi.org/10.1016/j.cmpb.2021.106236.

[13] S. Chen, K. Ma, and Y. Zheng. "Med3D: Transfer Learning for 3D Medical Image Analysis". *CoRR* abs/1904.00625 (2019). DOI: https://doi.org/10.48550/arXiv.1904.00625.

[14] L. Zhu. "Local filtration based scatter correction for cone-beam CT using primary modulation". *Medical Physics* 43.11 (2016), pp. 6199–6209. DOI: https://doi.org/10.1118/1.4965042.

[15] M. McCormick, X. Liu, L. Ibanez, et al. "ITK: enabling reproducible research and open science". *Frontiers in Neuroinformatics* 8 (2014). DOI: 10.3389/fninf.2014.00013.

Comparison of saddle and helical trajectories of a mobile cone beam computed tomography scanner with online calibration correction

Chengtao Wei[1,2], Simon Rit[3], Matthieu Laurendeau[3,4], Philipp Steininger[5], Felix Ginzinger[5], Marco Riboldi[2], Christopher Kurz[1], and Guillaume Landry[1]

[1]Department of Radiation Oncology, LMU University Hospital, LMU Munich, Munich, Germany
[2]Department of Medical Physics, Ludwig-Maximilians-Universität München, Garching, Germany
[3]Univ Lyon, INSA-Lyon, Université Claude Bernard Lyon 1, UJM-Saint Etienne, CNRS, Inserm, CREATIS UMR 5220, U1294, F-69373 Lyon, France
[4]Thales AVS, Moirans, France
[5]Research & Development, medPhoton GmbH, Salzburg, Austria

Abstract In cone-beam computed tomography (CBCT), the cone-beam (CB) artefact causes object-dependent image distortions away from the central plane. Non-circular trajectories such as saddles or helices can mitigate the CB artefact, but are challenging to execute with scanners not specifically designed for non-circular scanning. In this work, we made use of a recently available mobile robotic CBCT scanner (mobile ImagingRing, medPhoton, Salzburg, Austria) which allows sinusoidal motion perturbation of several degrees of freedoms during a source and detector rotation, coupled with infrared (IR) motion tracking to measure the geometry of each trajectory. We scanned a disk phantom designed to highlight CB artefacts with circular, saddle and helical trajectories executed by driving motion with wheels. We computed an incompleteness metric, $\tan(\psi)$, and correlated it with the region free from CB artefacts in the images. Using a separate cylindrical phantom, we additionally computed the modulation transfer function at 10 percent (MTF_{10}) to assess the accuracy of the calibration correction. We found regions free of CB artefacts of up to 190 mm in the superior-inferior direction for the saddle trajectory, and of 120 mm to 140 mm for the helical trajectory. MTF_{10} was 1.39 lp/mm, 0.98 lp/mm and 1.03 lp/mm for the circular, saddle and helical trajectories with IR tracking, while without IR tracking the saddle and helical trajectories had MTF_{10} less than 0.4 lp/mm.

1 Introduction

In cone-beam computed tomography (CBCT), non-circular scanning trajectories can eliminate the well known cone-beam (CB) artefact, which stems from data incompleteness when moving away from the plane containing the source trajectory. For points away from that plane, Tuy's condition is not satisfied since there will always be planes through these points which do not intersect the trajectory [1]. This issue can reduce image quality in a wide range of medical imaging applications, including dentistry, image-guided surgery and image-guided radiotherapy. The degree of incompleteness can be quantified with the $\tan(\psi)$ metric introduced by Clackdoyle et al. [2], which provides a continuous value which was shown to correlate with the severity of the CB artifact [3]. It is well known that using non-circular trajectories in CBCT scanning can mitigate the CB artifact, with publications demonstrating sinusoidal trajectories [4–6], circle-and-line or circle-and-arc trajectories [7, 8], or even freely optimized trajectories [9–11]. While several works have demonstrated improvements from such trajectories, often they have been performed at devices not specifically designed for such a purpose, entailing complex trajectory programming or long scanning times. Recently, a mobile robotic CBCT scanner (mobile ImagingRing, medPhoton GmbH, Salzburg, Austria) has become available with the option to execute sinusoidal perturbations along various axes of motion. This allows implementing time-efficient non-circular trajectories which can be executed with one rotation of the source and detector, such as saddle [12] and helical [13, 14] trajectories. The device is further equipped with infrared (IR) tracking technology which allows mitigating geometric inaccuracies in recorded trajectories, which can degrade spatial resolution. In this work we compared CB artefact reduction achievable with saddle and helical trajectories using scans of a disk phantom and $\tan(\psi)$ evaluation, in addition to modulation transfer function at 10 percent (MTF_{10}) analysis with a separate phantom.

2 Materials and Methods

2.1 Mobile robotic CBCT scanner and trajectories

All CBCT scans were acquired at the mobile robotic CBCT scanner described in Figure 1 and equipped with IR tracking based on a pair of cameras which monitor the position of a tracking object placed on the phantoms (Rigid Body Marker Base, MCP1145; NaturalPoint, Inc., Corvallis, USA). The scanner's motorized wheels allowed the execution of saddle and helical trajectories. For the saddle trajectory, a rotation with angle β around a vertical yaw rotation axis, which is displayed in Figure 1, was performed during source and detector rotation, expressed with rotation angle θ, according to

$$\beta(\theta) = A_{\text{saddle}} \sin(\omega_{\text{saddle}}\theta + \phi_{\text{saddle}}) \quad (1)$$

where A_{saddle}, ω_{saddle} and ϕ_{saddle} are the amplitude, the angular frequency and the phase shift of the yaw perturbation. For the helical trajectory a similar sinusoidal perturbation D displacing the scanner longitudinally in the superior-inferior (SI) direction followed

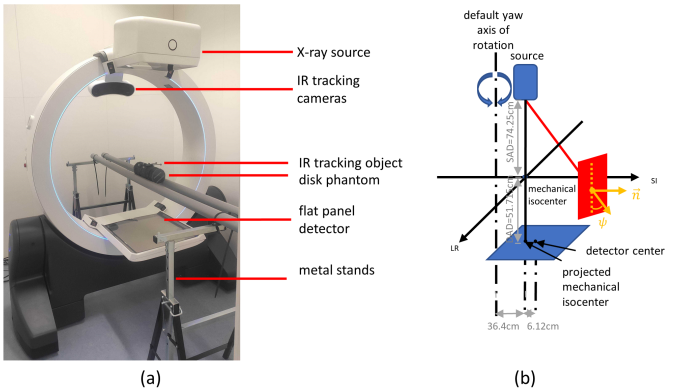


Figure 1: Setup and geometrical information of the CBCT scanner. (a) The whole setup for the experiments (note that the plastic support was replaced by a carbon fiber table). IR tracking markers are put on the phantom to be visible by both tracking cameras. (b) Basic geometrical information of the CBCT scanner and the rotation axis and source projection offset relative to the detector center. The geometry used to calculate $\tan(\psi)$ is shown in red and orange (see Ref [2] for details).

Table 1: Parameters of the saddle and helical trajectories defined by equations 1 and 2 with θ expressed in rad. A_{saddle} is expressed in degrees for clarity.

A_{saddle} [°]	ω_{saddle} [rad^{-1}]	ϕ_{saddle} [rad]	A_{helical} [cm]	ω_{helical} [rad^{-1}]	ϕ_{helical} [rad]
10	1	π	7.7	0.5	$\frac{\pi}{2}$

$$D(\theta) = A_{\text{helical}} \sin(\omega_{\text{helical}}\theta + \phi_{\text{helical}}) \quad (2)$$

where A_{helical}, ω_{helical} and ϕ_{helical} are the amplitude, the angular frequency and the phase shift of the longitudinal perturbation. The values used in this work are listed in Table 1.

The IR tracking system provided the position and direction of the rigid object for each projection, which were transformed into a rotation matrix R and translation t. The recorded source position s, detector position d and orthogonal detector direction vectors u and v (in the gantry's coordinate system) were transformed as follows to account for the perturbations executed following equations 1 and 2.

$$\begin{aligned} s' &= R_0 R^{-1}(s-t) + t_0, \\ d' &= R_0 R^{-1}(d-t) + t_0, \\ u' &= R_0 R^{-1} u, \\ v' &= R_0 R^{-1} v, \end{aligned} \quad (3)$$

where R_0 and t_0 are for the first frame and R and t for the current frame.

2.2 Phantoms, experiments and simulations

Two phantoms were used in this work. The first was a disk phantom consisting of eleven 1 mm thick acrylic disks of 100 mm diameter separated by 20 mm foam spacers, which was used to assess CB artefacts. The second was an acrylic cylinder of 152 mm diameter used to evaluate MTF_{10}. Both phantoms were scanned using circular, saddle and helical trajectories acquiring 720 projections with θ ranging from 0 to 2π, exposure of 20 ms and 5 mA and with IR tracking enabled. The detector was 432 mm $\times$ 432 mm with 1440 pixels $\times$ 1440 pixels. The geometrical information from equation 3 was additionally used to simulate a digitally reconstructed radiograph (DRR) for each projection of the disk phantom, using a digital voxelized version of the phantom with a voxel size of 0.5 mm^3. This was done using functions from the Reconstruction Toolkit (RTK) software package (version 2.3.0) [15].

2.3 Image reconstruction

All images were reconstructed using RTK. We used penalized least-squares minimization with the conjugate gradient algorithm with a regularization weight of $\gamma = 10$, a voxel size of 0.5 mm^3 and 50 iterations. We reconstructed all experimental and simulated acquisitions defined above.

2.4 Image analysis

For each experimental acquisition, the incompleteness metric $\tan(\psi)$ was calculated according to Clackdoyle et al. [2] for all positions in the reconstructed image. The value of $\tan(\psi)$ is 0 for complete data and has been shown to allow adequate reconstruction up to 0.02 [16], above which conspicuous CB artefacts can be expected. In more detail, for all the projection lines passing through a voxel, we calculated and kept those projection lines that fell on the detector, following the approach of Sun et al. [16]. Then, for one specific direction $\vec{n}$, we took the lowest $\tan(\psi)$ from all the source positions. Finally, we took the largest $\tan(\psi)$ value from all directions as the incompleteness value for this voxel. Here, ψ was the angle between the plane and the projection line. The red and orange parts of Figure 1(b) show the geometry of this calculation. For the saddle and helical trajectories, we quantified the extent of the region free from CB artefacts based on visual inspection and overlaid $\tan(\psi)$ isolines.

The MTF_{10} was calculated using an over-sampled edge obtained from the cylindrical phantom's outer diameter and was used to determine the performance of IR tracking. This was done slice-by-slice for an SI region of 100 mm and we calculated the mean and standard deviation of MTF_{10}. Comparison of simulations and experiments was used to determine whether potential image distortions stemmed from the trajectories themselves or from issues with IR tracking.

3 Results

Figure 2 shows the source trajectories performed in experiments. All the trajectories started at 0 mm in SI, the saddle

trajectory extended to −135 mm, and the helical trajectory extended to −154 mm in SI. The source executed a single rotation around the gantry for all the trajectories. Table 2 reports the MTF_{10} for the circular, saddle, and helical trajectories. The spatial resolution was the same for the circular trajectory with and without IR tracking. The spatial resolution of the saddle and helical trajectory both had important improvements when using IR tracking correction. The spatial resolution was improved from 0.37 mm to 0.98 mm for the saddle trajectory and from 0.34 mm to 1.03 mm for the helical trajectory.

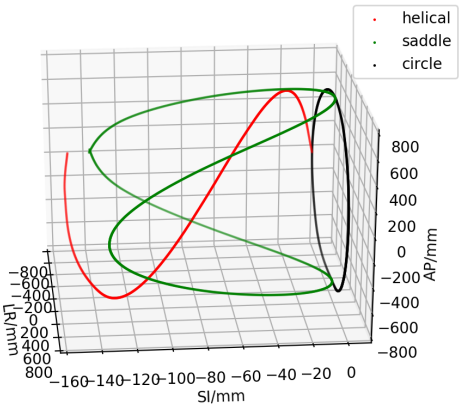


Figure 2: The three trajectories used in this work. The black line represents the circular trajectory, the green line is the yaw-based saddle trajectory, and the red line is the helical trajectory. Axes not to scale.

Table 2: MTF_{10} for different trajectories w/ and w/o IR tracking (mean ± standard deviation).

	with IR tracking	**w/o IR tracking**
	[lp/mm]	
circle	1.39 ± 0.06	1.39 ± 0.04
saddle	0.98 ± 0.07	0.37 ± 0.04
helical	1.03 ± 0.09	0.34 ± 0.01

The reconstructed images from the simulations of the three trajectories are shown in Figure 3. The disks inside the low incompleteness region ($\tan(\psi) < 0.02$) are all well reconstructed without CB artifacts and distorted into spindle shapes when lying outside of the low incompleteness region. For the saddle trajectory, the shape of the low incompleteness region was an irregular cylinder with 190 mm height in SI and a cross-section which had a short axis of 255 mm and a long axis of 270 mm. For the helical trajectory, the low incompleteness region had a round cross-section with diameter 255 mm, while the height in SI ranged from 120 mm to 140 mm as shown in Figure 3 (c).
Figure 4 shows the coronal plane of the reconstructed images from the experimental acquisitions with the three trajectories,

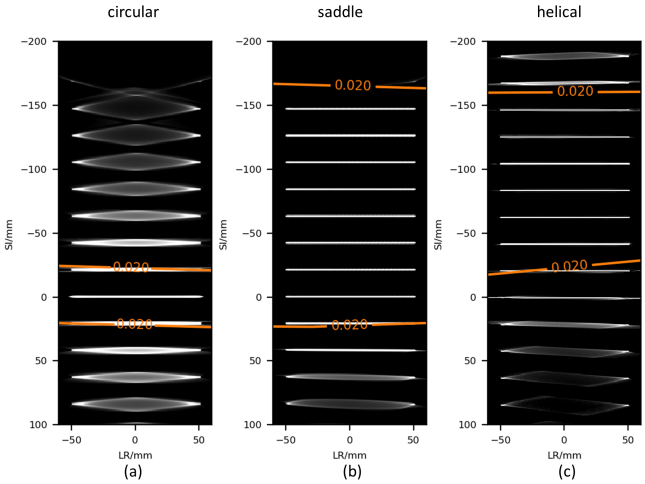


Figure 3: Reconstructed images of the digital disk phantom from the three simulated trajectories in a superior-inferior (IR) and left-right (LR) view. Orange lines indicate the edges of the low incompleteness region where $\tan(\psi) < 0.02$.

with convex hull and incompleteness overlays. The sizes and shapes of the low incompleteness regions are the same as the simulation, and whithin all disks are well reconstructed. We could confirm the observation from the simulation that the saddle trajectory provided a larger region where disks are well reconstructed. We can observe that the low incompleteness region is larger than the convex hull for all trajectories.

4 Discussion

Table 2 and Figure 4 show that the IR tracking correction is reliable and can provide accurate geometrical corrections online. The MTF_{10} of the saddle and the helical trajectories with IR tracking, although lower than for the circular scan, were comparable to that achievable at a linac-mounted CBCT scanner (XVI, Elekta, UK), for which we measured a MTF_{10} of 1.00 ± 0.01 lp/mm with a circular scan. Both the simulation (Figure 3) and the experiment (Figure 4) show that the helical trajectory provides a longer reconstructed region where more disks than for the saddle scan could be reconstructed, albeit in a distorted fashion. However, the low incompleteness region of the helical trajectory was smaller in SI than that of the saddle trajectory, despite the source having moved a larger distance along the SI direction for the helical trajectory. This was attributed to the non-closed nature of the trajectory, and helical trajectory simulations demonstrated that using $\omega_{\text{helical}} = 2\,\text{rad}^{-1}$ could provide a convex hull with an SI extent corresponding to $2A_{\text{helical}}$. We can also see in Figure 3 that at least one disk outside of the low incompleteness region is well reconstructed. These observations suggest that the SI length of the low incompleteness region is limited by a direction $\vec{n}$ other than that normal to the disks. One advantage of the helical trajectory over the saddle was that it was less likely to collide with the patient couch, which might occur when performing saddle trajectories with larger A_{saddle}.

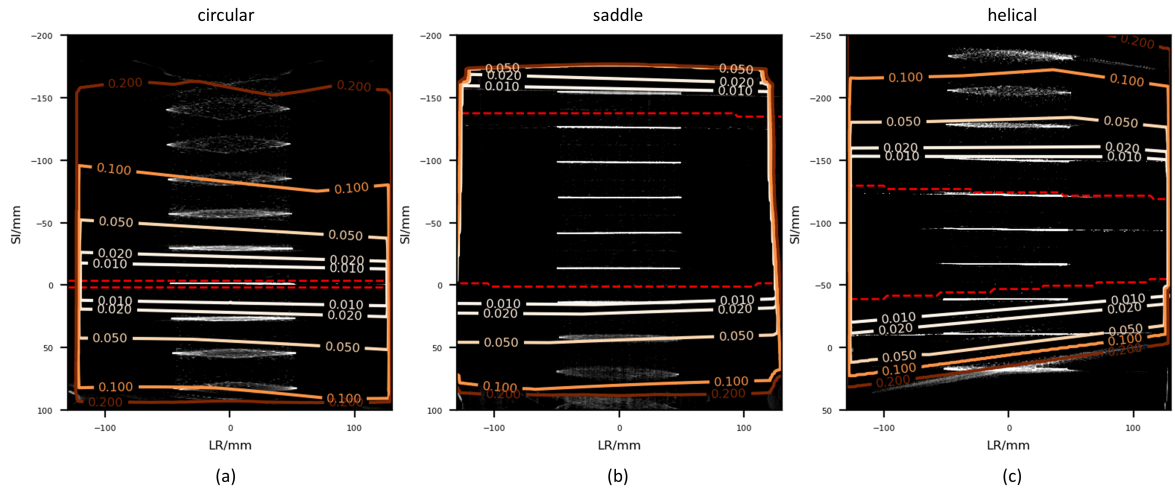


Figure 4: Reconstructed images of the disk phantom from experimental trajectories in a superior-inferior (IR) and left-right (LR) view, with convex hull and $\tan(\psi)$ isolines overlay. Red dashed lines are edges of the convex hull and the solid lines are $\tan(\psi)$ isolines.

5 Conclusion

In conclusion, both the yaw-based saddle and helical trajectories were feasible with online IR tracking geometrical corrections at the mobile robotic CBCT scanner used in this work. This allowed an important increase of the low incompleteness region ($\tan(\psi) < 0.2$) and elimination of the CB artifact.

References

[1] H. K. Tuy. "An Inversion Formula for Cone-Beam Reconstruction". *SIAM Journal on Applied Mathematics* 43.3 (1983), pp. 546–552.

[2] R. Clackdoyle and F. Noo. "Quantification of Tomographic Incompleteness in Cone-Beam Reconstruction". *IEEE Transactions on Radiation and Plasma Medical Sciences* 4.1 (Jan. 2020), pp. 63–80. DOI: 10.1109/TRPMS.2019.2918222.

[3] A. Tersol, P. Wu, R. Clackdoyle, et al. "Sampling effects for emerging cone-beam CT systems and scan trajectories: from Tuy's condition to system design and routine image quality tests". en. *Medical Imaging 2022: Physics of Medical Imaging*. Ed. by W. Zhao and L. Yu. San Diego, United States: SPIE, Apr. 2022, p. 55. DOI: 10.1117/12.2613139.

[4] Y. Q. Ma, G. J. Gang, T. Ehtiati, et al. "Non-Circular CBCT Orbit Design and Realization on a Clinical Robotic C-arm for Metal Artifact Reduction". *Medical Imaging 2022: Image-Guided Procedures, Robotic Interventions, and Modeling*. Ed. by C. A. Linte and J. H. Siewerdsen. San Diego, United States: SPIE, Apr. 2022, p. 8. DOI: 10.1117/12.2612448.

[5] H. Hosoo, Y. Ito, A. Marushima, et al. "Image Quality Improvements for Brain Soft Tissue in Neuro-Endovascular Treatments: A Novel Dual-Axis "Butterfly" Trajectory for Optimized Cone-Beam CT". *European Journal of Radiology* 160 (Mar. 2023), p. 110713. DOI: 10.1016/j.ejrad.2023.110713.

[6] T. Reynolds, Y. Ma, T. Wang, et al. "Revealing Pelvic Structures in the Presence of Metal Hip Prothesis via Non-Circular CBCT Orbits". *Proceedings of SPIE–the International Society for Optical Engineering* 12466 (Feb. 2023), 124660Y. DOI: 10.1117/12.2652980.

[7] A. Katsevich. "Image reconstruction for the circle and line trajectory". *Physics in Medicine and Biology* 49.22 (Nov. 2004), pp. 5059–5072. DOI: 10.1088/0031-9155/49/22/003.

[8] A. Katsevich. "Image reconstruction for the circle-and-arc trajectory". en. *Physics in Medicine and Biology* 50.10 (May 2005), pp. 2249–2265. DOI: 10.1088/0031-9155/50/10/005.

[9] G. J. Gang, J. H. Siewerdsen, and J. W. Stayman. "Non-circular CT orbit design for elimination of metal artifacts". en. *Medical Imaging 2020: Physics of Medical Imaging*. Ed. by H. Bosmans and G.-H. Chen. Houston, United States: SPIE, Mar. 2020, p. 79. DOI: 10.1117/12.2550203.

[10] J. W. Stayman, S. Capostagno, G. J. Gang, et al. "Task-driven source–detector trajectories in cone-beam computed tomography: I. Theory and methods". en. *Journal of Medical Imaging* 6.02 (May 2019), p. 1. DOI: 10.1117/1.JMI.6.2.025002.

[11] P. Wu, N. Sheth, A. Sisniega, et al. "C-arm orbits for metal artifact avoidance (MAA) in cone-beam CT". en. *Physics in Medicine and Biology* 65.16 (Aug. 2020), p. 165012. DOI: 10.1088/1361-6560/ab9454.

[12] J. D. Pack, F. Noo, and H. Kudo. "Investigation of saddle trajectories for cardiac CT imaging in cone-beam geometry". en. *Physics in Medicine and Biology* 49.11 (June 2004), pp. 2317–2336. DOI: 10.1088/0031-9155/49/11/014.

[13] F. Noo, J. Pack, and D. Heuscher. "Exact helical reconstruction using native cone-beam geometries". en. *Physics in Medicine and Biology* 48.23 (Dec. 2003), pp. 3787–3818. DOI: 10.1088/0031-9155/48/23/001.

[14] H. Yu and G. Wang. "Studies on artifacts of the Katsevich algorithm for spiral cone-beam CT". en. *Developments in X-Ray Tomography IV*. Ed. by U. Bonse. Denver, CO, Oct. 2004, p. 540. DOI: 10.1117/12.559286.

[15] S. Rit, M. V. Oliva, S. Brousmiche, et al. "The Reconstruction Toolkit (RTK), an open-source cone-beam CT reconstruction toolkit based on the Insight Toolkit (ITK)". *Journal of Physics: Conference Series* 489.1 (Mar. 2014), p. 012079. DOI: 10.1088/1742-6596/489/1/012079.

[16] T. Sun, R. Clackdoyle, J.-H. Kim, et al. "Estimation of Local Data-Insufficiency in Motion-Corrected Helical CT". en. *IEEE Transactions on Radiation and Plasma Medical Sciences* 1.4 (July 2017). rate: 2, pp. 346–357. DOI: 10.1109/TRPMS.2017.2710237.

Deep learning-based Normal Tissue Complication Models for Head and Neck Cancer patients

Lisanne V. van Dijk[1], Suzanne P.M. de Vette[1], Hendrike Neh[1], Luuk van der Hoek[1], Hung Chu[1], Johannes A. Langendijk[1], Peter M.A. van Ooijen[1], and Nanna M. Sijtsema[1]

[1]Department of Radiotherapy, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands

Abstract Accurate pre-treatment toxicity risk estimation in head and neck cancer (HNC) patients is essential to optimize radiotherapy and minimize long-term toxicities. This study utilizes the potential of deep learning (DL) to predict based on 3D data to improve the current conventional 1D Normal tissue complication probability (NTCP) models. Our results show that DL models that incorporating full 3D dose and image information improved the predictive performance for late xerostomia, dysphagia, and taste loss compared to conventional NTCP models. The attention maps showed that the DL model's focus is on the salivary gland dose for xerostomia, a broad scale of swallowing muscles for dysphagia, and parotid glands and anterior region of the tongue for taste loss. This research sets a new benchmark for toxicity prediction in current HNC radiotherapy, paving the way for more informed treatment decisions, minimizing toxicities, and enhancing patients' quality of life.

1 Introduction

Head and neck cancer (HNC) patients are often faced with severe toxicities during and after radiotherapy, the most prominent of which are xerostomia (dry mouth), swallowing, and taste dysfunction. These toxicities significantly impact the quality of life [1]. Combined with the increasing survival rates of HNC patients [2,3], this underscores the need to optimize treatment and minimize radiation-induced toxicities. The growth in recent radiotherapy advancements that aim to reduce toxicity while maintaining effective treatment (e.g., image-guided radiotherapy and proton therapy) has resulted in an increase in treatment options [4–7]. To select the most suitable treatment for individual patients, accurate pre-treatment toxicity risk estimation is required.

Normal tissue complication probability (NTCP) models predict radiation-induced toxicities based on discrete dose parameters of one or a couple organs (e.g., mean dose parotid gland) and clinical variables (e.g., baseline toxicity scores). NTCP models have been published for xerostomia [8–10], swallowing dysfunction [11–13], and recently Bosch et al. [14] developed the comprehensive toxicity risk prediction profile (CITOR) incorporating an array of swallowing, salivary, taste, speech, pain, and general toxicities[14]. These models have shown to be effectiveness for selecting HNC patients for proton therapy of in the Netherlands [15]. However, they lack the capacity to consider the 3D nature of radiation-induced damage. Incorporating complete dose distributions has the potential to significantly enhance their predictive performance.

Deep learning (DL) models, particularly convolutional neural networks (CNNs), directly extract features from 3D information, preserving spatial details in dose and image data without reducing them to 1D dose-volume or image parameters. Building upon Men et al.'s demonstration of a 3D CNN's potential in predicting late xerostomia in RTOG-0522 HNC patients[16], our research group developed DL models for more comprehensive HNC toxicities, utilizing a large-scale prospective toxicity dataset of patient treated with current radiation standards. Moreover, our models incorporate clinical variables, explore alternative DL architectures, employ attention maps to enhance interpretability, and assess them against the latest NTCP modeling standards.

The study aims to develop DL NTCP models for moderate-to-severe late xerostomia, swallowing dysfunction and taste loss in HNC radiotherapy patients, comparing them to the current state-of-the-art NTCP models.

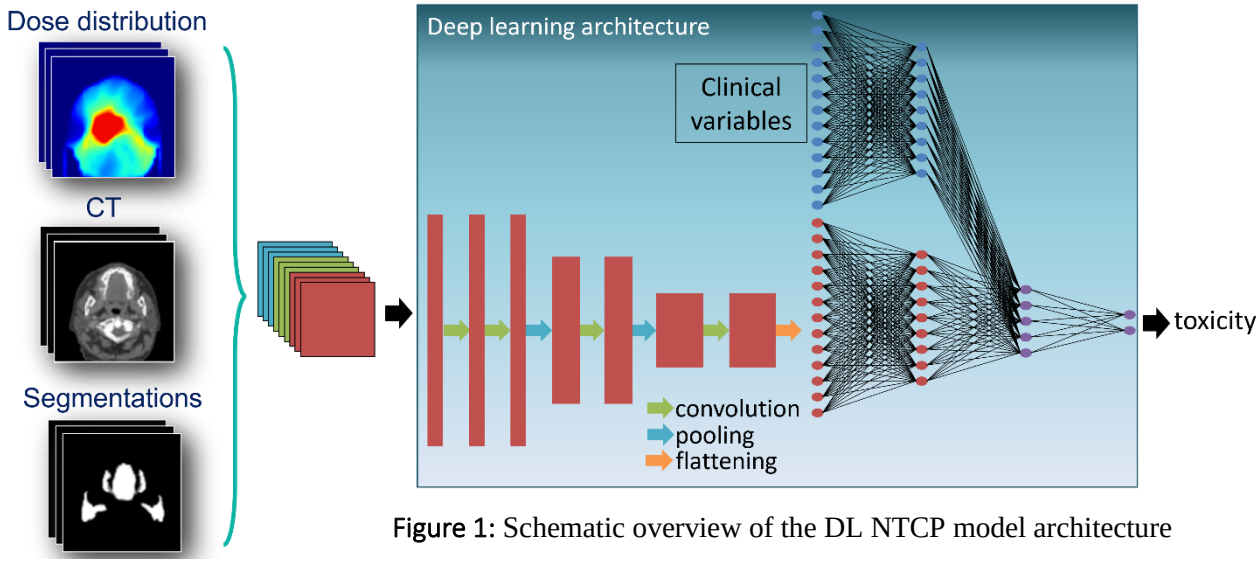


Figure 1: Schematic overview of the DL NTCP model architecture

2 Materials and Methods

2.1 Patient cohorts and treatment

The training and independent validation cohort included HNC patients treated with curative (chemo)radiotherapy for squamous cell carcinoma in the oropharynx, larynx, oral cavity, or nasopharynx at University Medical Centre Groningen from 2007-2021. Most patients received IMRT or VMAT (79.1), with a minority receiving IMPT (13.7%) or 3DCRT (7.2%). All patients had a planning CT approximately 2 weeks before treatment (> 80% contrast-enhanced).

2.2 NTCP endpoints

The NTCP model endpoints were moderate-to-severe:

1) patient-rated xerostomia 12 months after RT
2) patient-rated taste loss 6 months after RT
3) physician-rated dysphagia 6 months after RT

All toxicity scores were prospectively collected. Patient-rated endpoints were collected with the EORTC QLQ-HN35 questionnaire ('none,' 'any' vs. 'moderate,' 'severe'). Dysphagia was assessed using CTCAEv4 (grade ≥2).

2.3 Reference NTCP models

The CITOR NTCP models [14] for xerostomia, taste, and dysphagia (Table 1), developed/validated on 750/395 HNC patients, were fitted to the current datasets to serve as competitive reference for the DL NTCP models.

	Baseline Tox	oral cavity	parotid glands	Submandibular glands	PCM	sup/mid/inf	Buccal mucosa	Cricopharyng.	Smoking status	WHO score	T-stage	N-stage	Tumor location	Age	Sex	P16-status
Xerostomia	♦		♦	♦			⁕							⁕	⁕	
Taste		♦	♦	⁕										♦		
Dysphagia	♦	♦	⁕	⁕	♦			⁕	⁕	⁕	⁕	⁕	♦			⁕

Table 1: Predictors in NTCP models.
♦ = in reference and DL model, ⁕ = in DL model only

2.4 Deep learning (DL) Model development

The UMCG cohort was split into a training/cross-validation (70%/15%), and independent validation cohort (15%).

The DL model architecture consisted of convolutional layers encoding image information followed by fully connected layers to make the final prediction (Figure 1). The input into the convolutional layers was a three channel stack image – the *3D dose*, *CT*, *organ-at-risk* (*OAR) segmentations* (Table 1); while the clinical variables were fed into the fully connected layers. Model architectures were DCNN, ResNet and ResNet-DCNN. Hyperparameter tuning was performed per toxicity and was aided by the Optuna framework [17]. Cyclic learning rate was used for dysphagia and taste loss, while a cosine learning rate was used for xerostomia. The optimizer were MADGRAD for dysphagia, Adam for taste loss, and AdaBound for xerostomia. Internal validation was performed with 10-fold cross-validation.

The input data – CT, dose, segmentations – was registered to each other, normalized, resampled [2x2x2 mm^3] and cropped [96x96x96 voxels] based on anatomic landmarks. Data augmentation was applied to improve generalization: cropping, lateral flip, minor translation, zooming, rotation, and intensity shifting.

2.5 Model evaluation metrics

The performance of all NTCP models was assessed with the area under the curve (AUC) in the cross-validation sets and independent test set. Calibration was assessed in the test set.

3 Results

3.1 Patient cohort

A total of 1442 HNC patients were included (Table 2). Incidence rates for xerostomia, dysphagia, and taste loss were 35%, 25%, and 27%, respectively.

	Train set		Test set	
Total	**1226**		**216**	
Available scores				
Xerostomia	759	62%	138	64%
Taste	807	66%	142	66%
Dysphagia	941	77%	171	79%
*Clinical characteristics**				
Age (SD)	63.8	(10.1)	64.0	(10.3)
Site: OPC	354	38%	58	34%
T-stage >T2	488	52%	90	53%
N-stage=N+	497	53%	95	56%
HPV= pos	158	17%	26	15%

**for the largest set (dysphagia)*

Table 2: Patient numbers and characteristics

3.2 Model performance

The performance of the *xerostomia* DL model was good in the cross-validation (AUC_{train}=0.79), and was improved when compared to the reference model (AUC_{train}=0.75). When tested in the independent validation set, the performance of the DL model was consistently good (AUC_{test}=0.79), and again better than the performance of the reference NTCP model (AUC_{test}=0.74). The calibration plot of the predicted versus the observed xerostomia rates in the validation set demonstrated that the DL model gives a clearer line, yet with a slightly larger inclination than the identity line, where the reference model follows the identity line better, but the points are more scattered.

The performance of the *dysphagia* DL model was excellent in the cross-validation (AUC_{train}=0.85). The performance of the reference NTCP model was also high (AUC_{train}=0.79), yet still substantially lower than the DL model. This improvement remained present in the independent validation set: with an AUC_{test} of 0.84 for the DL model and AUC_{test} of 0.76 for the reference model. The calibration plot

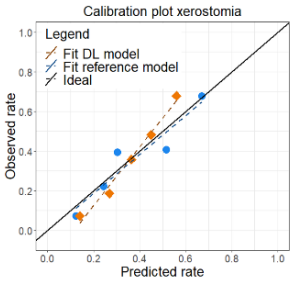
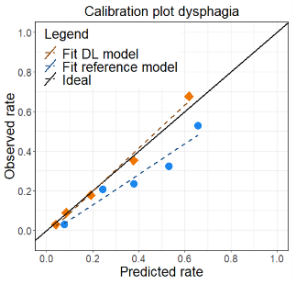
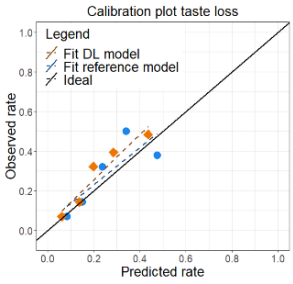
	Xerostomia		Dysphagia		Taste Loss	
	Reference	**DCNN**	**Reference**	**ResNet**	**Reference**	**ResNet-DCNN**
Cross-validation	0.75 [0.72 – 0.78]	**0.79** [0.76 – 0.82]	0.79 [0.76–0.82]	**0.85** [0.81–0.89]	0.70 [0.66–0.75]	**0.74** [0.70 - 0.77]
Independent validation	0.74 [0.66 – 0.83]	**0.79** [0.71 – 0.86]	0.76 [0.68–0.84]	**0.84** [0.77–0.90]	0.71 [0.70 - 0.72]	**0.72** [0.69 - 0.74]
Calibration plot on independent validation						

Table 3: Model performance of the reference model and deep learning (DL) NTCP models

showed a better fit to the observed dysphagia rates than the reference model (Table 3).

Both the DL model (AUC_{train}=0.74) and reference model (AUC_{train}=0.70) showed good performance for *taste loss*, yet was lower compared to the other two toxicities. While the performance of the DL model in the validation set (AUC_{test}=0.72) was still higher that of the reference (AUC_{test}=0.71), the improvement was less pronounced than for the other toxicities. Calibration was arguably marginally better for the DL than the reference model, as the calibration points show a more consistent incline.

3.3 Attention maps

The attention maps of the DL xerostomia NTCP model show that the main focus of the network was on the salivary glands (Figure 2a). Moreover, high focus was seen in the center of the parotid glands.

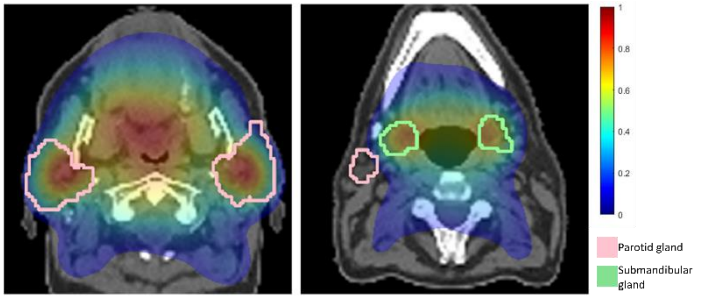


Figure 2a: Attention map for the DL xerostomia model

For the taste DL NTCP model, the attention was somewhat focused on the anterior site of the tongue and on the parotid glands (Figure 2b). The attention maps showed a more diffuse picture than for xerostomia and dysphagia.

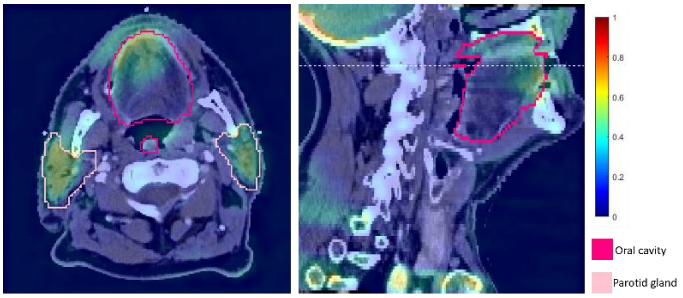


Figure 2b: Attention map for the DL taste NTCP model

The attention maps for the DL dysphagia NTCP model showed that the DL model focusses on the swallowing structures (Figure 2c). For instance, high focus was seen for the oral cavity and pharyngeal constrictor muscle (PCM).

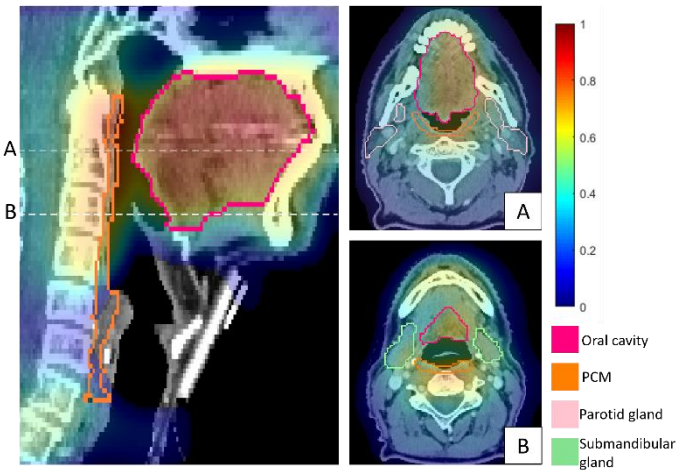


Figure 2c: Attention map for the DL dysphagia model. PCM = pharyngeal constrictor muscle.

4 Discussion

Our results suggest that using the full 3D dose and image information results in improved prediction performance over the conventional models that are based on average dose to distinct OARs. A large dataset with prospectively collected toxicity scores was used to develop and subsequently validate 3D DL models to predict late xerostomia, dysphagia and taste. The DL models performed well in both the training set and the independent test set.

The biggest improvement was seen for dysphagia, which is likely due to the many swallowing structures that are involved in maintaining swallowing functionality which makes conventional modelling approaches very complicated due to insuperable issues with multicollinearity between dose parameters of those structures. Also for xerostomia, the 3D DL model shows improved performance over the conventional NTCP model. The attention maps show that the model is still guided by the radiation dose administered to the salivary glands. Furthermore, higher

focus on the center of the parotid gland can be recognized, which corresponds to the stem cell-rich region of the gland [18]. The DL model improved the prediction of taste loss after treatment compared to the reference NTCP model, yet the increase in AUC was limited (Table 3). The prediction of taste loss is complex; the conventional NTCP model is based on parotid gland and oral cavity dose, yet the exact mechanisms causing taste loss are not fully understood.

The findings of this study are in line with the results of Men et al. [16], showing that DL outperforms logistic regression in predicting late xerostomia. In this study, we validated this in a more recent and larger cohort. Furthermore, DL models were developed and validated for other relevant HNC toxicities: dysphagia and taste. A fairer NTCP model comparison was ensured by using state-of-the-art conventional NTCP models as reference. Moreover, DL methodology was optimized by introducing a reproducible data pre-processing procedure, 10-fold stratified cross-validation, independent validation, and exploring attention maps. These improvements enhance understanding and applicability, setting a new benchmark for future clinical implementation. Nevertheless, more global external validation is needed to test the generalizability of the DL NTCP models.

5 Conclusion

Our study demonstrates superior prediction performance with DL NTCP models using full 3D dose and image information in predicting late xerostomia, dysphagia, and taste loss compared to conventional models based on average OAR dose. Clinical variables can be added to the DL models as well as the conventional models. Attention maps highlighted the relevance of salivary gland dose, as well as the central stem cell rich region for xerostomia. For dysphagia, the model focussed on a scala of swallowing muscles. The focus for taste was more diffuse, yet more attention was seen in the parotid glands and anterior part of the tongue. DL NTCP models show potential to improve the risk estimation of late HNC toxicities before treatment, facilitating better informed radiotherapy decisions (e.g., IMRT versus IMPT) to minimize toxicities and improve the quality of life of cancer survivors.

References

[1] Langendijk JA, Doornaert P, Verdonck-de Leeuw IM, et al. “Impact of Late Treatment-Related Toxicity on Quality of Life Among Patients With Head and Neck Cancer Treated With Radiotherapy.” Journal of Clinical Oncology 2008;26:3770–6. DOI:10.1200/JCO.2007.14.6647.

[2] Beadle BM, Liao K-P, Elting LS, et al. “Improved survival using intensity-modulated radiation therapy in head and neck cancers: A SEER-Medicare analysis.” Cancer 2014;120:702–10. DOI:10.1002/cncr.28372.

[3] Benson E, Li R, Eisele D, et al. “The clinical impact of HPV tumor status upon head and neck squamous cell carcinomas.” Oral Oncol 2014;50:565–74. DOI:10.1016/j.oraloncology.2013.09.008.

[4] Nutting C, Rooney K, Foran B, et al. “Results of a randomized phase III study of dysphagia-optimized intensity modulated radiotherapy (Do-IMRT) versus standard IMRT (S-IMRT) in head and neck cancer.” Journal of Clinical Oncology 2020;38:6508–6508. DOI:10.1200/JCO.2020.38.15_suppl.6508.

[5] Lagendijk JJW, Raaymakers BW, Raaijmakers AJE, et al. “MRI/linac integration.” Radiotherapy and Oncology 2008;86:25–9. DOI:10.1016/j.radonc.2007.10.034.

[6] Lomax A. “Intensity modulation methods for proton radiotherapy.” Phys Med Biol 1999;44:185–205.

[7] Castadot P, Lee JA, Geets X, et al. “Adaptive Radiotherapy of Head and Neck Cancer.” Semin Radiat Oncol 2010;20:84–93. DOI:10.1016/j.semradonc.2009.11.002.

[8] Beetz I, Schilstra C, Van Der Schaaf A, et al. “NTCP models for patient-rated xerostomia and sticky saliva after treatment with intensity modulated radiotherapy for head and neck cancer: The role of dosimetric and clinical factors.” Radiotherapy and Oncology 2012;105:101–6. DOI:10.1016/j.radonc.2012.03.004.

[9] Lee T-F, Liou M-H, Huang Y-J, et al. “LASSO NTCP predictors for the incidence of xerostomia in patients with head and neck squamous cell carcinoma and nasopharyngeal carcinoma.” Sci Rep 2014;4:6217. DOI:10.1038/srep06217.

[10] Moiseenko V, Wu J, Hovan A, et al. “Treatment planning constraints to avoid xerostomia in head-and-neck radiotherapy: An independent test of QUANTEC criteria using a prospectively collected dataset.” Int J Radiat Oncol Biol Phys 2012;82:1108–14. DOI:10.1016/j.ijrobp.2011.04.020.

[11] Mortensen HR, Jensen K, Aksglæde K, et al. “Late dysphagia after IMRT for head and neck cancer and correlation with dose-volume parameters.” Radiotherapy and Oncology 2013;107:288–94. DOI:10.1016/j.radonc.2013.06.001.

[12] Wopken K, Bijl HP, Van Der Schaaf A, et al. “Development of a multivariable normal tissue complication probability (NTCP) model for tube feeding dependence after curative radiotherapy/chemo-radiotherapy in head and neck cancer.” Radiotherapy and Oncology 2014;113:95–101. DOI:10.1016/j.radonc.2014.09.013.

[13] Christianen MEMC, Schilstra C, Beetz I, et al. “Predictive modelling for swallowing dysfunction after primary (chemo)radiation: results of a prospective observational study.” Radiother Oncol 2012;105:107–14. DOI:10.1016/j.radonc.2011.08.009.

[14] Van den Bosch L, van der Schaaf A, van der Laan HP, et al. “Comprehensive toxicity risk profiling in radiation therapy for head and neck cancer: A new concept for individually optimised treatment.” Radiotherapy and Oncology 2021;157:147–54. DOI:10.1016/j.radonc.2021.01.024.

[15] Langendijk JA, Lambin P, De Ruysscher D, et al. “Selection of patients for radiotherapy with protons aiming at reduction of side effects: the model-based approach.” Radiother Oncol 2013;107:267–73. DOI:10.1016/j.radonc.2013.05.007.

[16] Men K, Geng H, Zhong H, et al. “A Deep Learning Model for Predicting Xerostomia Due to Radiation Therapy for Head and Neck Squamous Cell Carcinoma in the RTOG 0522 Clinical Trial.” Int J Radiat Oncol Biol Phys 2019;105:440–7. DOI:10.1016/j.ijrobp.2019.06.009.

[17] Akiba T, Sano S, Yanase T, et al. “Optuna : A Next-generation Hyperparameter Optimization Framework” 2019:2623–31. DOI:10.1145/3292500.3330701.

[18] van Luijk P, Pringle S, Deasy JO, et al. “Sparing the region of the salivary gland containing stem cells preserves saliva production after radiotherapy for head and neck cancer.” Sci Transl Med 2015;7:305ra147. DOI:10.1126/scitranslmed.aac4441.

Causal Targeted Machine Learning for Real-World Evidence Generation in Oncology: Assessing Causal Effects of Radiation Toxicity on Overall Survival in Stage III Non-Small Cell Lung Cancer

Rachael V. Phillips[1],
Tomi Nano[2], Jessica Scholey[2], Efstathios D. Gennatas[3], Nasir Mohammed[4], Jing Zeng[5], Rupesh D. Kotesha[6], Pranshu Mohindra[7], Lane R. Rosen[8], John Chang[9], Henry K. Tsai[10], James J. Urbanic[11], Carlos E. Vargas[12], Nathan Y. Yu[12], Lyle H. Ungar[13], Eric Eaton[13], Charles B. Simone, II[14,*], Mark J. van der Laan[1,*], and Gilmer Valdes[2,3,*]

[*]Co-senior authorship
[1]Division of Biostatistics, School of Public Health, University of California, Berkeley, United States
[2]Department of Radiation Oncology, University of California, San Francisco, United States
[3]Department of Epidemiology and Biostatistics, University of California, San Francisco, United States
[4]University of Washington and Seattle Cancer Care Alliance Proton Therapy Center, Seattle, Washington, United States
[5]Northwestern Medicine Chicago Proton Center, Warrenville, Illinois, United States
[6]Miami Cancer Institute, Department of Radiation Oncology, Miami, Florida, United States
[7]University of Maryland School of Medicine and Maryland Proton Treatment Center, Baltimore, Maryland, United States
[8]Willis-Knighton Medical Center, Shreveport, Louisiana, United States
[9]Oklahoma Proton Center, Oklahoma City, Oklahoma, United States
[10]New Jersey Procure Proton Therapy Center, Somerset, New Jersey, United States
[11]California Protons Therapy Center, Department of Radiation Oncology, San Diego, California, United States
[12]Mayo Clinic Proton Center, Department of Radiation Oncology, Phoenix, Arizona, United States
[13]Department of Computer and Information Science, University of Pennsylvania, Philadelphia, United States
[14]Department of Radiation Oncology, Memorial Sloan Kettering Cancer Center, New York, New York, United States

Abstract This study explores the causal impact of radiation-induced toxicities on overall survival (OS) in stage III Non-Small Cell Lung Cancer (NSCLC) patients undergoing proton radiation therapy. Leveraging real-world data (RWD) from the Registry Study for Radiation Therapy Outcomes, and advanced causal inference methodologies that have garnered support from the US Food and Drug Administration's Real-World Evidence (RWE) Program, we estimated of the impact of pneumonitis/dyspnea and esophagitis/dysphagia on OS with targeted minimum loss-based estimation (TMLE). Our findings suggest there may be a time-dependent relationship between pneumonitis/dyspnea and decreased OS, emphasizing the need for personalized treatment strategies that balance treatment intensity with toxicity management. Additionally, our methodology demonstrates the utility of targeted learning and its accompanying step-by-step causal roadmap for generating reliable RWE, setting a precedent for applying such techniques in radiation oncology to support clinical and regulatory decision-making, and ultimately enhance patient care.

1 Introduction

Non-Small Cell Lung Cancer (NSCLC) is the predominant form of lung cancer and a leading cause of cancer-related mortality, with stage III patients facing particularly challenging prognoses. Accounting for approximately 85% of thoracic cancers, NSCLC represents a significant health burden [1]. The survival rates for unresectable stage III NSCLC patients are markedly lower compared to earlier stages, highlighting the need for more effective treatment strategies. Current treatment paradigms involving chemoradiotherapy (CRT) and targeted therapies like Durvalumab have improved progression-free survival (PFS) but locoregional and distant metastasis control still need to improve. Proton Beam Therapy (PBT), particularly with the advent of pencil beam scanning (PBS) technology, offers a promising avenue due to its superior organ-sparing capabilities [2]. However, the relationship between radiation-induced toxicities and patient survival remains underexplored. We aim to fill this gap using targeted machine learning estimation / targeted minimum loss-based estimation (TMLE), a statistical framework for efficient learning from data [3, 4], and causal inference methodologies. The focus of the present study is to understand the causal impact of radiation-induced toxicity—specifically, higher grade (≥ 2, warranting medical intervention) pneumonitis/dyspnea or esophagitis/dysphagia—on the overall survival (OS) of stage III NSCLC patients receiving PBT. Leveraging registry real-world data (RWD) for estimation of these causal effects, we aim to provide insights crucial for personalized radiation therapy planning.

2 Methods

Data description: The Registry Study for Radiation Therapy Outcomes (NCT01255748), led by the Proton Collaborative Group (PCG), collects RWD US-wide with the objective of understanding factors that may lead to reduced toxicity, and better OS and PFS. This ongoing trial has accrued over 23,000 patients, spanning 27 of the 43 proton centers in the US. Patient data has been collected and distributed under an Institutional Review Board study approved by each member institution. Protected health information is de-identified in compliance with the Health Insurance Portability and Accountability Act. Strict data quality control, submission, and adherence to treatment protocols are supervised by the PCG for each participating center.

After a successful application, we received access to 636 patients with stage III NSCLC with the objective to

Patient cohort characteristics	
Age at diagnosis	70 (8 – 97)
BMI	26.0 (13.4 – 15.3)
Sex: Female (50.4%), Male (49.6%)	
Race: White (783.8%), Other (16.2%)	
Tumor location: LLL (10.9%), LUL (22.7%), RLL (13.9%), RML (4.1%), RUL (31.7%), airways (3.1%), other (13.7%)	
Chemotherapy: Cisplatin (5.4%), Etoposide (3.4%), Carboplatin (20.6%), Paclitaxel (11.7%), Pemetrexed (2.1%), Other (9.0%), None (43.6%)	
Histology: Adenocarcinoma (39.3%), Squamous cell carcinoma (29.0%), Non-small cell lung cancer (13.5%), Small cell (7.6%), Other (10.2%)	
Dose Fractionation: 2Gy/fx (71.3%), 10-12Gy/fx (11.4%), Other (28.7%)	

39 Total Demographic Features	
Institution	12 Categories
Race	6 Categories
Ethnicity	3 Categories
Weight (kg)	75 (31 – 190)
Height (cm)	170 (65 – 215)
Days diagnosis to RT	105 (3 – 8148)
Smoking-years	40 (1 – 50)
45 Total Clinical Features	
Tx Interruption?	Yes (30%), No (70%)
Concurrent Chemo?	Yes (44%), No (56%)
Was it metastatic?	Yes (7.6%), No (92.4%1)
Primary tumor size (cm)	3.5 (0.4 – 9.8)
Grade	5 Categories
Clinical stage	Multi-categories for T,N,M
Pathological stage	Multi-categories for T,N,M
ECOG performance	5 Categories
53 Total Dosimetry Features	
Total proton dose (Gy)	60 (1.2 – 95.0)
Fraction numbers	23 (1 – 57)
Dose metrics for OARS: Esophagus, Heart, lungs (contralat and ipsalat)	8 Numerical features/OAR: Dmin, Dmax, Davg, D98%, D95%, D50%, d1cc, d2cc

Table 1: Description of the data and features included in the analysis of the causal effect of radiation toxicities on overall survival. The left hand side shows the patient cohort characteristics from the initial 636 patients spanning 12 institutions, and the right hand side provides a breakdown of patient features into demographic, dosimetric, and clinical categories with examples.

understand the factors leading to life-threatening radiation toxicities, and their impact on OS and PFS. We summarize the information considered in our analysis in Table 1. Out of the 636 patients, 368 were treated with uniform or passive scattering (ST) and 268 were treated with PBS. Most patients received 60 Gy dose. Significant, grade $geq2$ esophagitis/dysphagia (19.7% of patients) and pneumonitis/dyspnea (25.6% of patients) developed within one year of treatment. Patients treated with ST were more likely to develop any grade $geq2$ toxicity (45%) than those treated with PBS (33%). Patients receiving ST also had worse OS, with 28% alive at 5 years with ST versus 43% with PBT.

Data structure: We introduce some notation to describe the data structure and will draw on this in subsequent sections.[1] The data consists of n independent and identically distributed observations $O = (W, A, \Delta, \tilde{T})$, where W is comprised of covariates (Table 1); A is a binary toxicity indicator of grade $geq2$; $\tilde{T}$ is the last time point at which a subject was monitored, the minimum of the censoring time (C) and outcome event time (T); and Δ is an event type indicator (0: censored, 1: death), $\Delta = I(T < C)$.

Realistic statistical model: We express the likelihood and statistical model ($\mathscr{M}$) for time-to-event data similarly to [8], in terms of the intensities of the failure and censoring time processes: $N(t) = I(\tilde{T} \leq t \mid \Delta = 1)$ and $A_C(t) = I(\tilde{T} \leq t, \Delta = 1)$, respectively. [2] For the sake of generality, we recode the observed data O in discrete time as a longitudinal structure ordered in time $O = (W, A, N(0), A_C(0), \ldots, N(\tau), A_C(\tau), N(\tau+1))$, where τ is a maximal follow up time so that each $\tilde{t} \geq \tau + 1$. This can be more succinctly expressed as $O = (W, A, \bar{N}(\tau+1), \bar{A}_C(\tau))$, where the over-bar denotes the entire history. We factorize the likelihood of the observed data according to the time ordering that generated it: $p(O) = q_W(W) g_A(A \mid W) \Pi_{t=0}^{\tau} g_{A_C(t)}(A_C(t) \bar{A}_C(t-1), \bar{N}(t), W, A) \Pi_{t=0}^{\tau} q_{N(t)}(N(t) \mid \bar{N}(t-1), \bar{A}_C(t-1))$, where each conditional density is conditioning on the variables realized prior to the variable in question. Further, $g_{A_C(t)}(1 \mid \bar{A}_C(t-1), \bar{N}(t), W, A) = I(A_C(t-1) = 0, N(t) = 0, W, A) \lambda_C(t \mid W, A)$, and $q_(N(t))(1 \mid \bar{N}(t-1), \bar{A}_C(t-1) = I(N(t-1) = 0, A_C(t-1) = 0) \lambda_T(t \mid W, A)$, where $\lambda_C(t \mid W, A) = P(C = t \mid C > t-1, N(t) = 0, W, A)$ and $\lambda_T(t \mid W, A) = P(T = t \mid T > t-1, A_C(t-1) = 0, W, A)$. Under the coarsening at random assumption on C and T, stating that the censoring and event times are conditionally independent given W, A, the intensities of $N(t)$ and $A_C(t)$ reduce to indicators of being at risk of jumping, multiplied by the conditional hazards of C and T, respectively. That is, the conditional hazards reduce to $\lambda_C(t \mid W, A) = P(C = t \mid C > t-1, W, A)$ and $\lambda_T(t \mid W, A) = P(T = t \mid T > t-1, W, A)$.

The density p of O can be parameterized as $p = p_(q_w, \lambda_C, \lambda_T, g)$. $\mathscr{M}$ is therefore determined by domain knowledge regarding these 4 nuisance parameters: λ_C and g, with q_w and λ_T. In the present study, it remains fully nonparametric. In Section 4, we discuss the potential to draw on natural experimentation of PBT type via instrumental variable analysis.

Causal estimand, causal model, and identification: It is thought that high grade toxicities impact OS and PFS, but their causal effect has not been quantified and is confounded by other covariates. Here, we aim to estimate the causal effect of radiation-induced toxicities on OS to further our understanding of the role of these toxicities in the management of cancer patients receiving radiation therapy. Figure 2 shows a directed acyclic graph (DAG), which illustrates the relationship between our covariates, exposure of interest (toxicities), and primary and secondary outcomes of interest. Examining the structure of a DAG can help identify sufficient sets of confounders to adjust for in the analysis [10, 11]. The study's causal model was formulated based on the relationships between the variables represented by the DAG.

The causal estimand is a mathematical quantity which represents the answer to the causal question of interest. Frameworks often used to guide evidence generation in clinical research—the International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH) E9(R1) Guidelines [12], Target Trial Emulation [11, 13, 14], PICOT [15], and Targeted Learning and Causal Roadmaps [5–9]—similarly prompt investigators to define components of a causal estimand. We review each attribute in the context of this study.

1. population: reflected by the PCG trial's eligibility criteria (inclusion criteria include planned treatment with proton radiation therapy and ability to understand and sign IRB approved consent), as well as stage III NSCLC patients, and restricted to those receiving PBT.

[1] This presentation of the methodology mirrors the Targeted Learning Roadmap / Causal Roadmap: a systematic guide to navigating the study design and analysis challenges inherent in real-world studies [5–9].

[2] A statistical model ($\mathscr{M}$) is a set of possible data distributions. Realistic $\mathscr{M}$ align with subject-matter knowledge, i.e., rule out distributions that are incompatible with known truth (e.g., outcome bounds or knowledge about the clinical decision-making), and avoid making overly restrictive (e.g., parametric) assumptions that preclude the true data distribution [5].

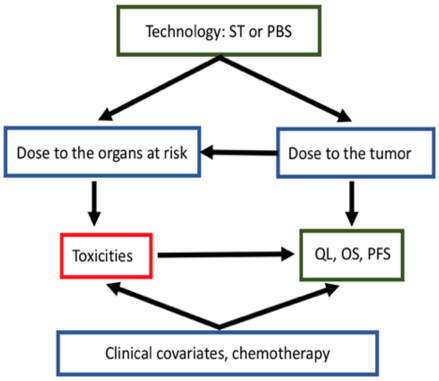


Table 2: Directed acyclic graph (DAG) showing the relationship between the covariates, exposure (toxicity), and outcomes of interest: overall survival (OS), progression free survival (PFS) and quality of life (QL). The direction of the arrow denotes the flow of information. Exogenous errors are not shown.

2. treatment/exposure variable: pneumonitis/dyspnea, dichotomized according to whether it is a high grade toxicity ≥ 2 (warranting medical intervention) [16]. We also consider esophagitis/dysphagia as another exposure variable in an independent analysis, and similarly dichotomize it according to whether it's a high grade toxicities ≥ 2.

3. endpoint/outcome variable: survival time, $\tilde{T}$.

4. summary measure:we are interested in contrasting 5-year cumulative incidence of death under grade ≥ 2 toxicity versus grade < 2 or no toxicity, and under no right censoring.

5. intercurrent events: defined as "events occurring after treatment[/exposure] initiation that affect either the interpretation or the existence of the measurements associated with the clinical question of interest" [12]. Our primary "treatment" of interest is actually toxicity, so issues like treatment non-compliance or discontinuation of treatment are not considered an intercurrent event here. In future studies that will evaluate PFS outcome, mortality will be a competing risk that precludes observing the time to disease progression and will need to be addressed. For instance, we may define a composite outcome: disease progression or death [8].

We define our causal estimand as the cumulative incidence ratio (CIR) of mortality by 5 years. Consider an intervention-specific causal estimand, $\psi_a^F(P_X) = P(T_a \geq 5)$. We have that $P(T_a \geq 5) = E_W P(T_a \geq 5 \mid W, A = 1) = E_W \Pi_{s \leq 60}[1 - \lambda_T(s \mid W, A = 1))]$. Denote the latter expression by $\psi_a(P)$, establishing the identification of $\psi_a^F(P_X)$ under exposure a. Note that $P(T_a \leq 5) = 1 - \psi_a^F(P_X)$.[3] Next we define the conditional survival function of T given A, W as $S(t \mid A, W) = \Pi_{s \leq t}[1 - \lambda_T(s \mid W, A)]$ so that $\psi_1(P) = E_W S(5 \mid A = 1, W)$ and $\psi_0(P) = E_W S(5 \mid A = 1, W)$. The CIR is a function of these two statistical estimands, $\psi_{CIR}^{obs}(P) = ([1 - \psi_1(P)])/([1 - \psi_0(P)])$.

Statistical estimation: For causal inference and in settings with complex data structures or where traditional modelling assumptions (e.g., proportional hazards) are unreasonable, TMLE is more appropriate. In general, TMLE typically consists of two stages: (1) generate initial estimates of the (relevant portions of the) underlying data-generating mechanism, then (2) modify them in such a way that endows the resulting plug-in estimator with desirable statistical properties, including efficiency and unbiased estimation in large samples as well as finite-sample robustness. The *survtmle* package in R is a software program for designed for TMLE survival analysis. We used version 1.1.3 to calculate the 5-year CIR of mortality [17, 18]. Super Learner (SL) was used to estimate the propensity score, censoring and failure time processes (step 1 of TMLE) [19]. The SL procedure was abridged in this initial analysis for the sake of time. We considered 2 cross-validation (CV) folds and a k-fold/V-fold CV scheme, and an SL library incorporating lasso regression. Upcoming analyses will follow SL specification guidelines [20], considering 20 CV folds and a rich and diverse SL library that captures a diversity of potential underlying functional forms.

[3] Causal estimands are defined in terms of the full-data P_X implied by the post-intervention distribution on the causal model, where counterfactual outcomes are generated. To learn from data we must define a statistical estimand, which can be observed in the real world. Expressing the causal estimand as a statistical estimand requires so-called causal "identification" assumptions like no unmeasured confounding / coarsening at random [5].

3 Results

The study preliminary findings (Table 3) present estimated CIRs for mortality and corresponding 95% confidence intervals—contrasting higher grade radiation toxicity with lower/no toxicity and considering two specific toxicities (pneumonitis/dyspnea and esophagitis/dysphagia)—after correcting for all other variables in Table 1. These results span various time points up to five years following radiation therapy initiation. For pneumonitis/dyspnea, the estimated CIRs (denoted by $\hat{\psi}$ in 3) ranged from 1.18 at 30 days to 1.00 at 1095 days and beyond, gradually converging to 1.00 over the 5-year period. This indicates an initially elevated risk of mortality associated with toxicity that diminishes over time. The 95% confidence intervals start wider and progressively narrow, reflecting increasing precision in the estimates as time progresses. In contrast, for esophagitis/dysphagia, the CIRs consistently hovered around 1.00 across all time points, with the exception of a notable increase to 1.46 at 182 days. This suggests that esophagitis/dysphagia does not consistently impact mortality risk over the five-year period, except for a transient increase at around six months post-therapy.

The observed temporal patterns in the mortality risks associated with pneumonitis/dyspnea and esophagitis/dysphagia offer important insights. The initial higher mortality risk in patients with pneumonitis/dyspnea could be attributed to the acute effects of this toxicity, which appears to resolve or be managed effectively over time. This underscores the importance of early detection and intervention for pneumonitis/dyspnea in the treatment of NSCLC patients. The transient increase in mortality risk at six months for patients with esophagitis/dysphagia warrants further investigation. It might be linked to specific treatment-related complications or other factors not captured in the analysis. The lack of a sustained impact of esophagitis/dysphagia on

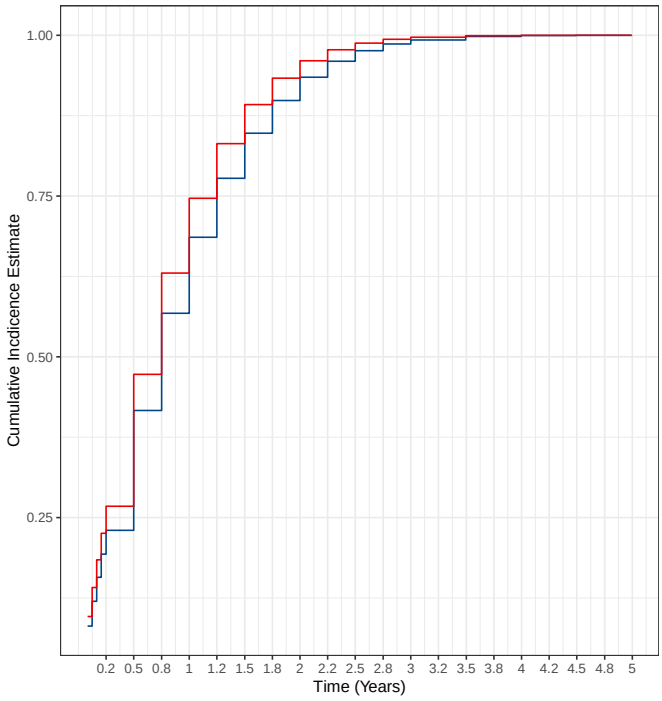


Figure 1: Counterfactual, covariate-adjusted cumulative incidence estimates under an intervention of higher grade ≥ 2 pneumonitis/dypsnea toxicity (red line) versus no or lower grade < 2 pneumonitis/dypsnea toxicity (blue line) over 5 years following proton beam therapy initiation for stage III Non-Small Cell Lung Cancer.

mortality across the other time points could indicate effective management strategies or a lesser severity of this toxicity in terms of long-term mortality risk.

4 Discussion

This study provides a detailed temporal understanding of the impact of two common radiation toxicities on mortality in NSCLC patients receiving PBT. Findings emphasize the importance of nuanced patient monitoring, especially in the early months following treatment. The differences in mortality risk profiles between pneumonitis/dyspnea and esophagitis/dysphagia highlight the need for toxicity-specific management strategies. These insights contribute to the growing body of knowledge on long-term outcomes in NSCLC patients undergoing advanced radiation therapies and underscore the value of continuous monitoring and intervention strategies tailored to specific treatment-related toxicities.

Future research will incorporate sensitivity analyses using instrumental variables (IV) to examine the impact of causal bias (e.g., unmeasured confounding) on the study conclusions. An IV analysis does not require that there are no unmeasured confounders but it assumes that if they exist and are correlated to the outcome then they cannot be correlated to the IV. We aim to leverage PBT type (PBS or ST) as an IV; access to technology is a commonly considered as an IV in causal inference [21]. For reassurance, we will use clinical and dosimetric covariates to determine if they are correlated to PBT type. If not, this will provide evidence that access to technology is acting as an IV in our case.

The use of negative control outcomes will also be explored to validate causal assumptions and identify potential biases in the analysis. If negative controls show an association with toxicity, this could indicate the presence of unaccounted biases, prompting a re-evaluation of the analysis methods or the causal assumptions made in the study. Some examples of potential negative control outcomes include unrelated health outcomes, such as conditions with a genetic basis or unrelated infections; early time-point events that occur too soon after the start of radiation therapy to be plausibly caused by it; and non-radiation responsive conditions, such as osteoarthritis or chronic skin conditions unrelated to radiation exposure. Additionally, in future analyses, we will further refine the SL implementations considered in *survtmle* to enhance the robustness of the findings.

This research contributes to the field of causal inference in oncology. An illustration of the power of TMLE and causal inference for rigorously analyzing complex RWD, this study offers insights that are generally applicable for generating reliable RWE in oncology. We strive to empower cancer researchers to use these tools in their work for causal targeted machine learning.

	Pneumonitis / Dyspnea			Esophagitis / Dysphagia		
time	$\hat{\psi}$	lb	ub	$\hat{\psi}$	lb	ub
30	1.18	1.01	1.38	1.00	0.87	1.15
45	1.18	0.92	1.52	1.00	0.76	1.32
60	1.17	0.95	1.44	1.00	0.78	1.28
75	1.17	0.90	1.51	1.00	0.76	1.31
91	1.16	0.93	1.46	1.00	0.78	1.28
182	1.13	0.96	1.34	1.46	1.20	1.77
274	1.11	0.97	1.27	1.00	0.87	1.15
365	1.09	0.98	1.21	1.00	0.89	1.13
456	1.07	0.98	1.17	1.00	0.90	1.11
548	1.05	0.98	1.13	1.00	0.92	1.08
639	1.04	0.98	1.10	1.00	0.93	1.08
730	1.03	0.97	1.09	1.00	0.95	1.06
821	1.02	0.98	1.06	1.00	0.96	1.04
912	1.01	0.95	1.07	1.00	0.96	1.04
1004	1.01	0.97	1.04	1.00	0.97	1.03
1095	1.00	0.98	1.03	1.00	0.95	1.05
1276	1.00	0.98	1.02	1.00	0.97	1.03
1458	1.00	0.99	1.02	1.00	0.98	1.02
1640	1.00	0.98	1.02	1.00	0.98	1.03
1823	1.00	0.99	1.01	1.00	0.99	1.01

Table 3: Covariate-adjusted estimates of the cumulative incidence ratio ($\hat{\psi}$) of mortality under two radiation toxicities—grade *geq*2 versus no or grade < 2 pneumonitis/dyspnea, and grade *geq*2 versus no or grade < 2 esophagitis/dysphagia—with accompanying 95% confidence interval lower bounds (lb) and upper bounds (ub). Estimates were obtained across various time points, spanning 30 days and up to 5 years following proton beam therapy radiation treatment for stage III Non-Small Cell Lung Cancer.

Acknowledgements

This work was supported in part by the National Institute of Biomedical Imaging and Bioengineering of the National Institutes of Health (Grant K08EB026500) and the National Institute of Allergy and Infectious Diseases (Grant 5R01AI074345-09).

References

[1] A. Casal-Mouriño, A. Ruano-Ravina, M. Lorenzo-González, et al. "Epidemiology of stage III lung cancer: frequency, diagnostic characteristics, and survival". *Translational Lung Cancer Research* 10.1 (2021), p. 506.

[2] M. J. LaRiviere, P. M. G. Santos, C. E. Hill-Kayser, et al. "Proton therapy". *Hematology/Oncology Clinics* 33.6 (2019), pp. 989–1009.

[3] M. J. Van der Laan, S. Rose, et al. *Targeted learning: causal inference for observational and experimental data*. Vol. 4. Springer, 2011.

[4] M. J. Van der Laan and S. Rose. *Targeted learning in data science*. Springer, 2018.

[5] S. Gruber, H. Lee, R. Phillips, et al. "Developing a targeted learning-based statistical analysis plan". *Statistics in Biopharmaceutical Research* 15.3 (2023), pp. 468–475.

[6] S. Gruber, R. V. Phillips, H. Lee, et al. "Targeted Learning: Toward a Future Informed by Real-World Evidence". *Statistics in Biopharmaceutical Research* (2023), pp. 1–15.

[7] S. Gruber, R. V. Phillips, H. Lee, et al. "Evaluating and improving real-world evidence with Targeted Learning". *BMC Medical Research Methodology* 23.1 (2023), p. 178.

[8] S. Gruber, H. Lee, R. Phillips, et al. "Causal Inference with Targeted Learning for Producing and Evaluating Real-World Evidence". *Real-World Evidence in Medical Product Development*. Springer, 2023, pp. 125–143.

[9] L. E. Dang, S. Gruber, H. Lee, et al. "A causal roadmap for generating high-quality real-world evidence". *Journal of Clinical and Translational Science* 7.1 (2023), e212.

[10] J. Pearl. *Causality*. Cambridge university press, 2009.

[11] M. A. Hernán and J. M. Robins. *Causal Inference: What If*. Chapman Hall/CRC, 2020.

[12] I. C. for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use. *ICH E9(R1) Addendum to Statistical Principles for Clinical Trials on Choosing Appropriate Estimands and Defining Sensitivity Analyses in Clinical Trials*. https://database.ich.org/sites/default/files/E9-R1_Step4_Guideline_2019_1203.pdf. 2019.

[13] M. A. Hernán and J. M. Robins. "Using big data to emulate a target trial when a randomized trial is not available". *American journal of epidemiology* 183.8 (2016), pp. 758–764.

[14] M. A. Hernán, W. Wang, and D. E. Leaf. "Target trial emulation: a framework for causal inference from observational data". *Jama* 328.24 (2022), pp. 2446–2447.

[15] L. P. Rios, C. Ye, and L. Thabane. "Association between framing of the research question using the PICOT format and reporting quality of randomized controlled trials". *BMC Medical Research Methodology* 10.1 (2010), pp. 1–8.

[16] U. D. of Health and H. Services. *Common Terminology Criteria for Adverse Events (CTCAE) Version 5.0*. https://ctep.cancer.gov/protocoldevelopment/electronic_applications/docs/ctcae_v5_quick_reference_5x7.pdf. 2017.

[17] D. C. Benkeser, M. Carone, and P. B. Gilbert. "Improved estimation of the cumulative incidence of rare outcomes". *Statistics in Medicine* (2017). DOI: 10.1002/sim.7337.

[18] D. C. Benkeser and N. S. Hejazi. *survtmle: Targeted Minimum Loss-Based Estimation for Survival Analysis in R*. 2017. DOI: 10.5281/zenodo.835868.

[19] E. C. Polley and M. J. Van der Laan. "Super learner in prediction" (2010).

[20] R. V. Phillips, M. J. van der Laan, H. Lee, et al. "Practical considerations for specifying a super learner". *International Journal of Epidemiology* 52.4 (2023), pp. 1276–1285.

[21] E. A. Vertosick, M. Assel, and A. J. Vickers. "A systematic review of instrumental variable analyses using geographic region as an instrument". *Cancer epidemiology* 51 (2017), pp. 49–55.

Rethinking the Elective Target Volume: Modelling Bilateral Lymphatic Head and Neck Tumour Progression Using Probability Theory

Kristoffer Moos[1,2], Jesper G. Eriksen[2,3], Anne I.S. Holm[3], and Stine Sofia Korreman[1,2]

[1]Institute of Clinical Medicine, Aarhus University, Aarhus, Denmark
[2]Danish Centre for Particle Therapy, Aarhus University Hospital, Aarhus, Denmark
[3]Department of Experimental Clinical Oncology, Aarhus University Hospital, Aarhus, Denmark

Abstract: In radiotherapy, a low-dose elective target volume is often defined to include lymph nodes that may be malignant. The prescribed dose to the elective target volume results in large irradiated volumes with associated risk of severe side-effects. In this study we establish a model for bilateral tumour progression in patients with oropharyngeal squamous cell carcinoma for personalised elective target volume definition. An expansion of a previously proposed Bayesian network model was employed to estimate the probability of bilateral nodal spread. This study highlights the potential for reconsidering the elective target volumes through probabilistic modelling, aiming to decrease unnecessary radiation dose to organs-at-risk and healthy tissue.

1 Introduction

Advancements in radiotherapy over the past decades have led to increasingly personalised irradiation volumes, with radiation beams that closely conform to the tumour, and high precision beam delivery. Despite efforts to minimize exposure of healthy tissue to unnecessary radiation, irradiation of such tissues remains inevitable and often leads to side-effects. A substantial part of patients with head and neck cancer (HNC) whose primary treatment is radiotherapy, suffer from late morbidity including xerostomia, oral discomfort, difficulties chewing, speaking, and swallowing, as well as loss of taste and reduced periodontal hygiene.

In current curatively intended RT of HNC, the target volumes include the gross tumour volume (GTV) and Clinical Target Volume (CTV) that confines the subclinical disease. The CTV is often subdivided accounting for factors of varying importance. Particularly a low-dose elective CTV (CTVe) is often defined which encompasses elective lymph nodes (LNs) that may involve malignant deposits. Although the prescribed dose to the elective CTV is relatively low, it still covers a substantial portion of the head and neck region which leads to undesired irradiation of adjacent organs-at-risk (OAR) and surrounding healthy tissue. Studies have established that large treatment margins and irradiated volumes, are the leading factors associated with increased acute and late side-effects. [1], [2], [3], [4], [5]

A recent study by Kristensen et.al [6] has shown that the majority of local-regional failures were found within the high-dose volume in head and neck cancer squamous cell carcinoma. This is supported by Bollen et. al [7] who found that the majority of recurrences were observed within the high-dose target volume following IMRT/VMAT. Likewise, a study by Razavian et.al. [8] investigated the rate of contralateral neck failure on tonsil cancer patients treated with unilateral radiotherapy. They found that unilateral radiotherapy is associated with low risk of contralateral LN failure and should be considered if patients are diagnosed with lateralised tonsil cancer. However, the characteristic across these studies is the emphasis on the potential for overtreatment under the current guidelines. Considering the severe side effects, which to some extent limits the level of high doses, an investigation into potential adjustments, such as a reduction of the CTVe, is justified. A strategy that could allow us to reduce the volume, is by defining the CTVe using a probabilistic approach. By statistical estimations of the risk of malignant formations in specific LNs, we can potentially selectively omit LNs with low risk without compromising the therapeutic effectiveness. Pouymayou et.al [8] have previously suggested a statistical framework to estimate lymphatic nodal involvement for patients receiving primary radiotherapy. The framework was based on the anatomical principles of the lymphatic drainage system, modelled as a Bayesian Network (BN). Lymph node levels (LNL) were represented as nodes connected with directed arcs between each other. The study established, through inference of the BN, the viability of estimating the probability of malignant LNs. They observed low risk of nodal involvement for specific LNLs with different states of macroscopic involved LNLs obtained from diagnostic imaging.

The approach outlined in Pouymayou et al [9], primarily focuses on ipsilateral tumour progression, which bears certain limitations. Firstly, focusing solely on ipsilateral tumour progression might neglect important parts of the disease dynamics. Secondly, bilateral involvement is common in many types of HNC, and neglecting this aspect could lead to an inaccurate assessment of the overall disease state, increasing the risk of missing metastatic developments. In this study, we expand to incorporate bilateral tumour progression into the model, to provide a more nuanced modelling of the disease dynamics.

2 Materials and Methods

2.1 Patients

In this retrospective study we included 460 patients with oropharyngeal squamous cell carcinoma treated at Aarhus University Hospital in the period 2013-2020. For calculating the probabilities of lymphatic nodal involvement, we used information on tumour characteristics. Firstly, T-stage for each patient was used as an approximation for greater risk of microscopic disease. Secondly, one of the factors linked with contralateral LN metastasis is whether the tumour is lateralised or not.

Therefore, in this study, we only included patients with central tumours, defined as those situated ≤ 1 cm from the midsagittal line. Lastly, location of both ipsilateral and contralateral LN metastasis was derived for the purpose of calculating the probabilities of nodal involvement. Each LN metastasis was identified through combination of PET/CT and MRI. In instances where the malignancy of LNs was uncertain, the identification process was supplemented with fine needle aspiration.

2.2 Bayesian Network for bilateral tumour progression

The BN presented in this study is based on the same principles as described by Pouymayou et.al. [9]. The graphical representation of the BN can be divided into three different phases: (1) the association between macroscopic and microscopic states, (2) spread from primary tumour directly to each LN level (LNL), and (3) spread between different LNL. The BN is based on the most relevant LNL, namely, I, II, III, IV, V, and VI. Furthermore, each equation as presented in the subsequent section holds true for both calculating the ipsilateral and contralateral lymph nodal involvement. We utilise two different BN for probability estimation; one for early-stage tumours and a second for advanced-stage tumours. The early-stage tumours are defined as low grade tumours (T1-T2) whereas advanced-stage is defined as high grade tumours (T3-T4). The graphical representation and calculations are identical for both BNs, the only difference lying in the dataset used for parameter learning. A visual representation of the BN employed in this study, can be seen in figure 1.

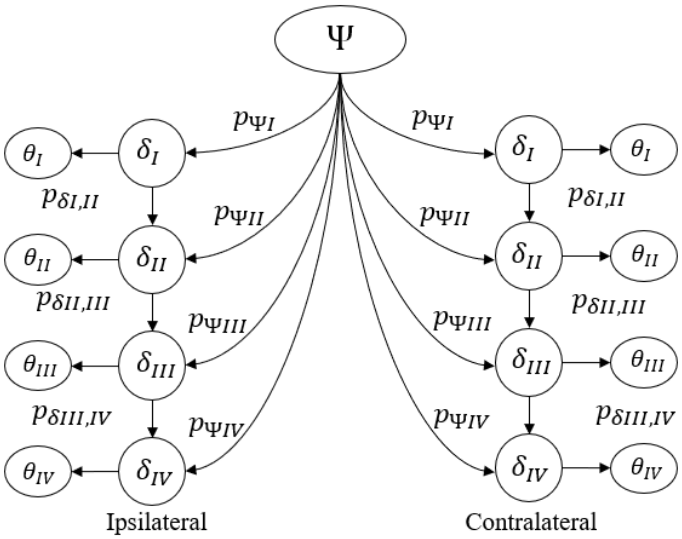


Figure 1: Bayesian Network of bilateral tumour progression. Each lymph node (δ_k) is connected through directed arcs. Between each lymph node, we have the parameter $p_{\delta k,\delta k+1}$ that describes the probability that δ_k spreads to δ_{k+1}. We assume that the primary tumour (Ψ) can spread directly to each δ_k with probability Ψ_k. The parameter θ_k determines the macroscopic state of the associated lymph node by the sensitivity and specificity of the diagnostic image modality.

Firstly, each LNL is associated with two random binary states: macroscopic and microscopic states. The macroscopic random state, defined as θ, determines the presence of macroscopic metastasis (from e.g. CT, MRI, PET) within the LNL. The microscopic random variable, represented as δ, specifies whether the LNLs contains microscopic metastasis. The association between these variables are determined by the sensitivity and specificity of the imaging modality. This is given by [8]:

$$\begin{pmatrix} P(\theta_k^1|\delta_k^1)\, P(\theta_k^0|\delta_k^1) \\ P(\theta_k^1|\delta_k^0)\, P(\theta_k^0|\delta_k^0) \end{pmatrix} = \begin{pmatrix} sens & 1-sens \\ 1-spec & spec \end{pmatrix}$$

Secondly, we assume that primary tumours can spread directly to each LNL. Thus, the tumour spread directly to each LNL can be estimated with the probability $P_{\Psi k}$. This is formally written as [7]:

$$\begin{pmatrix} P(\delta_k^1|\Psi_k^1)\, P(\delta_k^0|\Psi_k^1) \\ P(\delta_k^1|\Psi_k^0)\, P(\delta_k^0|\Psi_k^0) \end{pmatrix} = \begin{pmatrix} P_{\Psi k} & 1-P_{\Psi k} \\ 0 & 1 \end{pmatrix}$$

Lastly, we consider the association between LNLs. It is essential to consider the causality of the lymphatic system. Therefore, we model the association between different LNLs based on the dynamics of lymphatic drainage. Thus, LNL II receives efferent lymphatics from LNL I, and so on. The probability of tumour cells spreading between two efferent LNLs, is defined as $P_{\delta ij}$ (e.g. $P_{\delta II-III}$ is the probability of LNL II spreading to LNL III). We say that if a single parent Δ_i is in state 1, it will cause δ_k to be 1 with probability u_i. The Noisy-OR model is then described as:

$$P(\delta_k^0|\Delta) = \prod_{\lambda_i \in \lambda^+} (1-u_i)$$

$$P(\delta_k^1|\Delta) = 1 - \prod_{\lambda_i \in \lambda^+} (1-u_i)$$

Where Δ denotes the parents of the respective δk and the subset of parents that are in state 1, are denoted as λ^+. For instance, if one wish to calculate $P(\delta_{II} = 1\,|\delta_I = 1, \Psi = 0)$, we describe the Noisy-OR as $1 - P(1 - P_{\delta I-II})$. Calculating the probability of harbouring microscopic disease for a set of observed macroscopic involved LNs, is described by Baye's law. The conditional probability, with a set of parameters Φ and macroscopic involvement θ, is then given by:

$$P(\delta_i|\theta, \Phi) = \frac{\sum_{k:\delta_{ki}=1} P(\theta|\delta_i, \Phi) P(\delta_i|\Phi)}{P(\theta|\Phi)}$$

Where the nominator represents a marginalisation over all hidden states, in which δ_i is macroscopically observed.

2.3 Parameter estimation

The parameterisation of the BN follows the principles outlined in Ludwig et al [10]. This approach computes the likelihood of observing a patient cohort given a set of model

parameters, namely, the spread probabilities ($P_{\psi k}$, $P_{\delta ij}$). However, due to the model simplification using logical OR gate truth tables, standard maximum likelihood estimation can no longer be directly obtained through a simple ratio of occurrences. Therefore, to address this, we established an optimization scheme to correctly estimate the spread probabilities. In this study, we employ the affine-invariant ensemble sampler to draw samples from the likelihood function [11].

3 Results

Of 460 included patients, 428 patients were suitable for inclusion after sorting out patients that either had missing information of LN metastasis and/or T-stage or did not have a central tumour. We further divided the 428 patients into two groups: early-stage (275 patients) and advanced-stage tumours (153 patients). The estimated probabilities for observing microscopic disease for all LNs are reported in table 1. For patients with N_0, we observe that ipsilateral LNL II for both early-stage and advanced-stage has the highest probability of finding microscopic disease (37.5% and 46.3%, respectively). For ipsilateral LNL I, III and IV, we observe probabilities of 0.8%, 5.2%, and 0.8% for early-stage tumours and 2.1%, 7.8%, and 1.2% for advanced-stage tumours, respectively. However, an increase in probability of harbouring microscopic disease is observed when the LNLs are macroscopically involved. For instance, if ipsilateral LNL II is macroscopically observed for early-stage tumours, the probability of finding microscopic disease in ipsilateral LNL II increases from 37.5% to 93.3%, LNL III from 5.2% to 11.3%, and LNL IV from 0.8% to 1.1%. Correspondingly, for advanced-stage tumours the probabilities increase from 46.3% to 95.2%, 7.8% to 14.3%, and 1.2% to 1.5%, respectively. Similar tendencies are seen for ipsilateral as well as contralateral macroscopically involved LNL II/III and LNL III, for both early-stage and advanced-stage tumours.

Macroscopic involvement of LNL	*Microscopic involvement of LNL (%)*			
	Ipsi. I	**Ipsi. II**	**Ipsi. III**	**Ipsi. IV**
Ipsi. I	14.8/**32.8**	38.5/**48.0**	5.3/**8.0**	0.8/**1.2**
Ipsi. II	0.9/**2.3**	93.3/**95.2**	11.3/**14.3**	1.1/**1.5**
Ipsi. II/II	0.9/**2.3**	97.7/**98.5**	74.7/**79.4**	4.4/**5.0**
Ipsi. III	0.8/**2.2**	64.3/**73.3**	55.7/**66.1**	3.4/**4.3**
N_0	0.8/**2.1**	37.5/**46.3**	5.2/**7.8**	0.8/**1.2**
	Cont. I	**Cont. II**	**Cont. III**	**Cont. IV**
Cont. I	2.7/**1.7**	2.8/**9.3**	0.4/**1.1**	0.1/**0.4**
Cont. II	0.3/**0.2**	38.3/**69.5**	2.9/**5.7**	0.2/**0.9**
Cont. II/III	0.3/**0.2**	61.0/**85.5**	40.6/**58.3**	1.86/**5.8**
Cont. III	0.1/**0.1**	6.3/**20.4**	7.4/**19.4**	0.4/**2.2**
N_0	0.1/**0.1**	2.6/**9.0**	0.3/**1.0**	0.1/**0.4**

***Table 1**: Probabilities for harbouring microscopic disease given macroscopic involved LNLs. Values reported in bold are for advanced-stage tumours.*

4 Discussion

In this study we have suggested a statistical framework for bilateral tumour progression based on the principles of Pouymayou et.al [9], including contralateral tumour progression. The framework may contribute as a decision tool for physicians, allowing for personalised CTVe definition with omission of low-risk LNLs.

Defining an appropriate risk threshold for LNL omission is not a straightforward process due to the risk/benefit trade-offs, including disease control vs irradiation to OAR and healthy tissue. However, a study by Grégoire et.al. [12] proposed that threshold values for risk assessment, falling within the range of 5-10% for the N_0 patients would be acceptable. Using a threshold of 10%, our study suggests that it might be viable to omit great parts of the CTVe for selected patients. For instance, in patients with N_0 status for both early-stage and advanced-stage tumours, the risk associated with ipsilateral LNL I, III, and IV, as well as contralateral LNL I, II, III, and IV, falls below the threshold of 10%. Similar tendencies are observed for patients with macroscopic findings of ipsilateral LNL II, LNL II/III, and LNL III regardless of tumour stage. In these instances, the risk associated with ipsilateral LNL I and IV, as well as contralateral LNL I, II, III, and IV, also falls below the risk threshold. Looking at the contralateral involvement, we observe that for both early-stage tumours, advanced-stage tumours, and all cases of macroscopic involvement, LNL IV is below the risk threshold. LNL III is below the risk threshold for instances of macroscopic involved LNL II and LNL III, in contrast to ipsilateral involvement where all LNL III must be included.

Several studies have modelled microscopic disease in the LNLs. Pouymayou et.al. [8] constructed a BN for ipsilateral tumour progression for investigating the feasibility of personalised CTVe definition. Comparing our results with theirs, we observe notable similarities across all instances of early-stage tumours and macroscopic involved ipsilateral LNLs, emphasising the consistency and reliability of findings between the two studies. However, it is worth mentioning that our ability to make comparisons is limited to early-stage tumours and ipsilateral LNL. To the best of our knowledge, no other study has modelled bilateral tumour progression, and as a result, we lack the opportunity to thoroughly validate the estimated probabilities for microscopic disease in the contralateral LNLs.

Nevertheless, this study not only establishes consistency with the results of Pouymayou et al. [8], it also proposes the potential for further reduction in the CTVe by taking into account the contralateral LNLs. For instance, each case of macroscopically involved ipsilateral LNLs, the prospect of omitting all contralateral LNLs for both early-stage and advanced-stage tumours is viable, given that these probabilities significantly fall below the 10% risk threshold. Likewise, in specific cases of ipsilateral macroscopic involvement, portions of the ipsilateral LNLs could be omitted. This approach could reduce the dose administered to the OAR and corresponding healthy tissues substantially. Likewise, Ludwig et.al. [9], proposed a Hidden Markov Model for incorporating T-stage. They assumed that T-stage is a surrogate for time and thus predicted lymphatic tumour progression for time dependent parameters. They found that advanced-stage tumours result in more advanced nodal progression than early-stage tumours, and thus the risk of finding microscopic disease increases with greater T-stage. We observe a similar tendency with the model for advanced-stage tumours resulting in larger probabilities for both ipsilateral and contralateral LNLs.

However, it is important to acknowledge the limitations in this study. Firstly, the graphical representation of the BN is based on an assumption regarding of the bilateral lymphatic drainage system. Without correctly considering the causality of the lymphatic drainage system, the BN may misrepresent or overlook important pathways of spread. For instance, our model does not assume a relation between ipsilateral and contralateral involvement. This implies that macroscopic observations in ipsilateral LNLs cannot be considered in assessing the probability of discovering microscopic disease in contralateral LNLs. Addressing this issue proves challenging due to a lack of knowledge about specific microscopic spread patterns. Secondly, testing these models in prospective studies is not straightforward since the CTVe definitions will fall in many categories for different patients. However, a prospective study is currently investigating whether omitting LNLs based on sentinel LN mapping could lead to lower-dose radiation therapy while preserving disease control. Such studies may pave the way for further tests of probabilistic definitions of the CTVe for sparing OAR and healthy tissue from unnecessary irradiation.

5 Conclusion

In this study we successfully presented a statistical framework for assessing the probability of bilateral nodal dissemination for early-stage and advanced-stage oropharyngeal tumours. This study emphasises the feasibility of rethinking the elective clinical target volume, which could reduce radiation dose delivered to OAR and healthy tissue. However, further studies are required to investigate the dosimetric effects and benefits.

References

[1] P. D. de Veij Mestdagh *et al.*, 'SPECT/CT-guided elective nodal irradiation for head and neck cancer: Estimation of clinical benefits using NTCP models', *Radiother. Oncol. J. Eur. Soc. Ther. Radiol. Oncol.*, vol. 130, pp. 18–24, Jan. 2019, doi: 10.1016/j.radonc.2018.07.023.

[2] J. A. Langendijk, P. Doornaert, D. H. F. Rietveld, I. M. Verdonck-de Leeuw, C. R. Leemans, and B. J. Slotman, 'A predictive model for swallowing dysfunction after curative radiotherapy in head and neck cancer', *Radiother. Oncol. J. Eur. Soc. Ther. Radiol. Oncol.*, vol. 90, no. 2, pp. 189–195, Feb. 2009, doi: 10.1016/j.radonc.2008.12.017.

[3] A. Al-Mamgani *et al.*, 'Single Vocal Cord Irradiation: Image Guided Intensity Modulated Hypofractionated Radiation Therapy for T1a Glottic Cancer: Early Clinical Results', *Int. J. Radiat. Oncol. Biol. Phys.*, vol. 93, no. 2, pp. 337–343, Oct. 2015, doi: 10.1016/j.ijrobp.2015.06.016.

[4] P. D. de Veij Mestdagh *et al.*, 'SPECT/CT-guided elective nodal irradiation for head and neck cancer is oncologically safe and less toxic: A potentially practice-changing approach', *Radiother. Oncol. J. Eur. Soc. Ther. Radiol. Oncol.*, vol. 147, pp. 56–63, Jun. 2020, doi: 10.1016/j.radonc.2020.03.012.

[5] A. Navran *et al.*, 'The impact of margin reduction on outcome and toxicity in head and neck cancer patients treated with image-guided volumetric modulated arc therapy (VMAT)', *Radiother. Oncol. J. Eur. Soc. Ther. Radiol. Oncol.*, vol. 130, pp. 25–31, Jan. 2019, doi: 10.1016/j.radonc.2018.06.032.

[6] M. H. Kristensen *et al.*, 'High-dose loco-regional pattern of failure after primary radiotherapy in p16 positive and negative head and neck squamous cell carcinoma – A DAHANCA 19 study', *Clin. Transl. Radiat. Oncol.*, vol. 46, p. 100772, Mar. 2024, doi: 10.1016/j.ctro.2024.100772.

[7] H. Bollen, J. van der Veen, A. Laenen, and S. Nuyts, 'Recurrence Patterns After IMRT/VMAT in Head and Neck Cancer', *Front. Oncol.*, vol. 11, p. 720052, Sep. 2021, doi: 10.3389/fonc.2021.720052.

[8] N. B. Razavian, R. B. D'Agostino, C. R. Steber, C. A. Helis, and R. T. Hughes, 'Association of Unilateral Radiotherapy With Contralateral Lymph Node Failure Among Patients With Squamous Cell Carcinoma of the Tonsil', *JAMA Netw. Open*, vol. 6, no. 2, p. e2255209, Feb. 2023, doi: 10.1001/jamanetworkopen.2022.55209.

[9] B. Pouymayou, P. Balermpas, O. Riesterer, M. Guckenberger, and J. Unkelbach, 'A Bayesian network model of lymphatic tumor progression for personalized elective CTV definition in head and neck cancers', *Phys. Med. Biol.*, vol. 64, no. 16, p. 165003, Aug. 2019, doi: 10.1088/1361-6560/ab2a18.

[10] R. Ludwig, B. Pouymayou, P. Balermpas, and J. Unkelbach, 'A hidden Markov model for lymphatic tumor progression in the head and neck', *Sci. Rep.*, vol. 11, no. 1, Art. no. 1, Jun. 2021, doi: 10.1038/s41598-021-91544-1.

[11] J. Goodman and J. Weare, 'Ensemble samplers with affine invariance', *Commun. Appl. Math. Comput. Sci.*, vol. 5, pp. 65–80, Jan. 2010, doi: 10.2140/camcos.2010.5.65.

[12] V. Grégoire, E. Coche, G. Cosnard, M. Hamoir, and H. Reychler, 'Selection and delineation of lymph node target volumes in head and neck conformal radiotherapy. Proposal for standardizing terminology and procedure based on the surgical experience', *Radiother. Oncol.*, vol. 56, no. 2, pp. 135–150, Aug. 2000, doi: 10.1016/S0167-8140(00)00202-4.

Three-dimensional visualization of DEBI-NNs for supporting interpretability of predictive models relying on neural networks

Amine Boukhari[1,2], Boglarka Ecsedi[2,3], Nassib Abdallah[1], David Haberl[4], Clemens Spielvogel[4], Laszlo Papp[2*], and Mathieu Hatt[1*]

[1]Laboratory of medical information processing (LaTIM), INSERM UMR 1101, Univ Brest, Brest, France
[2]Center for Medical Physics and Biomedical Engineering, Medical University of Vienna, Vienna, Austria
[3]Georga Institute of Technology, Atlanta, USA
[4]Division of Nuclear Medicine, Medical University of Vienna, Vienna, Austria
[*]Contributed equally

Abstract In the domain of medical imaging, model interpretability is required for experts to optimize their models and it is paramount for clinical trust as well. This work explores the interpretability potential of recently-proposed Distance-Encoding Biomorphic-Informational Neural Networks (DEBI-NN), using as an example application the diagnosis of human papillomavirus (HPV) status in the HECKTOR (head and neck tumor segmentation and outcome prediction in positron emission tomography / computed tomography (PET/CT) images) 2022 challenge dataset. DEBI-NNs train 3D spatial coordinates instead of weights, thus, they require significantly less parameters to train compared to conventional NNs and they are inherently 3D objects. This study relies on a 3D visualization approach to understand the spatial configurations of DEBI-NNs, shedding light on the network's consistency and aiding in identifying optimal configurations. Two training settings were tested with and without spatial dropout (Setting A and B, respectively), resulting in 90% and 57% test balanced accuracy respectively. Spatial dropout supports the approximation of a task-specific DEBI-NN during training. Results demonstrate the predictive power of DEBI-NNs and the added value of 3D DEBI-NN renderings in explaining them.

1 Introduction

In the context of medical imaging applications, where explainability and interpretability of models is crucial for trust and acceptance of the tools in clinical routine practice, relying on deep neural networks (NNs) results in "black box" models, with no explanation of the output prediction. In addition, these models are notoriously data-hungry and dealing with typically small and imbalanced medical imaging datasets poses significant challenges when employing NNs. In this paper we report preliminary results regarding the potential for interpretability, of a recently published concept called DEBI-NN (for Distance-Encoding Biomorphic-Informational Neural Networks) [1]. This new NN architecture has already demonstrated its ability to provide similar classification accuracy to conventional NNs, on small datasets, while requiring a significantly reduced number of parameters to train. Unlike a conventional NNs where weights are trained, DEBI-NNs train 3D coordinates for each neuron, so that the weight of the information carried by a neuron is represented by the distance between a neuron and its input. In this way, training 3 coordinates of a single neuron is equivalent to modifying the weights of all its inputs, thus drastically reducing the number of parameters to be trained [1]. Figure 1 shows a rendering of an example 3D trained DEBI-NN. Our hypothesis is that 3D rendering of DEBI-NNs could facilitate user interaction and understanding of the architecture training and use of input data, and help in identifying suboptimal configurations and training hyperparameters.

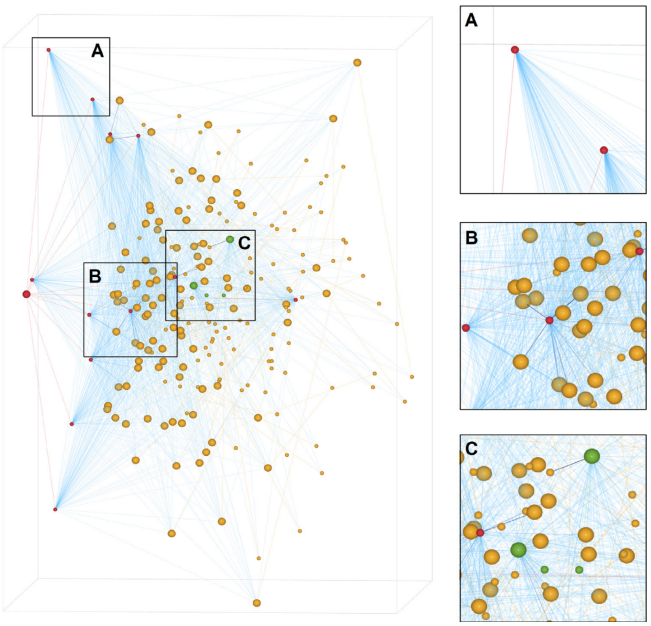


Figure 1: A visual representation of a trained DEBI-NN. The neurons in the input, hidden, and output layers are color-coded as red, yellow, and green respectively [1].

2 Methods

2.1 Dataset and application

In this study, we used the HECKTOR (head and neck cancer segmentation and outcome prediction from positron emission tomography (PET)/computed tomography (CT) images) challenge 2022 dataset [2], and used as a modeling task the diagnosis of the Human Papillomavirus (HPV) status as a binary outcome. Out of the more than 800 patients of the HECKTOR 2022 dataset, the HPV status is known for 563 of them. The training dataset consists of 279 patients with a positive HPV and 61 patients with a negative HPV, resulting in a minority subgroup imbalance of 22 %. Similarly, the

test dataset comprises 203 positive and 20 negative, hence a minority subgroup imbalance of 9.8 %.

2.2 Input features for modeling

IBSI (imaging biomarker standardization initiative)-compliant Radiomic features (shape descriptors, intensity and textural metrics) were extracted from each patient's lesions in PET and CT images using Pyradiomics and based on the provided expert-derived tumor contours.
Available clinical variables and the extracted PET/CT radiomic features were used as input data for modeling HPV status.

2.3 Features preprocessing

The features preprocessing phase involved employing the minimum Redundancy Maximum Relevance (mRMR [3]) method to identify and select the top 36 features with the highest ranking and non-redundancy from the training set. These selected features were then consistently extracted from the test set to maintain uniformity in the dataset. The choice of N=36 was deliberate, representing approximately 10 % of the total sample counts, which amounted to 340 samples in this study.
To ensure that the data were on a common scale, each feature within the training set underwent z-score standardization. This involved calculating the mean and standard deviation values for each feature in the training set and transforming the feature values into their respective z-scores. The calculated mean and standard deviation values from the training set were then used to standardize the corresponding features in the test set.

2.4 DEBI-NN training

Two DEBI-NN model variants (with Settings A and B) were built with common parameters listed in the Table 1. The difference between the two models was the reliance upon spatial dropout (Setting A: ON with 50% threshold, Setting B: OFF). Spatial dropout refers to disconnecting connections between neurons within a layer if the given spatial distance is higher than the defined percentage of the maximum distance within that layer.

2.5 Performance Evaluation

Model instances were evaluated on test sets in both settings A and B model variants (table 1). Sensitivity (SNS), specificity (SPC) and balanced accuracy (BACC) were calculated and reported as performance metrics.

3 Interpretability Analysis

Output neuron clusters: the number of output neurons equals the number of label outcomes. The spatial training approach allows to result in spatial configurations where the two output neurons could be very close to each other like we can see in the figure 1. In this case the network structure implies that the given prediction is not task-specific. If however, the two output neurons are far away from each other, as observed with in Figure 2 (where the output neurons are represented by the green spheres), it implies that the outputs depend on different data and that the network shall be task-specific (currently not implemented in DEBI-NNs). In addition, we evaluated the ability to rank input features importance according to their spatial distance (in the network architecture after training) with respect to the ouput neuron position. The distance between each of the 36 input neurons (one for each feature) and the output neuron (HPV status positive) was calculated as the distance of the entire path within the architecture between the two. The shorter this distance, the more important is the weight or contribution of this feature in the network's decision.

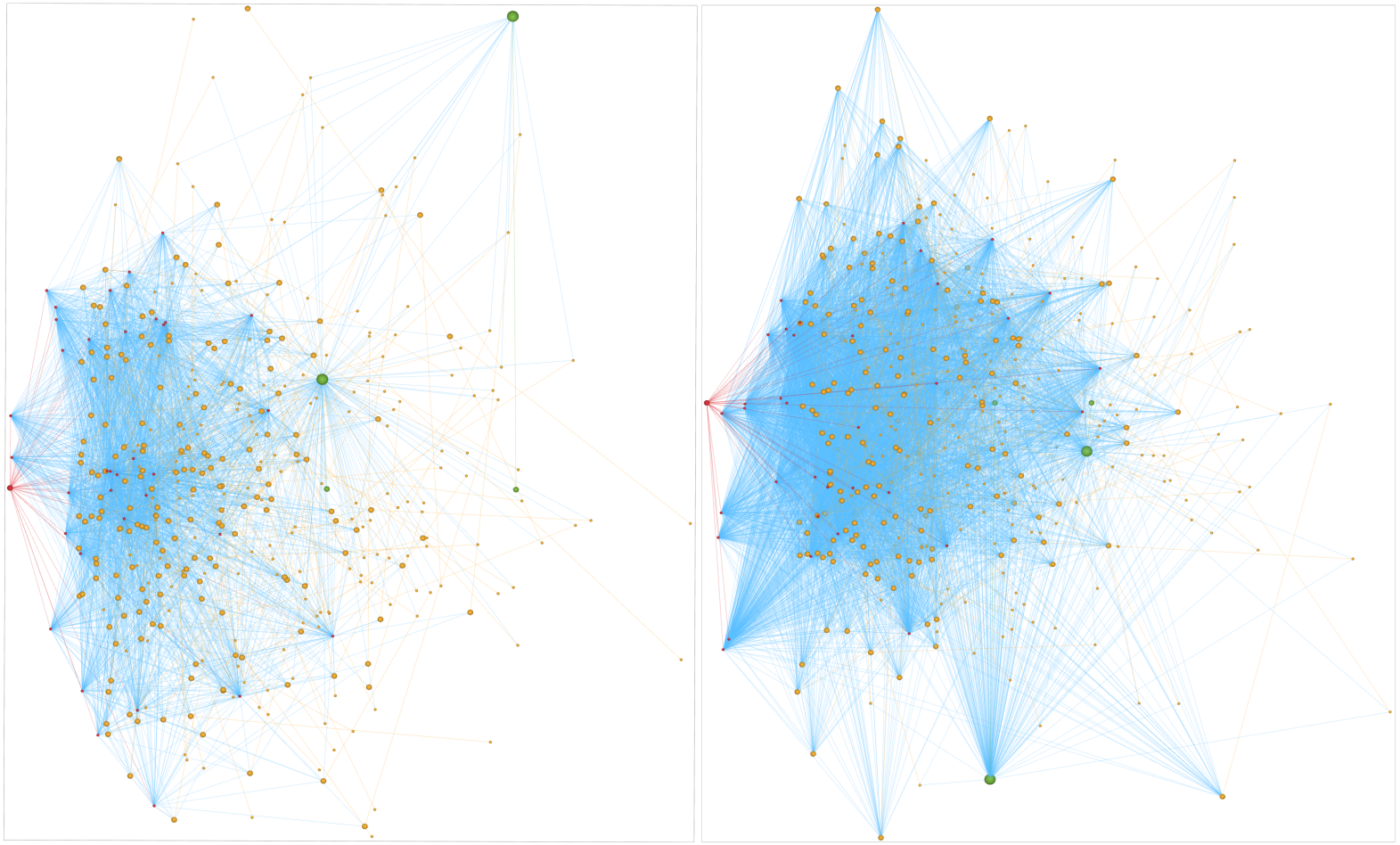
Parameter	Value
Layer delta	-0.7
Hidden neuron count	-180
Hidden layer group count	23
Activation function	LReLU
Shift-Scale optimization	ON
Maximum iteration	1000
Early stopping margin	300
Loss method	Balanced entropy loss
Label smoothing	ON
Loss regularization	OFF
Group normalization	OFF
Weight standardization	OFF

Table 1: DEBI-NN training and model parameters. Layer delta -0.7 denotes that the next layer is pushed back into the previous layer with 70% overlap across the z-axis. Only one hidden layer was utilized. LReLU - Leaky Rectified Linear Unit.

4 Results

Performance evaluation (figure 2) revealed 85% SNS, 95% SPC and 90% BACC for Setting A with spatial dropout. In contrast, Setting B resulted in near-random performance, with very high SNS (94%) at the cost of very low SPC (20%), resulting in BACC of 57%. The interpretability analysis consistently showed a clear clustering of output neurons in both DEBI-NN configurations, where the spatial clustering was more prominent in case of Setting A.

Figure 3 provides the number of connections and the distance metrics for each feature between input and output for identifying HPV+ patients, thereby ranking the importance of each feature in the positive diagnosis. This provides some explainability of the trained model to end-users as well.



(a) Setting A: (50% spatial dropout), 90% test BACC

(b) Setting B: (no spatial dropout), 57% test BACC

Figure 2: DEBI-NN network configurations (Settings A and B) with their 3D rendering views and their test Balanced Accuracy (BACC) values. Green spheres denote the output neurons that are distinct in both configurations, indicating the attempt by the DEBI-NN learner to evolve towards a task-specific network. Setting A allows to approximate such a network.

5 Discussion

DEBI-NNs have the potential to result in high-performing predictive models, especially considering the low sample count and the high imbalance ratio typically encountered in medical imaging applications, as it is the case with HECKTOR. Our visual rendering analysis implies that the HPV classification task should be handled with a task-specific network configuration and not with one general classification network scheme. Since DEBI-NN output neurons are individual per prediction outcome, their spatial distribution across the two network configurations demonstrated a visible spatial distinction. This is a counteraction of the network training process, which attempts to end up with a spatial configuration individually handling both label outcomes. While the hidden layers of the current DEBI-NN implementation do not explicitly support task-specific sub-network clusters, activating spatial dropout does allow to approximate such behavior, where distinct connections - and thereby clusters - can be detached during the training process. Our initial findings help in understanding the nature of the given input data while relying on DEBI-NN 3D visualization, and in alignment with findings reported in the DEBI-NN paper [1], further investigations are required regarding the ability of DEBI-NNs to model task-specific networks when spatial dropout is utilized. Figure 3 shows how the distance between input and output can make the model less of a black box and understand the basis on which the model makes predictions. In the same figure, we observe that the number of connections is lower for the HPV+ output neuron compared to the HPV-, and the cause can be related either to the nature of the data (class imbalance or one class with more distinctive features than the other) or to the architecture itself, which gives an advantage to one class over the other because of the limitation of the dropout to mimic a task specific behavior or the process of training itself. This aspect of the research is still ongoing.

6 Conclusions

Our work show how the DEBI-NN's inherently 3D architecture allows interactive visualization and enhances understanding of its network consistency, as well as helping users in identifying optimal network configurations for the corresponding tasks. This visualization can assist in subsequent

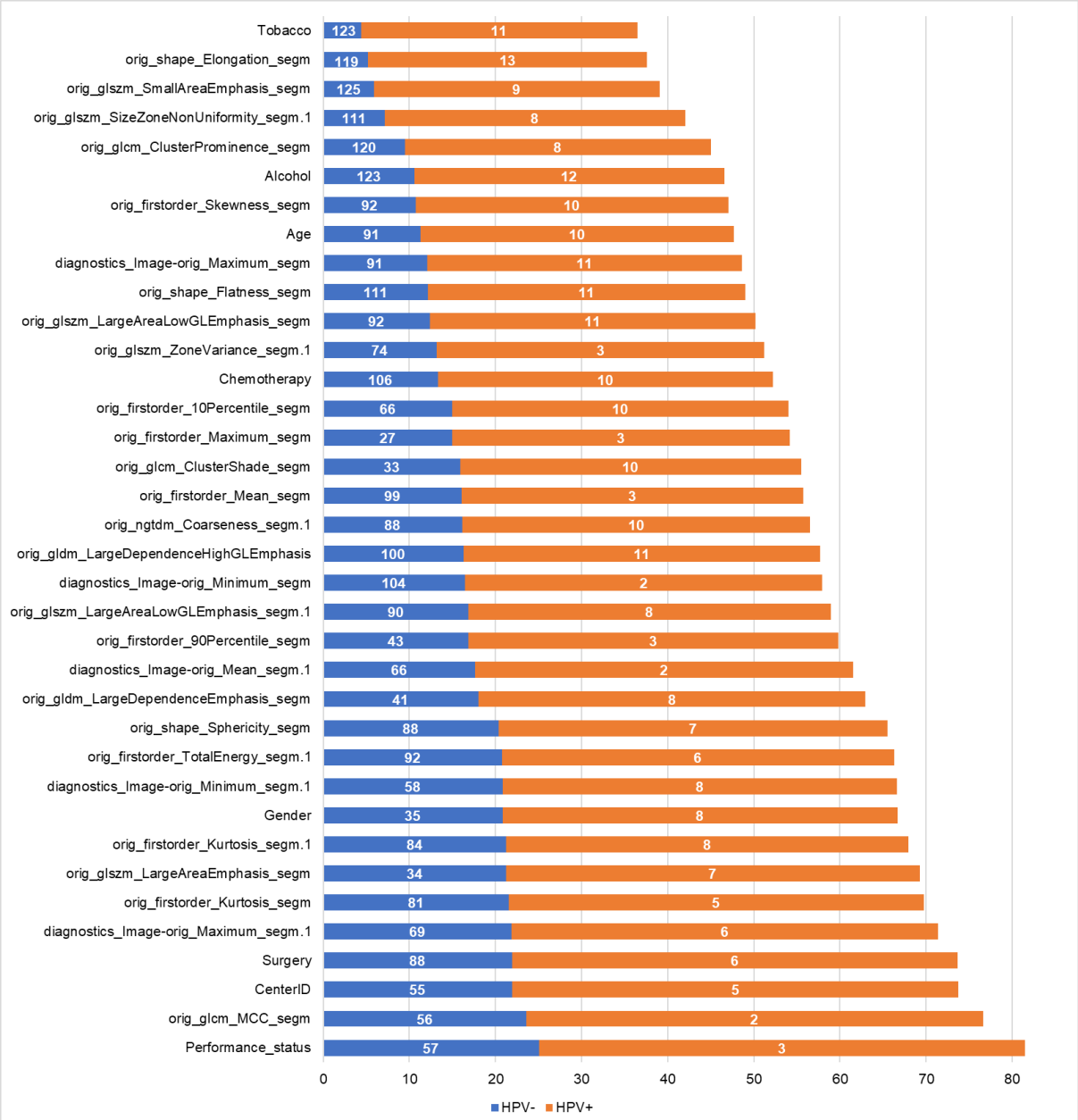


Figure 3: Setting A: Shortest path lengths (bars) and number of connections (numbers in bars) per input feature

network implementation processes, particularly for handling task-specific clinical scenarios. In addition to the contribution of input data ranking to interpretability, our findings underscore the potential of novel DEBI-NN training schemes combined with visualization.

7 Acknowledgments

The present research was partly funded by the City of Brest (Brest Métropole Océane), by La Ligue Contre le Cancer, the CHIST-ERA grant [CHIST-ERA-19-XAI-007] with project acronym INFORM, General Secretariat for Research and Innovation (GSRI) of Greece, National Science Centre (NCN) of Poland [2020/02/Y/ST6/00071] and Agence Nationale de la Recherche (ANR) of France.

References

[1] L. Papp, D. Haberl, B. Ecsedi, et al. “DEBI-NN: Distance-encoding biomorphic-informational neural networks for minimizing the number of trainable parameters”. *Neural Networks* 167 (2023), pp. 517–532.

[2] V. Andrearczyk, V. Oreiller, M. Abobakr, et al. “Overview of the HECKTOR Challenge at MICCAI 2022: Automatic Head and Neck Tumor Segmentation and Outcome Prediction in PET/CT”. *Head and Neck Tumor Segmentation and Outcome Prediction*. Ed. by V. Andrearczyk, V. Oreiller, M. Hatt, et al. Cham: Springer Nature Switzerland, 2023, pp. 1–30.

[3] C. Ding and H. Peng. “Minimum redundancy feature selection from microarray gene expression data”. *Journal of bioinformatics and computational biology* 3.02 (2005), pp. 185–205.

Robustness In Image-Based Data Mining: Bias In Training Data Can Affect The Identification Of Significant Regions

Tanwiwat Jaikuna[1,2], Fiona Wilson[1], Marcel van Herk[1], Peter Hoskin[1], Catharine M L West[1], the REQUITE Consortium, Marianne Aznar[1], and Eliana Vasquez Osorio[1]

[1]Division of Cancer Sciences, School of Medical Sciences, Faculty of Biology, Medicine and Health, The University of Manchester, Christie NHS Foundation Trust Hospital, Manchester, United Kingdom

[2]Division of Radiation Oncology, Department of Radiology, Faculty of Medicine Siriraj Hospital, Mahidol University, Bangkok, Thailand

Abstract Image-based data mining (IBDM) is a complex technique that requires robust technical and clinical validation to ensure repeatability. This study proposes a new method to assess the robustness of the IBDM-identified region using intra-cohort k-mean and k-fold cross-validation. We investigated the impact of biases in patient characteristics or outcome distribution using the k-mean method and stratified k-fold resampling, with k = 5. IBDM was performed using four groups in each cluster/fold, and the overlap of their significant regions was compared to IBDM for the complete cohort. We found that bias in outcome distribution impacts the volume of identified significant regions substantially, even more than bias in patient characteristics. Assessing the effect of potential biases in the training data can increase confidence in the robustness of the significant region identification in IBDM.

1 Introduction

Image-based data mining (IBDM) is a voxel-based analysis technique that can identify the significant region where 3D radiotherapy (RT) dose deposition is associated with treatment outcomes such as acute and late radiation-induced toxicity. IBDM techniques have previously been applied to various treatment sites and successfully identified the sensitive region related to a range of toxicities, e.g., trismus and dysphagia in head and neck cancer [1, 2]. IBDM has two main steps: *i*) spatial normalisation of 3D RT dose to a common coordinate system (CCS) and *ii*) voxel-based statistical inference [3]. However, IBDM requires a solid validation process to ensure the repeatability, reproducibility, and generalisability of modelling outcomes [4]. This study proposes a new method to investigate the robustness of the identified region in IBDM to different biases using stratified k-fold resampling and k-means techniques.

2 Materials and Methods

Radiotherapy data of 912 breast cancer patients from the REQUITE study (www.requite.eu) were included. All patients were treated in a supine position following breast-conserving surgery. The severity of breast oedema was scored on a four-point scale by clinicians and patients (CRO and PRO) before RT (baseline) and immediately post-RT following the Common Terminology Criteria for Adverse Events (CTCAE) v4.0 and the BR23 questionnaire for quality of life in breast cancer patients developed by the European Organisation for Research and Treatment of Cancer (EORTC). The cumulative burden of oedema was assessed by combining CRO and PRO and calculating the area under the curve starting from baseline to post-RT of oedema (oAUC) using Simpson's rule to characterise the severity pattern.

2.1 Applying IBDM

To bring all patient data into a CCS, 3D dose distribution and Computed tomography (CT) images from left-side breast patients were mirrored, so the complete cohort mimicked a right-sided treatment. 3D dose distributions were converted to equivalent dose in 2 Gy-fraction (EQD_2) using the Linear Quadratic model (α/β =1.7 Gy) to compensate for fractionation effects. We selected the '*most average patient*' as our CCS by considering the distribution of patient characteristics, including breast cup size, breast band, and body mass index (BMI). The 3D EQD_2 distributions were deformed to the selected CCS using deformable image registration (DIR) via NiftyReg. The performance of DIR was evaluated by the normalised correlation coefficient (NCC).

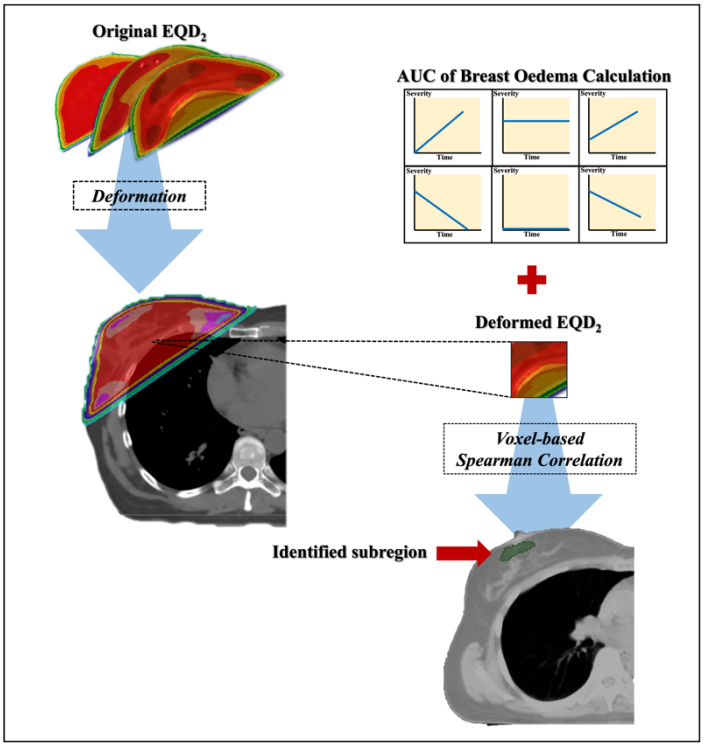


Figure 1. IBDM workflow for breast radiotherapy.

A voxel-based Spearman correlation between the deformed EQD_2 and oAUC post-RT was calculated. The permutation testing (n=1000) was applied to identify the

significant region associated with acute breast oedema (with a significance level corresponding to $p<0.05$). The overview of the IBDM workflow is shown in Figure 1. Regions of interest (ROIs) in the thorax region, including breast, muscle, and lymph nodes, were delineated by an experienced clinician on the CCS. We report the volume, the estimate of mean (EQD_{2mean}) and maximum EQD_2 (EQD_{2max}).

2.2 Intra Cohort Cross-Validation

The complete cohort was divided into five groups using three different methods, Figure 2. We used *i*) the 3D k-mean method, which allowed us to investigate potential biases in patient characteristics, i.e. different distributions of breast volume, BMI, and age, *ii*) the stratified k-fold method by resampling the population while maintaining the normal distribution of patient characteristics within cluster to investigate the impact of biases in outcome (oAUC) distribution, and *iii*) the stratified k-fold method by maintain a similar proportion of outcome and patient characteristics in each cluster (unbiased). To validate the performance of IBDM, four clusters were enrolled for five iterative IBDM model generations (k=5 fold), and one cluster was left out. The identified region overlaps with ROIs in each fold/experiment was segmented. The volume and overlapping regions were compared.

3 Results

DIR results produced NCC = 0.93 ± 0.03 (ideal value is 1), indicating negligible uncertainty in spatial RT dose deformation between the complete cohort and the CCS.

3.1 Identified Sensitive Region for Breast Oedema

IBDM for the complete cohort study generated a significant region overlapping with ROIs sparsely in the central breast region, the major pectoralis muscle (MPM), and axillary lymph node level 1 (ALNL1), Figure 3A-C. The volume of overlap of the significant regions and OARs was 169.03 cm^3, 2.78 cm^3, and 1.41 cm^3 for breast, major pectoralis, and ALNL1, respectively. In addition, the $EQD2_{mean}$ and $EQD2_{max}$ of region intersect with breast were 52.22 ± 4.75 Gy, 60.26 ± 7.94 Gy, 49.31 ± 4.29 Gy, and 52.56 ± 5.85 Gy for major pectoralis, and 35.81 ± 14.99 Gy, 41.30 ± 14.09 for ALNL1.

3.2 The Robustness of IBDM Modelling Outcome

All analyses found a significant region within the breast. However, the size of the identified region was impacted by biases in patient characteristics and outcome distribution, Figure 3D-E. For the unbiased investigation, volume consistency was more consistent within folds and versus the complete cohort (Figure 3F). A significant variation of volume, EQD_2mean and EQD_2max from the identified region was observed when biases in outcome distribution were present, while a low variation was found in unbiased data. The volume and EQD_2 of the region within the breast contour from each experiment are presented in Table 1.

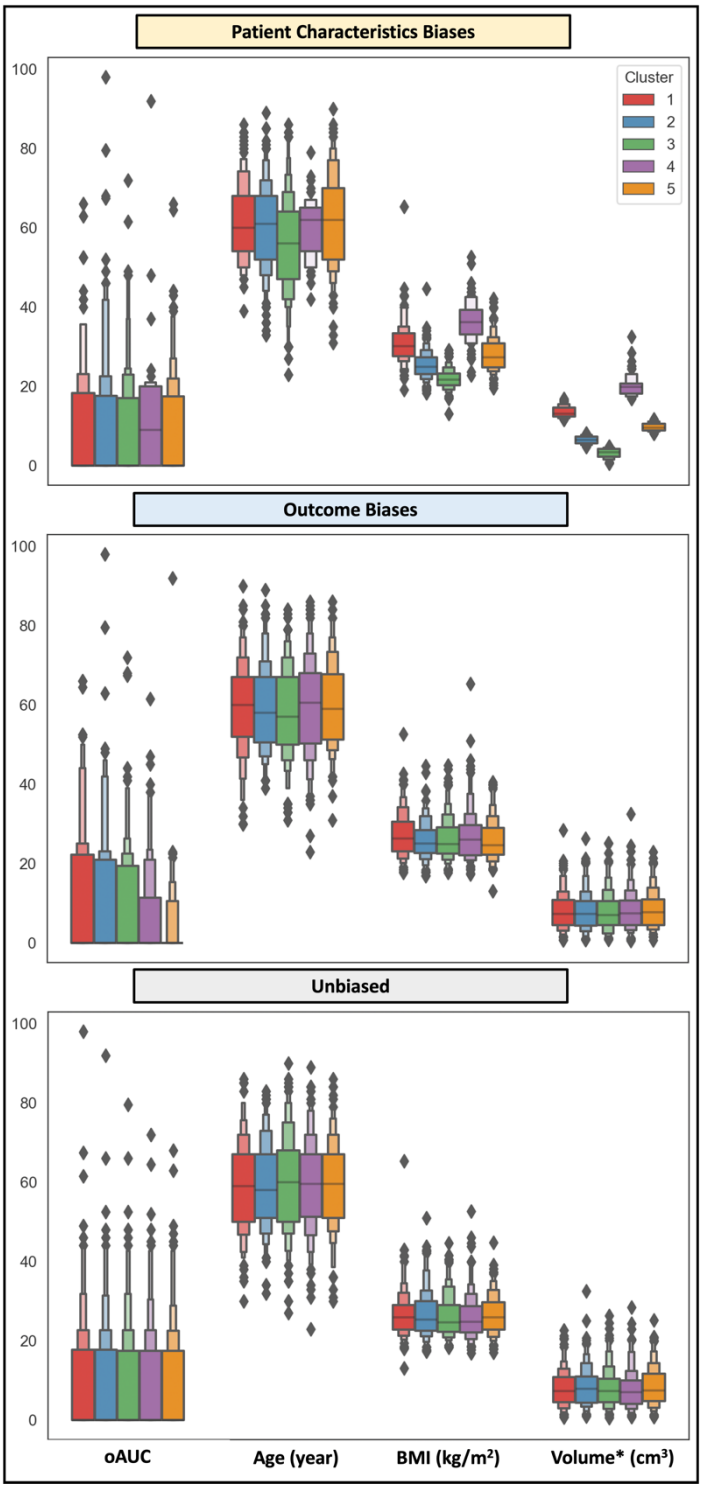


Figure 2. The distribution of patient characteristics and outcome (area under the curve) for oedema for each group and for each resampling method which produced data with biases on patient characteristics (top), on the outcome (mid) plus a non-biased set (bottom)
*The scale of volume is x100 cm^3

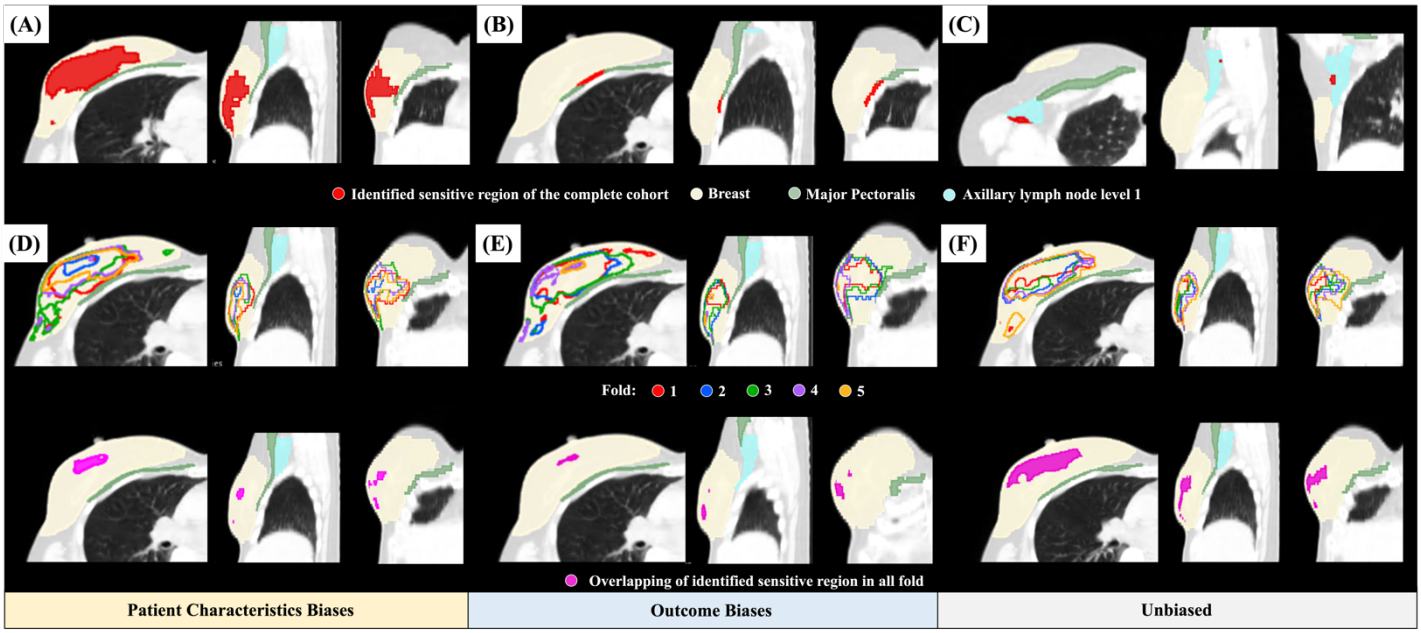


Figure 3. The identified sensitive region (●) intersects with (A) the breast (○), (B) the major pectoralis (◍), and (C) the axillary lymph node level 1 (◎) contour for the complete cohort. The identified sensitive region intersects with breast contour and overlap volume of fivefold of (D) patient characteristics biases, (E) outcome biases, and (F) unbiased investigation.

Table 1. The volume and EQD_2 of the identified subregion intersect with the breast contour from each fold of the difference clustering method.

Cluster	Parameter	Cluster/Fold				
		1	2	3	4	5
Patient Characteristics Biases	Number of patients	127	244	268	57	216
	Volume of subregion (cm^3)	148.55	16.09	181.16	117.17	112.68
	Volume of intersection subregion between each fold and the complete cohort (cm^3)	126.02	16.05	121.98	113.58	103.57
	Volume of intersection subregion from all fold (cm^3)			15.28		
	EQD_2mean (Gy)	52.76±4.64	52.67±5.11	51.62±5.01	52.81±4.71	52.51±4.59
	EQD_2max (Gy)	59.93±7.76	57.51±7.50	60.00±8.11	60.01±7.85	59.66±7.59
Outcome Biases	Number of patients	183	183	182	182	182
	Volume of subregion (cm^3)	109.84	180.5	212.02	49.57	7.32
	Volume of intersection subregion between each fold and the complete cohort (cm^3)	89.88	144.21	156.86	38.16	7.32
	Volume of intersection subregion from all fold (cm^3)			4.83		
	EQD_2mean (Gy)	51.58±4.75	52.40±4.76	52.46±4.64	49.48±5.92	53.06±4.84
	EQD_2max (Gy)	58.22±7.39	59.69±8.21	60.47±8.21	58.75±7.41	57.65±6.79
Unbiased	Number of patients	183	183	182	182	182
	Volume of subregion (cm^3)	81.67	99.63	71.99	148.12	170.82
	Volume of intersection subregion between each fold and the complete cohort (cm^3)	78.56	95.26	70.32	134.86	134.70
	Volume of intersection subregion from all fold (cm^3)			38.41		
	EQD_2mean (Gy)	52.36±4.54	52.44±5.19	53.05±5.18	52.72±4.82	52.45±4.47
	EQD_2max (Gy)	59.46±7.56	59.86±7.83	59.27±7.68	60.12±7.90	59.96±7.89

4 Discussion

Technical and clinical validation are mandatory processes for IBDM study to assess the robustness of the identified region and prediction model. The roadmap of IBDM for clinical implementation suggested validating the IBDM modelling in reproducibility and generalisability by using an internal or external validation cohort [4]. This study showed the impact of input training data biases on patient characteristics and outcomes for internal resampling cross-validation, which affected the size and location of the identified region compared with the unbiased investigation.

Previous studies evaluated the robustness of the identified region by considering the effect of DIR uncertainty in the internal cohort analysis by varying a CCS [2] and clinical repeatability validation of the identified region using external cohorts [5, 6]. However, high-quality external cohort data is not always available. Therefore, in the initial step of clinical validation in spatial dose-response maps, validation using internal resampling is beneficial.

This study is the first to evaluate input biases in training data and their impact on the modelling outcome of IBDM. Our proposed method considers some essential parameters (patient characteristics and outcome), but these might depend on the cohort type and outcome being assessed.

The proposed method can be used to increase the confidence of repeatability for technical and clinical validation of the IBDM pipeline. Investigating the identified region from the patient characteristics biases (skewness and normal distribution) can imply the registration performance of the spatial normalisation process. Moreover, our results showed the possible risk of using the unbalanced cluster, which resulted in substantial differences in subregion size, for internal cohort cross-validation.

However, clinical translation and other cross-validation methods still need to be investigated more and are essential for future research to prove the repeatability and reliability of IBDM modelling outcomes. In addition, the percentage of outcomes for the training and testing dataset still needs a threshold for the IBDM study.

5 Conclusion

In our cohort, biases in outcome distribution for the input training dataset strongly impacted IBDM-sensitive region identification more than biases in patient characteristics. Applying our proposed method to test the impact of training data biases in clinically spatial dose-response maps can provide internal validation and increase confidence in the robustness of region identification from IBDM prior to model development.

References

[1] Beasley W, Thor M, McWilliam A, et al. Image-based Data Mining to Probe Dosimetric Correlates of Radiation-induced Trismus. Int J Radiat Oncol Biol Phys. 2018;102(4):1330-1338. doi:10.1016/j.ijrobp.2018.05.054

[2] Vasquez Osorio E, Abravan A, Green A, et al. Dysphagia at 1 Year is Associated With Mean Dose to the Inferior Section of the Brain Stem. Int J Radiat Oncol Biol Phys. 2023;117(4):903-913. doi:10.1016/j.ijrobp.2023.06.004

[3] Palma G, Monti S, Cella L. Voxel-based analysis in radiation oncology: A methodological cookbook. Phys Med. 2020;69:192-204. doi:10.1016/j.ejmp.2019.12.013

[4] McWilliam A, Palma G, Abravan A, et al. Voxel-based analysis: Roadmap for clinical translation. Radiother Oncol. 2023;188:109868. doi:10.1016/j.radonc.2023.109868

[5] McWilliam A, Abravan A, Banfill K, Faivre-Finn C, van Herk M. Demystifying the Results of RTOG 0617: Identification of Dose Sensitive Cardiac Subregions Associated With Overall Survival. J Thorac Oncol. 2023;18(5):599-607. doi:10.1016/j.jtho.2023.01.085

[6] Craddock M, Nestle U, Koenig J, et al. Cardiac Function Modifies the Impact of Heart Base Dose on Survival: A Voxel-Wise Analysis of Patients With Lung Cancer From the PET-Plan Trial. J Thorac Oncol. 2023;18(1):57-66. doi:10.1016/j.jtho.2022.09.004

Modelling occult lymph node metastases in HNSCC patients with a trinary state hidden Markov model

Yoel Pérez Haas[1], Roman Ludwig[1], Esmée Lauren Looman[1], Vincent Grégoire[2], Laurence Bauwens[2], Panagiotis Balermpas[1], and Jan Unkelbach[1]

[1]Department of Radiation Oncology, University Hospital of Zurich, Zurich, Switzerland
[1]Centre Léon Bérard, Radiation Oncology, Lyon, France

Abstract Head and Neck squamous cell carcinomas often spread through the neck's lymphatic system. Apart from clinically detected lymph node metastases, treatment includes elective irradiation of lymph node levels (LNL) at risk of harbouring occult metastases. We previously developed models for personalized risk of occult metastases predictions. These models relied on literature values for sensitivity and specificity of MRI, CT and PET for detecting metastases, ranging between 70-90% averaged over all LNLs. However, recent data for oropharyngeal SCC indicate that the effective sensitivity varies between LNLs. Whereas LNL II conforms with the literature values, lower sensitivities are observed for LNL III (around 50%) and IV (around 30%). We developed a Hidden Markov Model of lymphatic tumor progression, which describes the difference between pathological and clinical lymph node involvement via growth of metastases from a microscopic to a macroscopic state, and thereby addresses the limitations originating from universal sensitivity and specificity values.

1 Introduction

Head and Neck Squamous Cell Carcinomas (HNSCC) spread through the lymphatic system [1]. Consequently, lymph nodes in the lymph drainage often harbour metastases that need to be treated. To increase the probability of controlling all metastases, lymph nodes levels (LNLs) without clinical evidence of involvement but at risk of occult metastases are irradiated prophylactically as part of the elective clinical target volume (CTV-N).

Current clinical guidelines for the definition of elective CTV-N are primarily based on observed prevalences of metastases in the LNLs [2, 3] rather than a conditional probability of occult disease, given that the LNL of interest is clinically negative. To address this problem, Ludwig et al. proposed a Hidden Markov Model (HMM) for lymphatic tumor progression in HNSCC patients [4, 5]. The model learns the probabilities of tumor spread from datasets of lymph node involvement. The probabilities of spread are combined with literature values of sensitivity and specificity for clinically detecting lymph node metastases, leading to a personalized prediction of occult LNL involvement.

Clinical assessment of LNLs involves various imaging modalities including MRI, PET, and CT. Pathology after neck dissection is the gold standard for assessing lymph node involvement. The difference between clinical and pathological involvement is often described by the sensitivity and specificity of clinical assessment, assuming pathology to be the ground truth. A meta analysis by Bondt et al. [6] proposed sensitivity and specificity values for different imaging modalities in the order of 80%, i.e. a probability of 20% to miss occult metastases and a probability of 20% to incorrectly assume suspicious nodes to be metastatic. A comparison with data of orpharyngeal cancer patients collected at the Centre Léon Bérard (Lyon, FR) ([7, 8]), where both pathological and clinical data was available, showed varying sensitivity between LNLs. Whereas LNL II conforms with the literature values, lower sensitivities are observed for LNL III (around 50%) and IV (around 30%). In this work we extend the previously developed HMM of lymphatic tumor progression by Ludwig et al. [4, 5]. We describe the difference between pathological and clinical lymph node involvement via the growth of metastases from a microscopic to a macroscopic state, and thereby address the limitations originating from universal sensitivity and specificity values.

2 Materials and Methods

2.1 Hidden Markov Model with Trinary Variables

The HMM initially proposed by Ludwig et al. [4] used binary variables to describe the state of each LNL: healthy (0) and involved (1). In the model extension developed here, each LNL $v \in 1,2,...,V$ is represented by a trinary random variable X_v: healthy (0), microscopically involved (1), and macroscopically involved (2). The entire state of a patient is encapsulated in a random vector $\mathbf{X} = [X_1, X_2, ..., X_V]$. Lymphatic metastatic progression is described by the transition matrix $\mathbf{A}$ of the HMM, whose elements A_{ij} encompass the transition probabilities from state $\mathbf{X}[t] = \xi_i$ at time step t to a subsequent state $\mathbf{X}[t+1] = \xi_j$ at $t+1$. $\mathbf{A}$ is thus a $3^V \times 3^V$ dimensional matrix. Transition probabilities are parameterized by the graphical model illustrated in Figure 1 containing three distinct types of parameters:

- The base probabilities b_v describe the probability of direct spread from the primary tumor to a LNL v;
- The transmission probabilities $t_{v \to v+1}$ describe the spread from LNL v to a downstream LNL $v+1$;
- The growth parameter g models the probability of a LNL to transition from a microscopically to a macroscopically involved state.

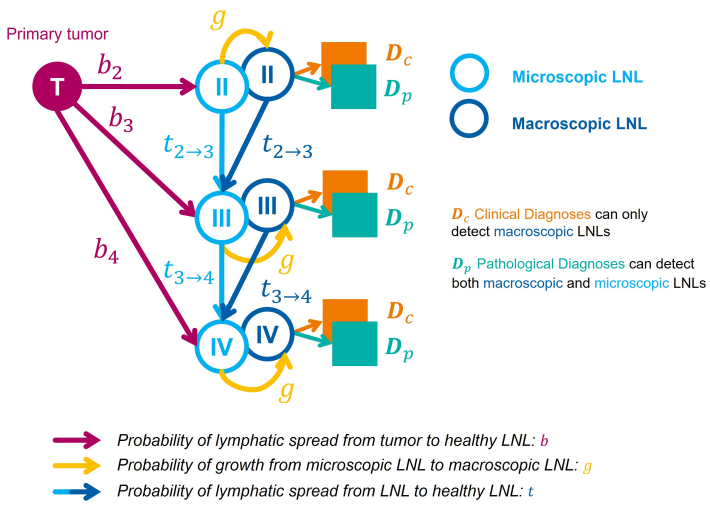


Figure 1: Graph consisting of a primary tumor T and LNLs II, III and IV. The LNLs are connected by arcs modelling the spread to downstream LNLs that receive efferent lymphatics from an involved upstream level.

To illustrate the entries in the transition matrix, we consider a model containing the three ipsilateral LNLs II, III and IV described by the state $[X_2, X_3, X_4]$. As an example, let us consider the state $[1,0,0]$ where LNL II is microscopically involved, and LNLs III and IV are healthy. The probability to transition to state $[2,1,0]$ is:

$$P(\mathbf{X}[t+1] = [2,1,0] | \mathbf{X}[t] = [1,0,0]) = g \cdot (1-(1-b_3)(1-t_{2\to3})) \cdot (1-b_4) \quad (1)$$

Equation 1 consists of a multiplication of transition probabilities for each LNL. The first term (g) is the probability of LNL II to grow from microscopic to macroscopic; the second term $(1-(1-b_3)(1-t_{2\to3}))$ is the probability for LNL III getting microscopically involved due to spread from the primary tumor or LNL II; the third term $(1-b_4)$ is the probability of LNL IV staying healthy. Generally, a LNL may stay in the same state, or its state increases by one. Thus, self-healing and skipping microscopic involvement is not allowed.

The HMM models the state of lymph node involvement as a dynamic system. Let us consider the time evolution of a state probability vector whose i-th element contains the probabilities of being in states ξ_i at time t. We assume that each patient starts with probability one in the healthy state where $X_v = 0$ for all LNLs, which we denote by a vector π which is one for the healthy state and zero elsewhere. Evolving the system from one time step to the next is equivalent to multiplying the current state probability vector with the transition matrix $\mathbf{A}$. Thus, the probability of being in state ξ_i at time t is the i-th element of the vector $\pi \cdot \mathbf{A}^t$. To obtain the probability over the states at diagnosis, we have to marginalize over the unknown diagnose time

$$P(\mathbf{X} = \xi_i | \theta) = \sum_{t=0}^{t_{max}} P_T(t) (\pi \cdot \mathbf{A}^\mathbf{t})_i \quad (2)$$

where $P_T(t)$ is a prior distribution over the diagnose time, which we model by a Binomial distribution. This marginalization enables us to model higher prevalence of LNL involvement for more advanced T-stage via a T-stage dependent prior $P_T(t)$. Early T-stage tumors are assumed to be diagnosed earlier, i.e. in earlier time steps, while late T-stage tumors have a probability distribution which gives more weight to later time steps (see [4] for details).

2.2 Training Data

To train the model, a dataset including both clinical and pathological information for the relevant LNLs is required. The publication by Bauwens et al. [7] reported clinical and pathological involvement for 263 oropharyngeal SCC patients. The dataset is publicly available [8] and can be viewed through the web-based graphical interface at www.LyProx.org [9]. 259 of these patients have a complete diagnosis for ipsilateral LNLs II, III and IV and were included in the training and evaluation dataset of the model.
Clinical and pathological diagnosis are modeled as binary variables $\mathbf{Z}^c$ and $\mathbf{Z}^p$ and are linked to the states $\mathbf{X}$ as follows

- Pathological diagnosis $\mathbf{Z}^p$ detects microscopic and macroscopic involvement, but does not discriminate between them;
- Clinical diagnosis $\mathbf{Z}^c$ detects macroscopic involvement, but but does not discriminate microscopic involvement from healthy.

state	early		late	
[0,0,0]	48	(27.9%)	16	(18.4%)
[0,1,0]	9	(5.2%)	2	(2.3%)
[0,2,0]	2	(1.2%)	2	(2.3%)
[1,0,0]	12	(7.0%)	5	(5.8%)
[1,1,0]	3	(1.7%)	2	(2.3%)
[2,0,0]	65	(37.8%)	34	(39.1%)
[2,1,0]	7	(4.1%)	10	(11.5%)
[2,1,1]	3	(1.7%)	1	(1.2%)
[2,2,0]	13	(7.6%)	9	(10.3%)
[2,2,1]	3	(1.7%)	1	(1.2%)
[2,2,2]	3	(1.7%)	1	(1.2%)

Table 1: Prevalence of most common states of involvement of ipsilateral LNLs II, III, and IV, $[X_2, X_3, X_4]$, for early (T1&T2) and late (T3&T4) T-stages in absolute number of patients and percentages. Only states with more than one patient are listed. False positive clinical diagnoses are considered healthy.

Thus, a LNL that is positive at pathology but clinically negative is in state $X_v = 1$, a LNL that is positive on both modalities is in state $X_v = 2$, and a LNL that is negative at pathology is in the healthy state $X_v = 0$ (i.e. negative pathology overwrites a (false) positive clinical diagnosis). The prevalence of most common involvements are listed in table 1.

Training our model corresponds to learning the model parameters θ. This is achieved through Likelihood inference. Assuming a uniform prior in the interval $[0,1]$ for all parameters, the posterior distribution of the parameters is proportional to the Likelihood of the data. We employ a Markov chain Monte Carlo (MCMC) algorithm, as detailed in Ludwig et al. [4] to sample model parameters from the posterior.

2.3 Estimating Risks of Occult Metastases

Pathological diagnosis is only available for surgically treated patients after neck dissection, but not for CTV-N definition in definitive radiotherapy patients. With a set of inferred model parameters θ, the HMM can be used to predict risks of occult metastases, given a clinical diagnosis $\mathbf{Z}^c$. Through Bayes' theorem we can compute the probability of a specific involvement $\mathbf{X} = \xi_k$ given a diagnosis $\mathbf{Z}^c = \zeta_l$:

$$P(\mathbf{X} = \xi_k | \mathbf{Z}^c = \zeta_l) = \frac{P(\mathbf{Z}^c = \zeta_l | \mathbf{X} = \xi_k) P(\mathbf{X} = \xi_k | \theta)}{\sum_{r=1}^{3^V} P(\mathbf{Z}^c = \zeta_l | \mathbf{X} = \xi_r) P(\mathbf{X} = \xi_r | \theta)} \tag{3}$$

where the conditional probability of a clinical diagnose $\mathbf{Z}^c$ given the true state $\mathbf{X}$ is parameterized by sensitivity and specificity. Macroscopic involvement is (by definition) detected with sensitivity one. However, we have to consider false positive diagnosis, i.e. the healthy state $X_v = 0$ is diagnosed as clinically positive with probability (1 - specificity). Based on the dataset from the Bauwens et. al. [7], we set the clinical specificity to 0.94.

To illustrate the formalism of risk calculations, we consider a clinical diagnose for a patient with an early T-stage tumor where only LNL II was clinically diagnosed as metastatic, $\mathbf{Z}_c = [1,0,0]$, corresponding to the results discussed later in Figure 2b. The diagnosis is consistent with only 12 out of 27 states:

$$\begin{aligned}&[2,1,0],[2,1,1],[1,1,0],[1,1,1],[0,1,0],[0,1,1],\\&[2,0,0],[2,0,1],[1,0,0],[1,0,1],[0,0,0],[0,0,1]\end{aligned} \tag{4}$$

The remaining 15 states have probability zero. Equation (3) provides the respective probabilities for each state. To compute the probability of occult metastases in LNL III, we sum up the probabilities for all states where LNL III is microscopically involved (corresponding to the states in the first row in equation (4)).

3 Results

3.1 Risks of Occult Metastases

With the sampled model parameters we can compute the risks of occult involvement for each LNL. In figure 2, risks of occult involvement is plotted for the most common diagnosis and compared to the observed prevalence. The distributions describe the uncertainty in the model prediction (due to uncertainty in the model parameters) and the data (due to the limited number of patients). In figure 2a, N0 patients without clinically diagnosed metastases and early T-stage (T1&T2) are shown. LNL III has an involvement risk around 17%, LNL IV a risk around 3%. In figure 2b, LNL II was diagnosed as clinically positive, corresponding to the example discussed in the methods section. The risk for LNL III increases to approximately 22%, which is dominated by the state $[2,1,0]$ which contributes with a risk of 18.0%, followed by the state $[2,1,1]$ with a probability of 2.6%. The remaining states (corresponding to a false positive diagnose in LNL II) are unlikely. The risk for LNL IV is approximately 5%. In figure 2c, both LNLs II and III are diagnosed as involved and the tumor is of advanced T-stage (T3&T4), which increases the risk of LNL IV involvement to approximately 20%.

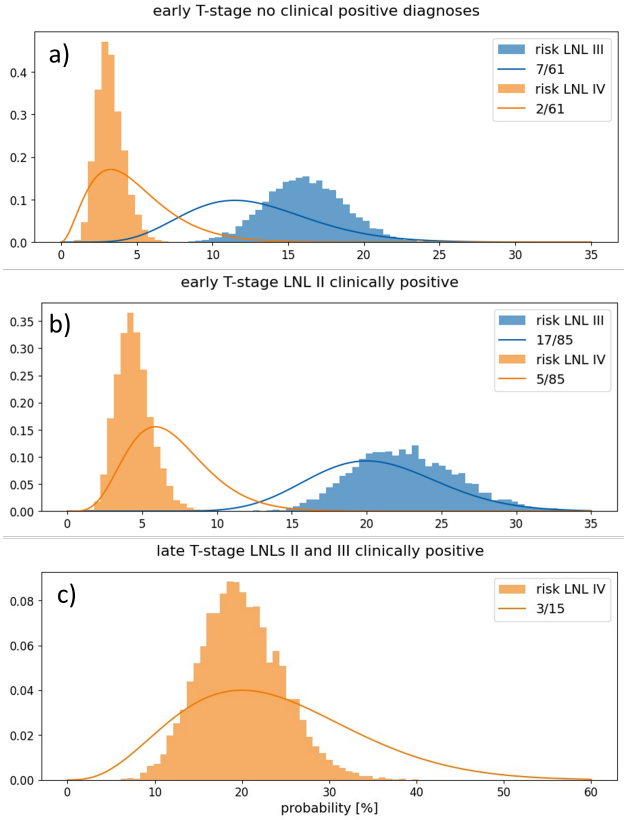


Figure 2: Comparison of predicted risk of involvement (shaded histograms) vs the Beta posterior of the observed prevalence (solid line) for different clinical diagnosis and T-categories. The legend indicate the number of patients with occult involvement in the respective LNL divided by the number of patients matching the clinical diagnose.

3.2 Effective sensitivity of clinical diagnoses

The probability of microscopic and macroscopic involvement obtained from the trinary state HMM allow the computation of an "effective sensitivity" for each LNL to clinically detect metastases. Equation (2) yields the probability of each state. The total probability of occult disease in LNL v is then obtained by summing the probabilities of all states where $X_v = 1$, and the total probability for macroscopic involvement is obtained by summing over all states where $X_v = 2$. The effective sensitivity is then given by

$$\text{sens}(X_v) = \frac{P(X_v = 2|\theta)}{P(X_v = 1|\theta) + P(X_v = 2|\theta)} \tag{5}$$

In figure 3, the effective sensitivities for LNL III and IV are presented. The dataset contains 75 patients with metastases in LNL III out of which 35 were clinically detected, corresponding to an observed sensitivity of 48% (analogously 29% for LNL IV). The trinary state HMM explains most of the difference in the effective sensitivity.

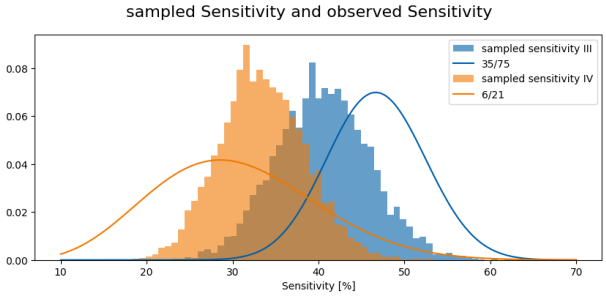


Figure 3: Comparison of predicted sensitivities (shaded histograms) with the Beta posterior of the observed sensitivities (solid lines) in the dataset of Bauwens et al. [7].

4 Discussion

We present a statistical model to estimate the risk of occult metastases in a LNL using a HMM with a trinary state. The model parameters describing the probability of the tumor to spread to a LNL and the probability of clinically detecting metastases are learned from a dataset of 259 oropharyngeal SCC patients in whom both clinical involvement and pathological involvement after neck dissection was reported. The results shown in Figure 2 demonstrate agreement between predicted and observed risks of occult LNL involvement for the most common clinical diagnoses.

Typically, the difference between true LNL involvement and clinically detected involvement is described by sensitivity and specificity of the diagnostic procedure, a concept that is widespread in clinical research and which has been used in previous works on estimating risk of occult disease [4]. However, occult lymph node metastases are not randomly distributed across LNLs. A typical situation is that the tumor first spreads to LNL II, and while the metastases in LNL II grows to a clinically detectable size, the tumor may further spread to LNL III, but the metastasis in LNL III may still be occult when the patient is diagnosed. Therefore, the fraction of occult metastases is larger in more rarely involved LNLs, corresponding to a lower apparent sensitivity. The trinary state HMM addresses this problem by describing the difference between microscopic and macroscopic involvement via the growth parameter g, the probability for a metastasis to grow from occult to clinically detectable. This better reflects the nature of lymphatic metastatic spread compared to the classic model of sensitivity and specificity, and it describes the observation that the fraction of occult metastases is larger in more rarely involved LNLs.

Here, the model is demonstrated for ipsilateral LNLs II, III, and IV. These levels are resected for almost all surgically treated patients, allowing for a comparison of the model's predictions to the prevalence of occult metastases in the data. The methodology has been extended to all relevant LNLs. This requires model training based on incomplete pathology data since not all LNLs are resected for all patients, which is not discussed in this short paper.

In the future, the model may inform clinical trials on volume deescalated radiotherapy for HNSCC and eventually contribute to further personalized guidelines of the definition of the elective nodal CTV-N. The results in figure 2 suggest including LNL III into the CTV-N in all cases, given its involvement probability exceeding 15%. Conversely, for patients without clinical of LNL III, one may consider excluding LNL IV from the CTV-N.

5 Acknowledgements

This work was supported by the clinical research priority program (CRPP) "Artificial intelligence in oncological imaging" of the University of Zurich and by the Swiss cancer research foundation under grant KFS 5645–08–2022.

References

[1] S. K. Mukherji, D. Armao, and V. M. Joshi. "Cervical nodal metastases in squamous cell carcinoma of the head and neck: What to expect". *Head & Neck* 23 (11 Nov. 2001), pp. 995–1005. DOI: 10.1002/hed.1144.

[2] V. Grégoire, E. Coche, G. Cosnard, et al. "Selection and delineation of lymph node target volumes in head and neck conformal radiotherapy. Proposal for standardizing terminology and procedure based on the surgical experience". *Radiotherapy and Oncology* 56 (2 Aug. 2000), pp. 135–150. DOI: 10.1016/S0167-8140(00)00202-4.

[3] J. Biau, M. Lapeyre, I. Troussier, et al. "Selection of lymph node target volumes for definitive head and neck radiation therapy: a 2019 Update". *Radiotherapy and Oncology* 134 (May 2019), pp. 1–9. DOI: 10.1016/j.radonc.2019.01.018.

[4] R. Ludwig, B. Pouymayou, P. Balermpas, et al. "A hidden Markov model for lymphatic tumor progression in the head and neck". *Scientific Reports* 11 (1 June 2021), p. 12261. DOI: 10.1038/s41598-021-91544-1.

[5] R. Ludwig. "Modelling Lymphatic Metastatic Progression in Head and Neck Cancer". PhD thesis. University of Zurich, 2023.

[6] R. de Bondt, P. Nelemans, P. Hofman, et al. "Detection of lymph node metastases in head and neck cancer: A meta-analysis comparing US, USgFNAC, CT and MR imaging". *European Journal of Radiology* 64 (2 Nov. 2007), pp. 266–272. DOI: 10.1016/j.ejrad.2007.02.037.

[7] L. Bauwens, A. Baltres, D.-J. Fiani, et al. "Prevalence and distribution of cervical lymph node metastases in HPV-positive and HPV-negative oropharyngeal squamous cell carcinoma". *Radiotherapy and Oncology* 157 (Apr. 2021), pp. 122–129. DOI: 10.1016/j.radonc.2021.01.028.

[8] R. Ludwig, A. Schubert, D. Barbatei, et al. "A multi-centric dataset on patient-individual pathological lymph node involvement in head and neck squamous cell carcinoma". *Data in Brief* 52 (Feb. 2024), p. 110020. DOI: 10.1016/j.dib.2023.110020.

[9] R. Ludwig, J.-M. Hoffmann, B. Pouymayou, et al. "Detailed patient-individual reporting of lymph node involvement in oropharyngeal squamous cell carcinoma with an online interface". *Radiotherapy and Oncology* 169 (Apr. 2022), pp. 1–7. DOI: 10.1016/j.radonc.2022.01.035.