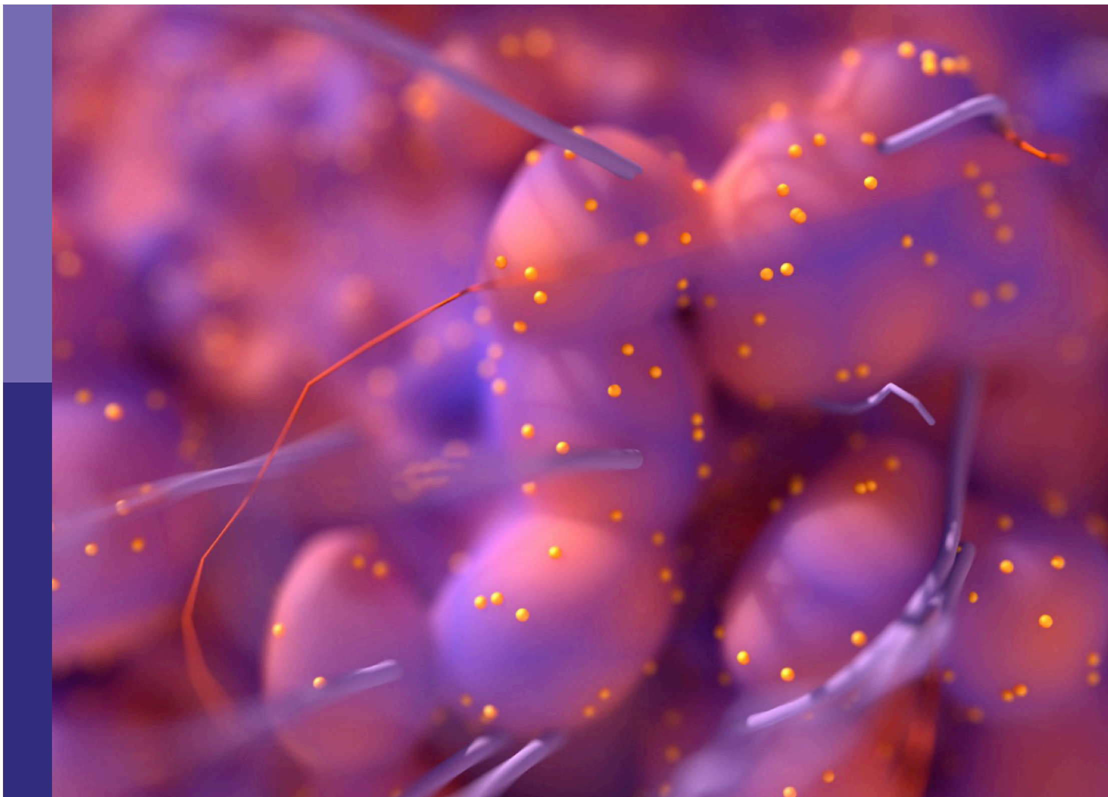
frontiers | Research Topics

# Machine learning-based adaptive radiotherapy treatments: From bench top to bedside

**Edited by**
Jiahan Zhang, Yang Sheng, Justin Roper and Xiaofeng Yang

**Published in**
Frontiers in Oncology



ISSN 1664-8714
ISBN 978-2-8325-2331-5
DOI 10.3389/978-2-8325-2331-5

## About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

## Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

## Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

## What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

# Machine learning-based adaptive radiotherapy treatments: From bench top to bedside

**Topic editors**
Jiahan Zhang — Emory University, United States
Yang Sheng — Duke University Medical Center, United States
Justin Roper — Emory University, United States
Xiaofeng Yang — Emory University, United States

**Citation**
Zhang, J., Sheng, Y., Roper, J., Yang, X., eds. (2023). *Machine learning-based adaptive radiotherapy treatments: From bench top to bedside*.
Lausanne: Frontiers Media SA. doi: 10.3389/978-2-8325-2331-5

# Table of contents

04 Editorial: Machine learning-based adaptive radiotherapy treatments: From bench top to bedside
Jiahan Zhang, Yang Sheng, Justin Roper and Xiaofeng Yang

07 A Comparison Study Between CNN-Based Deformed Planning CT and CycleGAN-Based Synthetic CT Methods for Improving iCBCT Image Quality
Bo Yang, Yankui Chang, Yongguang Liang, Zhiqun Wang, Xi Pei, Xie George Xu and Jie Qiu

19 A Feasibility Study of Deep Learning-Based Auto-Segmentation Directly Used in VMAT Planning Design and Optimization for Cervical Cancer
Along Chen, Fei Chen, Xiaofang Li, Yazhi Zhang, Li Chen, Lixin Chen and Jinhan Zhu

29 Impact of Using Unedited CT-Based DIR-Propagated Autocontours on Online ART for Pancreatic SBRT
Alba Magallon-Baro, Maaike T. W. Milder, Patrick V. Granton, Wilhelm den Toom, Joost J. Nuyttens and Mischa S. Hoogeman

40 Deep learning auto-segmentation of cervical skeletal muscle for sarcopenia analysis in patients with head and neck cancer
Mohamed A. Naser, Kareem A. Wahid, Aaron J. Grossberg, Brennan Olson, Rishab Jain, Dina El-Habashy, Cem Dede, Vivian Salama, Moamen Abobakr, Abdallah S. R. Mohamed, Renjie He, Joel Jaskari, Jaakko Sahlsten, Kimmo Kaski and Clifton D. Fuller

51 Adaptive dose painting for prostate cancer
Emil Fredén, David Tilly and Anders Ahnesjö

67 A novel approach for eliminating metal artifacts based on MVCBCT and CycleGAN
Zheng Cao, Xiang Gao, Yankui Chang, Gongfa Liu and Yuanji Pei

79 Automatic segmentation of nasopharyngeal carcinoma on CT images using efficient UNet-2.5D ensemble with semi-supervised pretext task pretraining
Jansen Keith L. Domoguen, Jen-Jen A. Manuel, Johanna Patricia A. Cañal and Prospero C. Naval Jr

94 An attention base U-net for parotid tumor autosegmentation
Xianwu Xia, Jiazhou Wang, Sheng Liang, Fangfang Ye, Min-Ming Tian, Weigang Hu and Leiming Xu

103 Leveraging intelligent optimization for automated, cardiac-sparing accelerated partial breast treatment planning
Joel A. Pogue, Carlos E. Cardenas, Yanan Cao, Richard A. Popple, Michael Soike, Drexell Hunter Boggs, Dennis N. Stanley and Joseph Harms

114 Dosiomics and radiomics to predict pneumonitis after thoracic stereotactic body radiotherapy and immune checkpoint inhibition
Kim Melanie Kraus, Maksym Oreshko, Denise Bernhardt, Stephanie Elisabeth Combs and Jan Caspar Peeken

frontiers | Frontiers in Oncology

TYPE Editorial
PUBLISHED 19 April 2023
DOI 10.3389/fonc.2023.1188788

OPEN ACCESS

EDITED AND REVIEWED BY
Timothy James Kinsella,
Brown University, United States

*CORRESPONDENCE
Jiahan Zhang
jiahan.zhang@emory.edu

SPECIALTY SECTION
This article was submitted to Radiation Oncology, a section of the journal Frontiers in Oncology

RECEIVED 17 March 2023
ACCEPTED 27 March 2023
PUBLISHED 19 April 2023

CITATION
Zhang J, Sheng Y, Roper J and Yang X (2023) Editorial: Machine learning-based adaptive radiotherapy treatments: From bench top to bedside.
*Front. Oncol.* 13:1188788.
doi: 10.3389/fonc.2023.1188788

# Editorial: Machine learning-based adaptive radiotherapy treatments: From bench top to bedside

Jiahan Zhang[1]*, Yang Sheng[2], Justin Roper[1] and Xiaofeng Yang[1]

[1]Department of Radiation Oncology, Emory University, Atlanta, GA, United States, [2]Department of Radiation Oncology, Duke University, Durham, NC, United States

KEYWORDS

**adaptive radiotherapy, machine learning, image segementation, treatment planning, image registration**

Editorial on the Research Topic
Machine learning-based adaptive radiotherapy treatments: From bench top to bedside

## Introduction

Radiation therapy aims to control malignant and less commonly benign diseases while preserving the surrounding healthy tissues. Standard courses of radiation therapy last up to six weeks, during which time anatomical changes are often anticipated due to tumor shrinkage and the day-to-day variations of organ filling and patient positioning. Historically, clinicians have compensated for these variations by adding generous margins around target volumes to prevent a geometric miss but at the expense of increased radiation dose to the healthy tissues. One alternative is adaptive radiotherapy where the patient receives customized treatment based on the "anatomy of the day." This approach reduces the need for large margins by directly accounting for the inter-fraction variations and consequently better spares the healthy tissues. Adaptive radiotherapy has been an active research area for some time and finally has been commercialized and implemented in some radiotherapy clinics, due in large part to machine learning. In this Research Topic "Machine Learning-Based Adaptive Radiotherapy Treatments: From Bench Top to Bedside", machine learning applications in various stages of the adaptive radiotherapy workflow are covered, including image registration, segmentation, treatment planning, and clinical decision support.

## Topics covered in this research topic

AI-driven image segmentation: Naser et al., Domoguen et al., Xia et al.
Treatment-time image processing/correction: Yang et al., Cao et al.
Automated treatment planning: Fredén et al., Pogue et al.
Clinical decision support *via* dosiomics and radiomics: Kraus et al.

Online adaptive planning workflow validation: Chen et al., Magallon-Baro et al.

## Papers included in this research topic

Naser et al. used auto-segmentation to help define the skeletal muscle index (SMI) and calculate sarcopenia, a prognostic factor for head-and-neck cancer (HNC) patients. The auto-segmentation approach substantially improved the efficiency in determining sarcopenia and proved effective in predicting patient survival. The proposed model could potentially assist in clinical decision-making for HNC treatments.

Domoguen et al. designed a deep learning architecture for the task of auto-segmenting nasopharyngeal carcinoma (NPC) target volumes. This model is based on UNet-2.5D and has been enhanced with multi-scale training and semi-supervised pre-training to improve training efficiency. With a small training/validation dataset, the proposed method demonstrated improved performance at NPC target volume segmentations as compared to the current state-of-the-art methods, indicating efficient use of limited data.

Xia et al. developed an attention-based UNet model to auto-segment parotid neoplasms, a relatively rare form of HNC. The authors reported an average Dice similarity coefficient (DSC) of 0.88 for both parotids. The performance of the proposed model was comparable to human observers (3 radiologists).

Yang et al. compared two approaches for enhancing the image quality of cone-beam CT (CBCT) images: 1) deformable registration of planning CT images to CBCT images, and 2) synthetic CT images derived from CBCT images. The authors found that the auto-segmented contours based on synthetic CT images achieved significantly higher DSCs for bladder, rectum, spinal cord, and femoral heads, compared with contours segmented on deformed planning CT images. This study validated the efficacy of synthetic CT images for auto-segmenting pelvic anatomy.

Cao et al. proposed a novel method to eliminate metal artifacts by synthesizing CT from mega-voltage CBCT (MVCBCT). They implemented a cycle-consistent generative adversarial network (CycleGAN) to synthesize metal artifact-free CT images. The process successfully eliminated metal artifacts. Further, Gamma analysis of the dose matrices calculated based on planning CT and synthesized CT confirmed the dose calculation accuracy.

Fredén et al. studied the effects of adaptive radiation therapy on dose painting treatments. The authors compared the tumor control probability (TCP) of the adaptive workflow and the conventional workflow and found that adaptive workflow consistently achieved target coverage, albeit with marginal improvements in the TCP.

Pogue et al. investigated an automated adaptive planning workflow for accelerated partial breast irradiation with an emphasis on cardiac sparing. A commercial adaptive radiotherapy platform, Varian Ethos, was used to test the workflow. Two physicians evaluated the auto-generated plans and found at least 95% of cases were clinically acceptable. Additionally, the auto-generated plans improved cardiac sparing compared with the previous manual plans.

Kraus et al. developed a model to predict radiation-induced pneumonitis using both dosiomic features and radiomic features. With both sets of features, the model achieved an area under the ROC curve (AUC) of 0.79, suggesting that the model could effectively predict pneumonitis before treatment and help guide clinical decision-making for at-risk patients.

Chen et al. studied the feasibility of using auto-segmented contours directly for cervical cancer VMAT planning. They evaluated plan metrics for plans created based on auto segmentations (AS-VMAT) and compared that with manual segmentations (MS-VMAT) results. While for most organs at risk (OARs), the difference between AS-VMAT and MS-VMAT was not significant, MS-VMAT plans achieved better rectum sparing. The study concluded that auto-segmented contours, especially for organs in close proximity to the target volume, need to be examined carefully to ensure plan quality.

Magallon-Baro et al. explored the feasibility of adaptive treatment planning for pancreas stereotactic body radiotherapy (SBRT) with contours deformed from planning CTs onto treatment CBCTs. Two commercial deformable registration methods were tested. Replanning with unedited, deformed contours resulted in slightly worse results due to inaccuracies in contours near the target volumes. However, the automated plans still outperformed the non-adapted plans.

## Conclusions and outlook

Adaptive therapy is in a crucial phase, transitioning from the bench top to the bedside. With the first generation of commercial adaptive treatment machines and solutions already in some radiation oncology centers, there is emerging expertise in the clinical implementation of adaptive planning workflows. This invaluable clinical knowledge from incorporating adaptive therapy into routine clinical practice will undoubtedly encourage related research activities to enhance accuracy and efficiency, which further promotes the clinical implementation of adaptive therapy. In parallel, researchers are making strides in developing advanced adaptive treatment technologies based on information from various imaging modalities. This Article Collection showcases both the practical aspects of clinical applications of AI-driven modern adaptive therapy workflows and cutting-edge technological advancements in this domain. The adaptive radiotherapy treatments that clinicians have long dreamed of are now gradually becoming a clinical reality.

## Author contributions

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

## Funding

Supported by Winship Cancer Institute # IRG-21-137-07 -IRG from the American Cancer Society.

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Oncology

ORIGINAL RESEARCH
published: 30 May 2022
doi: 10.3389/fonc.2022.896795

# A Comparison Study Between CNN-Based Deformed Planning CT and CycleGAN-Based Synthetic CT Methods for Improving iCBCT Image Quality

*Bo Yang*[1†], *Yankui Chang*[2†], *Yongguang Liang*[1], *Zhiqun Wang*[1], *Xi Pei*[2,3], *Xie George Xu*[2,4] *and Jie Qiu*[1*]

*[1] Department of Radiation Oncology, Chinese Academy of Medical Sciences, Peking Union Medical College Hospital, Beijing, China, [2] School of Nuclear Science and Technology, University of Science and Technology of China, Hefei, China, [3] Technology Development Department, Anhui Wisdom Technology Co., Ltd., Hefei, China, [4] Department of Radiation Oncology, First Affiliated Hospital of University of Science and Technology of China, Hefei, China*

OPEN ACCESS

**Edited by:**
*James Chow,*
*University of Toronto, Canada*

**Reviewed by:**
*Ruijie Yang,*
*Peking University Third Hospital, China*
*Wei Zhao,*
*Beihang University, China*

***Correspondence:**
*Jie Qiu*
*qj_ww@139.com*

*[†]These authors have contributed equally to this work*

**Specialty section:**
*This article was submitted to Radiation Oncology, a section of the journal Frontiers in Oncology*

**Received:** *15 March 2022*
**Accepted:** *27 April 2022*
**Published:** *30 May 2022*

**Citation:**
*Yang B, Chang Y, Liang Y, Wang Z, Pei X, Xu XG and Qiu J (2022) A Comparison Study Between CNN-Based Deformed Planning CT and CycleGAN-Based Synthetic CT Methods for Improving iCBCT Image Quality. Front. Oncol. 12:896795. doi: 10.3389/fonc.2022.896795*

**Purpose:** The aim of this study is to compare two methods for improving the image quality of the Varian Halcyon cone-beam CT (iCBCT) system through the deformed planning CT (dpCT) based on the convolutional neural network (CNN) and the synthetic CT (sCT) generation based on the cycle-consistent generative adversarial network (CycleGAN).

**Methods:** A total of 190 paired pelvic CT and iCBCT image datasets were included in the study, out of which 150 were used for model training and the remaining 40 were used for model testing. For the registration network, we proposed a 3D multi-stage registration network (MSnet) to deform planning CT images to agree with iCBCT images, and the contours from CT images were propagated to the corresponding iCBCT images through a deformation matrix. The overlap between the deformed contours (dpCT) and the fixed contours (iCBCT) was calculated for purposes of evaluating the registration accuracy. For the sCT generation, we trained the 2D CycleGAN using the deformation-registered CT-iCBCT slicers and generated the sCT with corresponding iCBCT image data. Then, on sCT images, physicians re-delineated the contours that were compared with contours of manually delineated iCBCT images. The organs for contour comparison included the bladder, spinal cord, femoral head left, femoral head right, and bone marrow. The dice similarity coefficient (DSC) was used to evaluate the accuracy of registration and the accuracy of sCT generation.

**Results:** The DSC values of the registration and sCT generation were found to be 0.769 and 0.884 for the bladder ($p < 0.05$), 0.765 and 0.850 for the spinal cord ($p < 0.05$), 0.918 and 0.923 for the femoral head left ($p > 0.05$), 0.916 and 0.921 for the femoral head right ($p > 0.05$), and 0.878 and 0.916 for the bone marrow ($p < 0.05$), respectively. When the bladder volume difference in planning CT and iCBCT scans was more than double, the accuracy of sCT generation was significantly better than that of registration (DSC of bladder: 0.859 vs. 0.596, $p < 0.05$).

**Conclusion:** The registration and sCT generation could both improve the iCBCT image quality effectively, and the sCT generation could achieve higher accuracy when the difference in planning CT and iCBCT was large.

**Keywords: iCBCT, registration, sCT generation, pelvic, CycleGAN**

# INTRODUCTION

Cervical cancer is an important factor that endangers women's lives (1), and radiotherapy is one of the main ways to treat cervical cancer. The most widely used radiotherapy techniques in clinical practice are IMRT (intensity modulated radiotherapy) (2) and VMAT (volumetric modulated radiotherapy) (3, 4), both of which can provide a high dose to the target area while protecting more organs at risk (OARs). Higher conformity requires higher accuracy of the patient's position during treatment; thus, image-guided radiotherapy (IGRT) is used to monitor changes in the patient's position and anatomical structure during clinical treatment. The acquisition of CT image again may increase the treatment burden and radiation, and CBCT image guidance is most widely accepted in clinical practice. However, the quality of CBCT images is poor due to the scattering and artifacts, which is typically not enough for dose calculation and adaptive radiotherapy. The iterative cone beam CT (iCBCT) combines the statistical reconstruction and Acuros CTS scattering correction algorithm (5, 6), which can achieve uniform imaging with less noise and higher quality. Nevertheless, the artifacts (cavity artifacts, etc.) still exist, which need to be improved.

In recent years, deep learning-based image processing methods have been widely applied to the field of medical imaging, including medical image segmentation (7–9), disease diagnosis (10, 11), medical image denoising (12), and medical image translation (13, 14). The development of deep learning technology has accelerated the process of clinical treatment and improved the mining of medical image information. For the inaccuracy of CBCT images, many scholars have made a lot of contributions to improve the quality of CBCT images based on deep learning methods; some of them used the planning CT (pCT) to be registered to the CBCT to obtain deformed planning CT (dpCT), which was used to approximately replace CBCT as the current treatment images. Duan et al. (15) proposed a patch-wise CT-CBCT registration unsupervised model for thoracic patients; Han et al. (16) used a segmentation similarity loss, in addition to the image similarity loss, to train the network to predict the transformation between the pancreatic CT and CBCT images. Liang et al. (17) developed a deep unsupervised learning (DUL) framework based on a regional deformable model for automated prostate contour propagation from pCT to CBCT. In addition, some scholars tried to generate sCT from CBCT images, which was used to replace CBCT as the current treatment images. Zhao et al. (18) used the modified CycleGAN to generate sCT from MV CBCT; the auto-segmentation and dose calculation based on sCT showed promising results. Liang et al. (19) compared the CycleGAN model with other unsupervised learning methods and demonstrated that CycleGAN (20) outperformed the other models on sCT generation. Chen et al. (21) retrained the head model in the pelvic region, and the improvement of the accuracy proved the generalization feasibility of sCT generation.

However, the registration accuracy of CT-CBCT depends more on the consistency of pCT and CBCT images. Deformable image registration (DIR) enabled accurate contour propagation and dose calculation for head and neck (22), but obtained lower accuracy in more complex anatomical regions such as the lung (23) and pelvis (24). Due to the daily deformation of the patient's anatomy, especially for cases with large differences in bladder volumes in cervical cancer patients, the accuracy of the registration can be greatly compromised. On the other hand, the sCT generation is obtained from the trained model parameters, which may produce some fake structure inconsistent with the CBCT images. Therefore, this study implemented image registration based on MSnet and sCT generation based on CycleGAN to better improve the quality of CBCT images, and analyzed the effect of anatomical structure changes in pCT and CBCT scans on the accuracy of registration and sCT generation.

In this paper, we introduce the dataset acquisition and image processing in *Section 2.1*, deformable image registration and data preprocessing in *Section 2.2*, and the CycleGAN-based CBCT to sCT generation in *Section 2.3*. Then, we present the experimental results in *Section 3* and discuss the experimental results and related research in *Section 4*.

# MATERIALS AND METHODS

## Dataset Acquisition and Image Processing

In this study, 115 cases of cervical cancer were retrospectively collected between June 2021 and October 2021 at Peking Union Medical College Hospital. The patients ranged in age from 32 to 73 years with a median age of 56 years. Among them, each patient includes 1–2 sets of pCT and the corresponding delineation information. The iCBCT was acquired when the patient underwent radiotherapy for the first time normally. Moreover, iCBCT could be obtained in each fraction when the radiotherapy was delivered in the Varian Halcyon 2.0 system. A total of 190 pairs of CT and first fraction iCBCT images were collected, of which 150 were used for model training, and 40 were used for model evaluation. The CT images were obtained on PHILIPS BrillianceTM Bigbore CT, which has a bore with a diameter of 85 cm. The plane resolution of the CT ranged from 0.962 mm × 0.962 mm to 1.365 mm × 1.365 mm, and the slice thickness was 5 mm. The iCBCT images were obtained from the Halcyon system, with a plane resolution ranging from 0.908 mm × 0.908 mm to 1.035 mm × 1.035 mm and a slice thickness of 2 mm. The range of iCBCT was mainly concentrated near the tumor target area, with a length of about 240 mm. Meanwhile, the scanning range of CT is longer than that of iCBCT and can completely cover the scanning range of iCBCT.

The data preprocessing was required before DIR and sCT generation. The common preprocessing is shown in **Figure 1**, which included removing couch, resampling, rigid alignment, and cropping; the specific preprocessing for registration and sCT generation will be introduced later. Firstly, the skin prediction model was combined with the image processing of expansion corrosion, which can quickly and accurately extract the skin mask. The interference of redundant information outside the body was removed, and the HU values outside the body were set to the HU value of the air (−1000). Secondly, the CBCT and CT images were resampled to 1 mm × 1 mm × 5 mm. Then, the CBCT images were set as fixed images, and the CT images were rigidly aligned to the CBCT images based on the ITK rigid registration method (25, 26). The redundant layers in the CT images were removed. Finally, the centroid of the skin mask was set as the image center; 400 × 288 voxels are cropped out of each layer of the image, which can completely contain the outline of the body. It should be emphasized that the entire image preprocessing is fully automatic without manual participation.

## Deformable Image Registration

Although common preprocessing was completed, additional data processing operations for registration required threshold cutoff and normalization. The threshold range of HU values is [−250, 200]; then, the pixel values of the image data were normalized and mapped to the range of (−1, 1).

The used registration method was a 3D multi-stage cascade registration network, which was shown in **Figure 2** and realized the registration of pCT images to CBCT images. The network expected a pair of CT and CBCT images with 400 × 288 × 48 × 2 voxels and output a deformation field with 400 × 288 × 48 × 3 voxels. The network consists of three stages of registration, which achieved accurate registration from coarse to fine. The network architecture is shown in **Figure 2B**, which included two down-sampling layers and two up-sampling layers. Six ResNet Blocks (27) were used to increase the depth of the network and make the model easier to optimize. The loss function of the registration included the MIND (modality-independent neighborhood descriptor) loss ($L_{MIND}$) (28, 29) and smoothing loss ($L_{smooth}$) (30). The model was trained and tested on Nvidia Geforce RTX 3090. The batch was set to 20 with the model in stage 1, 4 in stage 2, and 1 in stage 3. The training required approximately 24 h for 200 epochs.

## CycleGAN-Based CBCT to sCT Generation

Additional data processing operations for sCT generation was required, which was according to the formula.

$$x = Tanh\left(\frac{x}{400}\right)$$

where the Tanh function was the hyperbolic tangent function, defined as

$$Tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

Because the final activation function of the generator model was Tanh, the CBCT and CT images were preprocessed by Tanh, which could improve the accuracy of sCT generation.

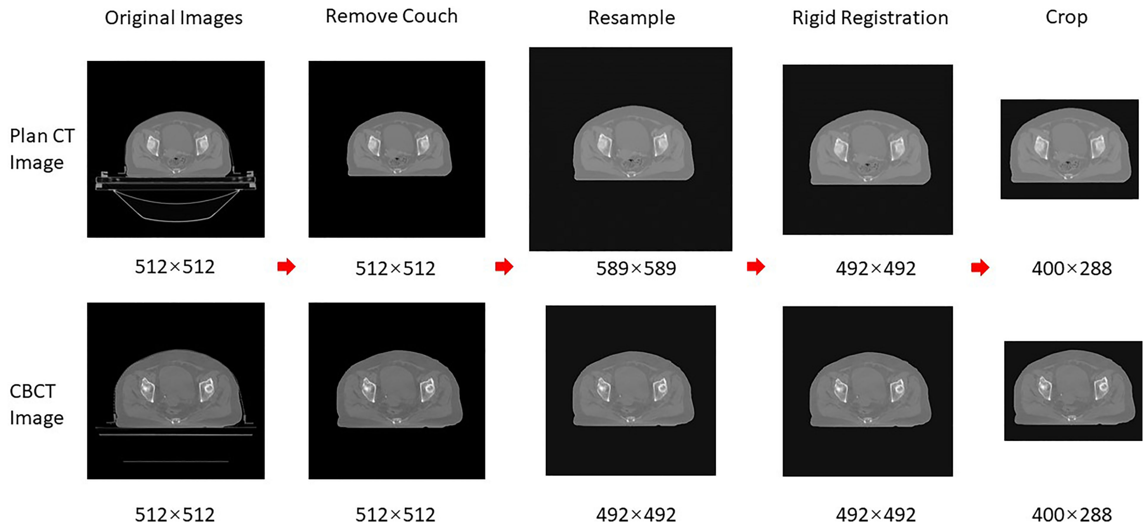


**FIGURE 1** | Schematic diagram of data preprocessing.

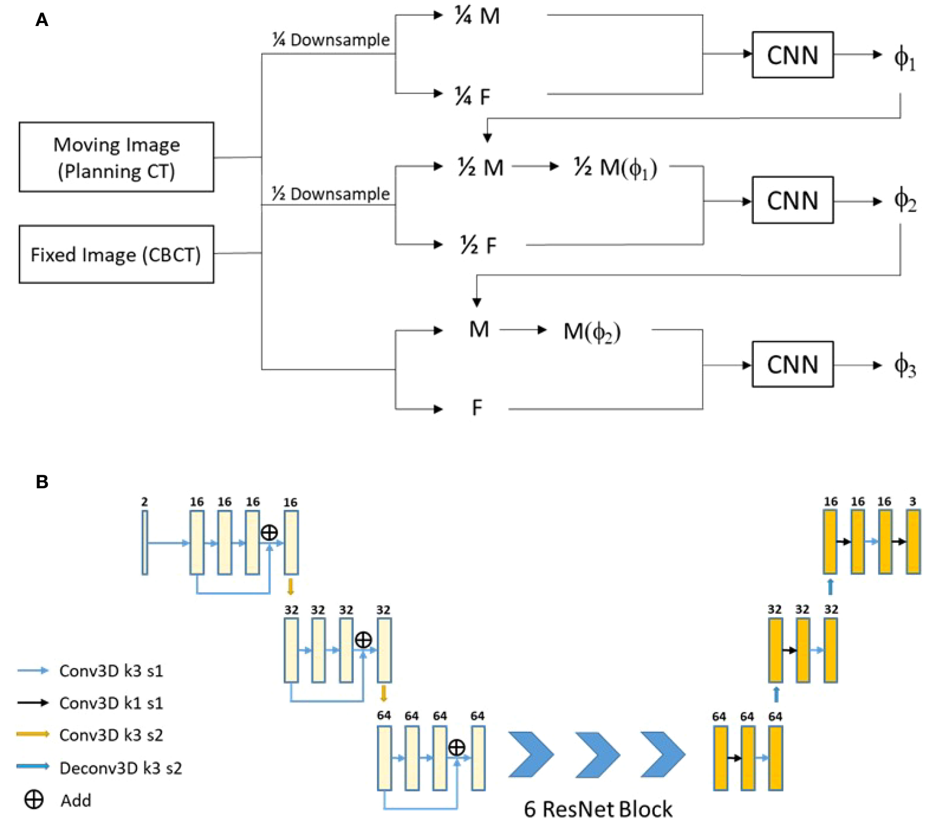


**FIGURE 2** | Our proposed registration method. **(A)** The network flow diagram. **(B)** The network architecture.

The architecture of CycleGAN is shown in **Figure 3A**, which mainly included two generators (Gcbct-ct and Gct-cbct) and two discriminators (Dct and Dcbct): Gcbct-ct generated sCT from the CBCT image, Gct-cbct generated sCBCT from the CT image, Dct identified the sCT image from the real CT image, and Dcbct identified the sCBCT image from the real CBCT image. During the training process, Gcbct-ct would try to generate an sCT that made Dct indistinguishable as much as possible, and then Gct-cbct would convert the sCT image generated in the previous step into the CBCT image, called cycle CBCT, so as to make the CBCT image and the cycle CBCT image as consistent as possible. We compared the accuracy of different networks as generators, such as the U-net (**Figure 3B**) and Resnet (**Figure 3C**). The discriminators used the same architecture as shown in **Figure 3D**.

The loss function of the sCT generation consisted of three parts: ① Adversarial Loss *Ladv*, which could facilitate the distribution of the synthetic images similar to that of the images in the target. ② Cycle-consistency Loss *Lcycle*, which could serve as an indirect constraint of structure between the input and synthetic images. ③ Similarity-constraint Loss *Lsc*, which used the MIND loss to enforce the structural consistency between synthetic images and real images. LG is defined as follows and the hyperparameters $\lambda$ and $\mu$ were set to 10.

$$LG = Ladv + \lambda Lcycle + \mu Lsc$$

The model was trained and tested on Nvidia Geforce RTX 3090. Verified by extensive experiments, the batch was set to 6, the initial learning rate was set to 0.002, and the discrimination rate was set to 0.02. The epoch number was set to 200, and the learning rate decreased linearly from 0.002 to 0 in last 100 epochs.

## Deformable Image Registration Evaluation

In this study, 40 pairs of CBCT and CT images were used to evaluate the registration. Due to the poor quality of CBCT images, the distribution of HU values was also different from CT images; thus, the single-modal similarity measure was not accurate to evaluate the registration. Firstly, objective evaluation criteria were used for images, including normalized mutual information (NMI) and normalized cross-correlation (NCC). Then, the dice similarity coefficient (DSC) was used to evaluate the registration accuracy. The manual contours delineated on CBCT (Mask_CBCT) were used as the ground truth, the contours on the pCT image were propagated to the CBCT image (deformed mask, dMask) through the deformation matrix, and the DSC values of Mask_CBCT and dMask could reflect the accuracy of the registration. The organs for contour comparison included the bladder, spinal cord, femoral head left, femoral head right, and bone marrow.

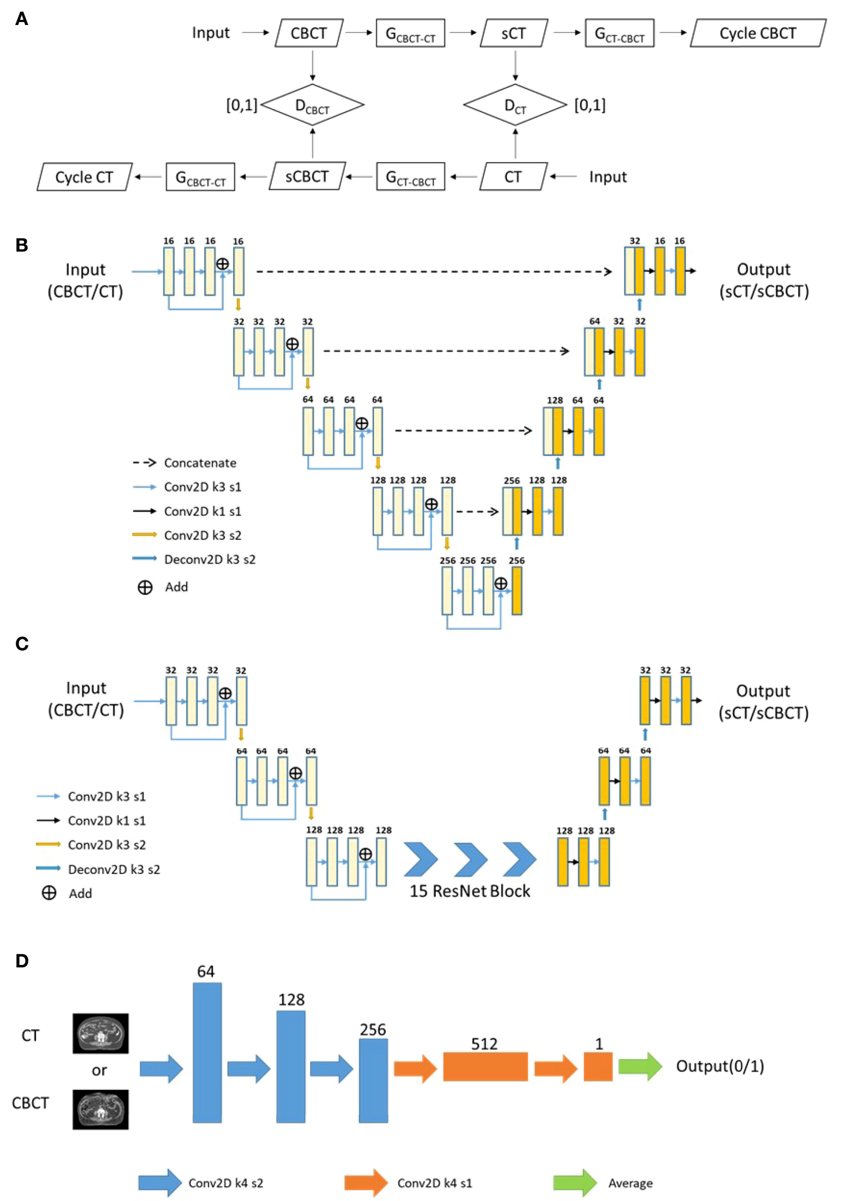


**FIGURE 3** | The flowchart and network architecture of sCT generation. **(A)** Architecture of CycleGAN. **(B)** U-net Generator. **(C)** Resnet Generator. **(D)** Discriminator.

$$NMI(I_1, I_2) = 2\frac{\sum_{i=1}^{I_1}\sum_{j=1}^{I_2} P(i,j)\log\left(\frac{P(i,j)}{P(i)P(j)}\right)}{\left(-\sum_{i=1}^{I_1} P(i)\log(P(i))\right) + \left(-\sum_{j=1}^{I_2} P(j)\log(P(j))\right)} \tag{1}$$

$$NCC(I_1, I_2) = \frac{1}{n_i n_j n_k}\sum_{x,y,z}^{n_i n_j n_k}\frac{(I_1(x,y,z)-\mu_{I_1})(I_2(x,y,z)-\mu_{I_2})}{(\sigma_{I_1}\sigma_{I_2})} \tag{2}$$

$$DSC(V_1, V_2) = \frac{2(V_1 \cap V_2)}{V_1 + V_2} \tag{3}$$

$I_1$ and $I_2$ represent two different images, $P(i)$ means the probability distribution of the variable $i$, $I(x, y, z)$ means the HU value of pixels $(x, y, z)$ in image $I$. $n_i n_j n_k$ is the total number of pixels in image $I$. $\mu$ and $\sigma$ represent the mean and the standard deviation of the HU value in an image. $V_1$ and $V_2$ represent the volume of the two contours for comparison, respectively

## Synthetic CT Image Quality Evaluation

The sCT evaluation criteria included mean absolute error (MAE), root mean square error (RMSE), peak signal-to-noise ratio (PSNR), and structural similarity (SSIM). The corresponding dpCT image with MSnet was used as the ground truth.

$$MAE(I_1, I_2) = \frac{1}{n_i n_j n_k} \sum_{x,y,z}^{n_i n_j n_k} |I_1(x, y, z) - I_2(x, y, z)| \quad (4)$$

$$RMSE(I_1, I_2) = \sqrt{\frac{1}{n_i n_j n_k} \sum_{x,y,z}^{n_i n_j n_k} |I_1(x, y, z) - I_2(x, y, z)|^2} \quad (5)$$

$$PSNR(I_1, I_2) = 10 \times \log_{10} (\frac{MAX^2}{RMSE(I_1, I_2)^2}) \quad (6)$$

$$SSIM(I_1, I_2) = \frac{(2\mu_{I_1}\mu_{I_2} + c_1)(2\sigma_{I_1,I_2} + c_2)}{(\mu_{I_1}2 + \mu_{I_2}2 + c_1)(\sigma_{I_1}2 + \sigma_{I_2}2 + c_2)} \quad (7)$$

MAX was the maximum HU value in the selected image, and other parameters are similar to the above.

Considering the difference in the anatomical structure of the pCT and CBCT images, it is not complete to use the above evaluation criteria to evaluate sCT generation. The DSC was also used for sCT evaluation. The manual contours delineated on CBCT (Mask_CBCT) were regarded as the ground truth, and the physicians re-delineated the contours based on the generated sCT (Mask_sCT). The overlap between Mask_CBCT and Mask_sCT was calculated to evaluate the sCT accuracy. The organs for contour comparison included the bladder, spinal cord, femoral head left, femoral head right, and bone marrow.

# RESULTS

## Deformable Image Registration

The DIR result of pCT and CBCT is shown in **Table 1**. Rigid registration was used for setup verification in the clinic and used for rigid alignment in our experiments, and we wanted to observe further improvement of DIR compared with rigid registration. MSnet registration was compared with the Elastix B-spline registration method (31, 32). It could be seen that both MSnet and the Elastix had improved the registration accuracy to some degree. In addition to the DSC of the bladder, MSnet was better than the Elastix in the evaluation of various indicators. **Figure 4** showed the difference between CT images and CBCT images before and after registration; MSnet had better skin contour alignment. In terms of time, it took 0.15 s for MSnet to get the dpCT for one case, while the Elastix method needed 30–50 s, about two hundred times faster.

**TABLE 1** | The registration result of pCT and CBCT (Ave ± Std).

| | | Rigid | Elastix | MSnet |
|---|---|---|---|---|
| NMI | | 0.350 ± 0.034 | 0.379 ± 0.033 | 0.397 ± 0.033 |
| NCC | | 0.959 ± 0.009 | 0.969 ± 0.008 | 0.980 ± 0.005 |
| DSC | Bladder | 0.738 ± 0.120 | 0.769 ± 0.125 | 0.755 ± 0.121 |
| | Spinal_Cord | 0.631 ± 0.145 | 0.741 ± 0.075 | 0.765 ± 0.088 |
| | Femoral_Head_L | 0.882 ± 0.061 | 0.913 ± 0.022 | 0.918 ± 0.028 |
| | Femoral_Head_R | 0.878 ± 0.052 | 0.891 ± 0.142 | 0.916 ± 0.022 |
| | Bone_Marrow | 0.796 ± 0.071 | 0.858 ± 0.036 | 0.878 ± 0.031 |

## Synthetic CT Generation

**Table 2** shows the CBCT image quality improvement from CBCT images to sCT images, where CBCT images and sCT images were compared with dpCT by metrics including MAE, RMSE, PSNR, and SSIM. **Figure 5** showed visualization of sCT generation for one example. It can be seen from the results that the generator of Resnet with 15 ResNet blocks had a better effect than the generator of U-net with 5 down-sampling layers, which had significant improvement over CBCT in various indicators and less difference with real CT images. The results showed that the ResNet blocks could use feature combinations at different levels to improve CBCT image quality more accurately. Limited by the busy work and manpower, physicians only re-delineated the contours of organs on the sCT produced by the Resnet, which was compared with the contours of CBCT. The DSC results are shown in **Table 3**. It can be seen from the results that the accuracy of sCT was higher than the accuracy of registration. Except for the femoral head left and femoral head right, the remaining three organs had significant differences, which also showed that the sCT had higher structural consistency with CBCT images compared with dpCT. **Figure 6** showed the boxplot of DSC values for registration and sCT generation.

We analyzed the cases with poor registration performance, and found that these cases' anatomical structures of pCT and CBCT were quite different, especially the bladder volume difference. When the volume difference was large, it was difficult to achieve good registration performance. Therefore, we calculated the volume difference of organs in pCT and CBCT (including the bladder, spinal cord, femoral head left, femoral head right, and bone marrow), and then statistically summarized the accuracy of registration and sCT with increasing volume difference. The results are shown in **Table 4**, in which it can be seen that the volume difference of bony structures (femoral head and pelvic) was small, most of the volume difference is less than 1%, and a small part may have a volume difference of less than 1% due to inconsistent delineation levels between the upper and lower ends. The main reason for the lower accuracy of the spinal cord was the different layers delineated in pCT and CBCT images. The bladder volume difference of pCT and CBCT was relatively large among the 40 cases in this study, only 9 had a volume difference of less than 20%, and 11 had a doubled volume difference (Diff > 100%). The DSC value of registration also changed from 0.874 to 0.587. The bladder volume difference was caused by the different degree of bladder filling during pCT scan and CBCT scan, which may be related to factors such as drinking water and waiting time. The above results showed that the volume difference had almost no effect on the accuracy of sCT, and had relatively little effect on the registration accuracy of bony structures (femoral head and pelvis). The volume difference had a great influence on the registration of soft tissues, especially the

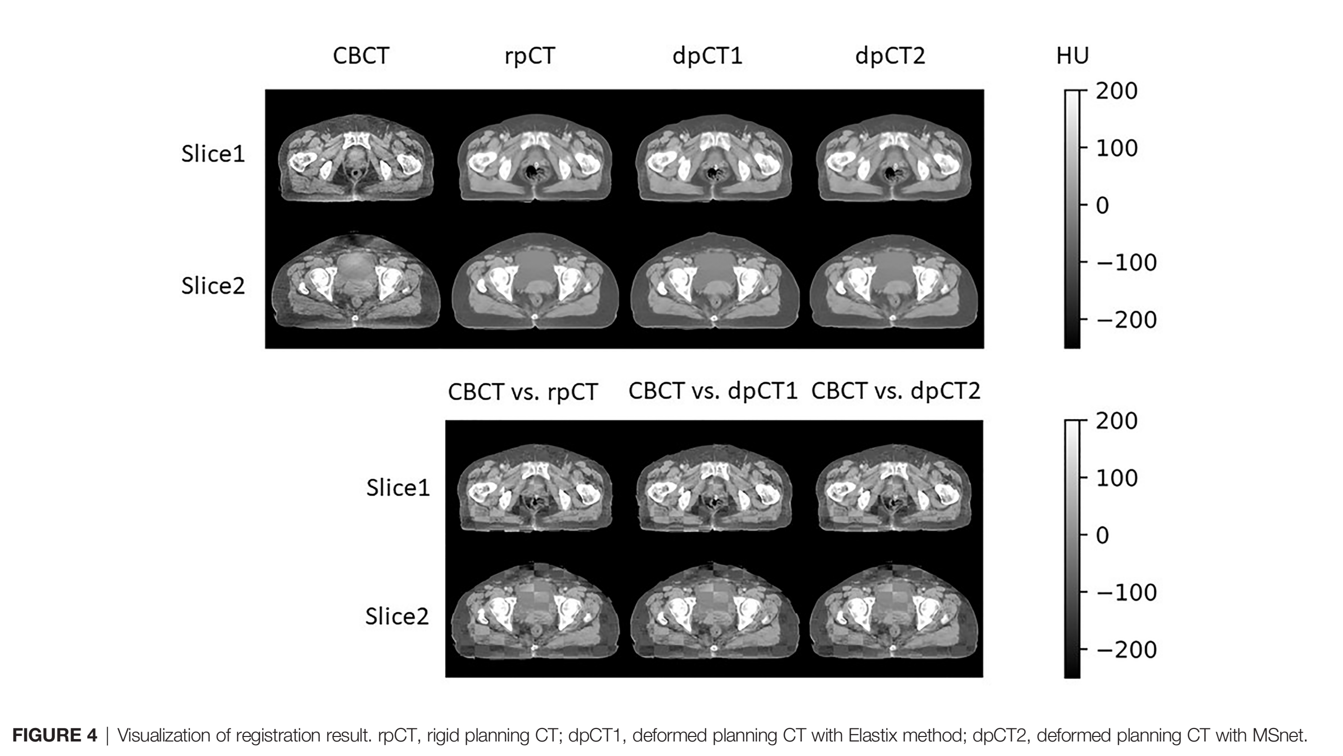


**FIGURE 4** | Visualization of registration result. rpCT, rigid planning CT; dpCT1, deformed planning CT with Elastix method; dpCT2, deformed planning CT with MSnet.

bladder in this study. **Figure 7** shows the effect of bladder volume difference on registration and sCT accuracy. With the increase of bladder volume difference, the delineation accuracy of the bladder in sCT was relatively stable, but the registration accuracy had dropped significantly.

## DISCUSSION

Due to the poor quality of CBCT images, which were often used for patient setup correction before radiotherapy in the current clinical practice, they cannot be used directly for accurate dose calculation. In this study, we had implemented two ways to improve the quality of CBCT images, including the registration of pCT to CBCT and the generation of sCT from CBCT. There existed many studies on CBCT-based dose calculations and CBCT-guided adaptive radiotherapy, which demonstrated that registration and sCT generation were acceptable within error tolerances (33–38). However, few studies had compared the accuracy difference of registration and sCT generation when the anatomical structure changes in pCT and CBCT scans. We conducted this study on cervical cancer cases; 150 pairs of CT and CBCT images were used for model training and 40 independent pairs were used to compare the accuracy. The manual contours delineated on CBCT images were regarded as the ground truth to evaluate the accuracy of registration and sCT generation.

For deformable image registration, we compared our proposed registration method (MSnet) with the Elastix B-spline method. MSnet achieved higher registration accuracy than the Elastix from the analysis of comprehensive indicators, and the time was significantly improved. It could be clearly seen from **Figure 4** that MSnet had higher accuracy in the alignment of skin and bony structures, and **Table 1** also presented the same result. If the bladder volume difference in CT and CBCT images was large, the registration could not be accurate. For the worst case, the DSC of bladder was less than 0.5, which might cause errors on dose calculation and be not eligible for precision radiotherapy. According to the AAPM TG 132 (39), the DSC of registration in the range 0.8–0.9 was acceptable. When the bladder volume difference was more than 50%, the registration was not satisfied.

**TABLE 2** | The result of sCT generation (Ave ± Std).

| | dpCT-CBCT | dpCT-sCT (Resnet) | dpCT-sCT (U-net) |
|---|---|---|---|
| MAE(HU) | 51.23 ± 13.67 | 43.98 ± 10.74 | 46.71 ± 12.71 |
| RMSE | 121.09 ± 30.23 | 117.58 ± 28.22 | 127.96 ± 30.76 |
| PSNR | 20.01 ± 2.74 | 22.23 ± 2.61 | 20.00 ± 3.77 |
| SSIM | 0.623 ± 0.084 | 0.680 ± 0.050 | 0.685 ± 0.055 |

*dpCT, deformed planning CT with MSnet.*

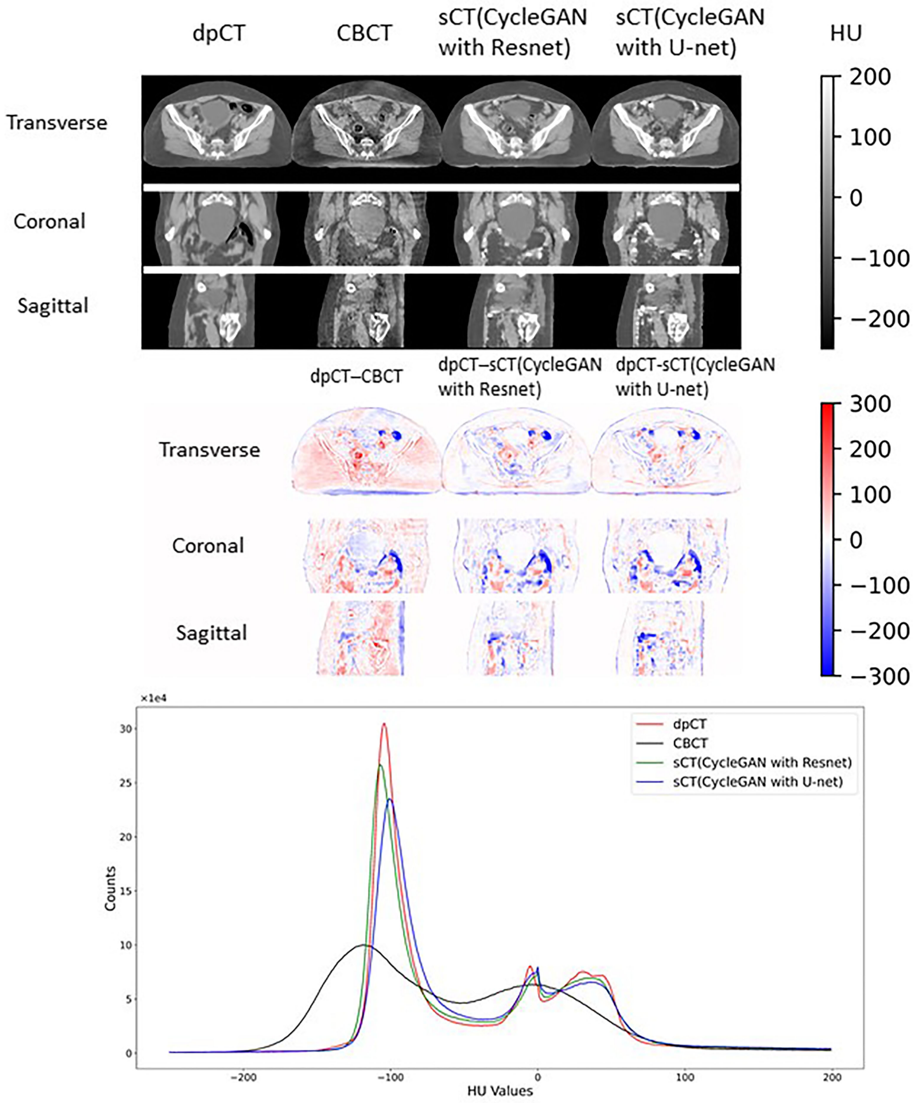


**FIGURE 5** | Visual comparison of dpCT, CBCT, sCT (CycleGAN with Resnet), and sCT (CycleGAN with U-net). The HU difference between two image sets. The HU histogram comparison of dpCT, CBCT, sCT (CycleGAN with Resnet), and sCT (CycleGAN with U-net).

CycleGAN was used to generate sCT from CBCT, which had aroused the interest of many researchers, including KV CBCT and MV CBCT. There are also related studies using different CNN structures as generator models. In this study, the U-net and Resnet were compared as generators to evaluate the accuracy of sCT; the Resnet achieved higher accuracy on our data for metrics such as MAE. Therefore, we generated sCT with the Resnet generator for the testing cases, and the physician re-delineated

**TABLE 3** | The comparison of registration and sCT (Ave ± Std).

| | DSC (sCT, CBCT) | DSC (dpCT1, CBCT) | *p*1-values | DSC (dpCT2, CBCT) | *p*2-values |
|---|---|---|---|---|---|
| Bladder | 0.884 ± 0.071 | 0.769 ± 0.125 | $p < 0.001$ | 0.755 ± 0.121 | <0.001 |
| Spinal_Cord | 0.850 ± 0.039 | 0.741 ± 0.075 | $p < 0.001$ | 0.765 ± 0.088 | <0.001 |
| Femoral_Head_L | 0.923 ± 0.010 | 0.913 ± 0.022 | 0.011 | 0.918 ± 0.028 | 0.265 |
| Femoral_Head_R | 0.921 ± 0.023 | 0.891 ± 0.142 | 0.217 | 0.916 ± 0.022 | 0.238 |
| Bone_Marrow | 0.916 ± 0.009 | 0.858 ± 0.036 | $p < 0.001$ | 0.878 ± 0.031 | <0.001 |

*dpCT1, deformed planning CT with Elastix. p1-value, DSC (sCT, CBCT) vs. DSC (dpCT1, CBCT). dpCT2, deformed planning CT with MSnet. p2-value: DSC (sCT, CBCT) vs. DSC (dpCT2, CBCT).*

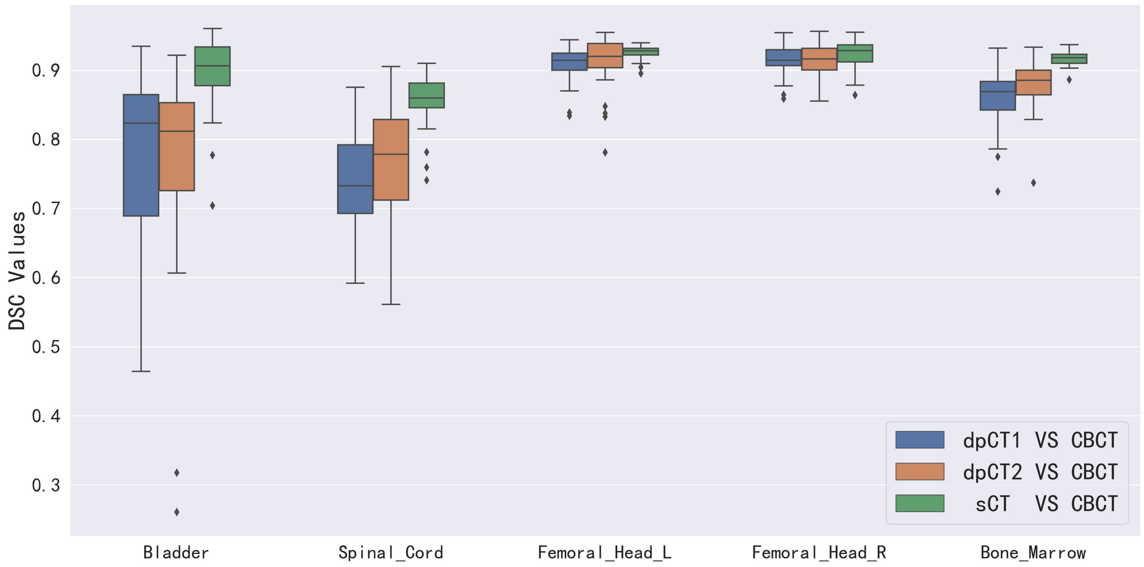


**FIGURE 6** | Boxplot of DSC values for registration and sCT generation. dpCT1, deformed planning CT with Elastix; dpCT2, deformed planning CT with MSnet.

the contours on the sCT images. The results in **Table 3** show that the sCT accuracy was comparable with the registration on bony material, and the sCT had achieved obvious advantages in bladder and spinal cord. **Table 4** further illustrates that the volume difference had little effect on the delineation accuracy of the sCT, but gradually reduced the accuracy of the registration. When the anatomical structure greatly changes, the accuracy of the sCT is higher than that of the registration.

From the analysis of the above results, the sCT generated based on CBCT was superior to dpCT in terms of anatomical structure similarity with the CBCT structure. If the anatomical difference between pCT and CBCT was small, there was little difference between the two methods. Although sCT had higher accuracy, we thought that if the difference between the pCT and CBCT was small, the registration could better reflect the real structure of the case; after all, it was a real CT image. The sCT was generated by a series of parameters obtained from continuously optimizing the data in the training set, which may appear out of nothing compared with the CBCT image. For example, the cavity artifact in the CBCT image was very serious, and the information of the CBCT images was

**TABLE 4** | The effect of volume difference on registration and sCT accuracy (DSC: Ave ± Std).

| | $Diff(V_{CBCT}, V_{pCT})$ | Counts | dpCT1 | dpCT2 | sCT |
|---|---|---|---|---|---|
| Bladder | <20% | 9 | 0.874 ± 0.045 | 0.874 ± 0.043 | 0.898 ± 0.034 |
| | 20%–50% | 13 | 0.846 ± 0.032 | 0.815 ± 0.029 | 0.905 ± 0.034 |
| | 50%–100% | 7 | 0.750 ± 0.046 | 0.737 ± 0.427 | 0.858 ± 0.106 |
| | >100% | 11 | 0.596 ± 0.079 | 0.587 ± 0.066 | 0.859 ± 0.088 |
| Spinal_Cord | <20% | 13 | 0.763 ± 0.060 | 0.805 ± 0.079 | 0.854 ± 0.031 |
| | 20%–50% | 18 | 0.768 ± 0.062 | 0.793 ± 0.041 | 0.854 ± 0.046 |
| | 50%–100% | 5 | 0.680 ± 0.043 | 0.700 ± 0.026 | 0.848 ± 0.029 |
| | >100% | 4 | 0.620 ± 0.023 | 0.582 ± 0.244 | 0.814 ± 0.013 |
| Femoral_Head_L | <1% | 7 | 0.925 ± 0.012 | 0.931 ± 0.015 | 0.924 ± 0.007 |
| | 1%–3% | 14 | 0.910 ± 0.027 | 0.902 ± 0.038 | 0.920 ± 0.011 |
| | 3%–5% | 14 | 0.905 ± 0.018 | 0.925 ± 0.017 | 0.926 ± 0.009 |
| | >5% | 5 | 0.927 ± 0.010 | 0.924 ± 0.013 | 0.923 ± 0.006 |
| Femoral_Head_R | <1% | 6 | 0.911 ± 0.017 | 0.930 ± 0.018 | 0.926 ± 0.011 |
| | 1%–3% | 10 | 0.917 ± 0.022 | 0.925 ± 0.018 | 0.926 ± 0.019 |
| | 3%–5% | 14 | 0.915 ± 0.024 | 0.913 ± 0.022 | 0.912 ± 0.024 |
| | >5% | 10 | 0.910 ± 0.013 | 0.901 ± 0.021 | 0.922 ± 0.025 |
| Bone_Marrow | <2% | 6 | 0.863 ± 0.016 | 0.889 ± 0.019 | 0.920 ± 0.011 |
| | 2%–5% | 17 | 0.871 ± 0.020 | 0.885 ± 0.019 | 0.917 ± 0.011 |
| | 5%–10% | 11 | 0.857 ± 0.039 | 0.885 ± 0.018 | 0.917 ± 0.005 |
| | >10% | 6 | 0.818 ± 0.054 | 0.834 ± 0.047 | 0.909 ± 0.006 |

$Diff(V_{CBCT}, V_{pCT}) = \frac{|V_{CBCT} - V_{pCT}|}{MIN(V_{CBCT}, V_{pCT})} \times 100\%.$

*dpCT1, deformed planning CT with Elastix; dpCT2, deformed planning CT with MSnet.*

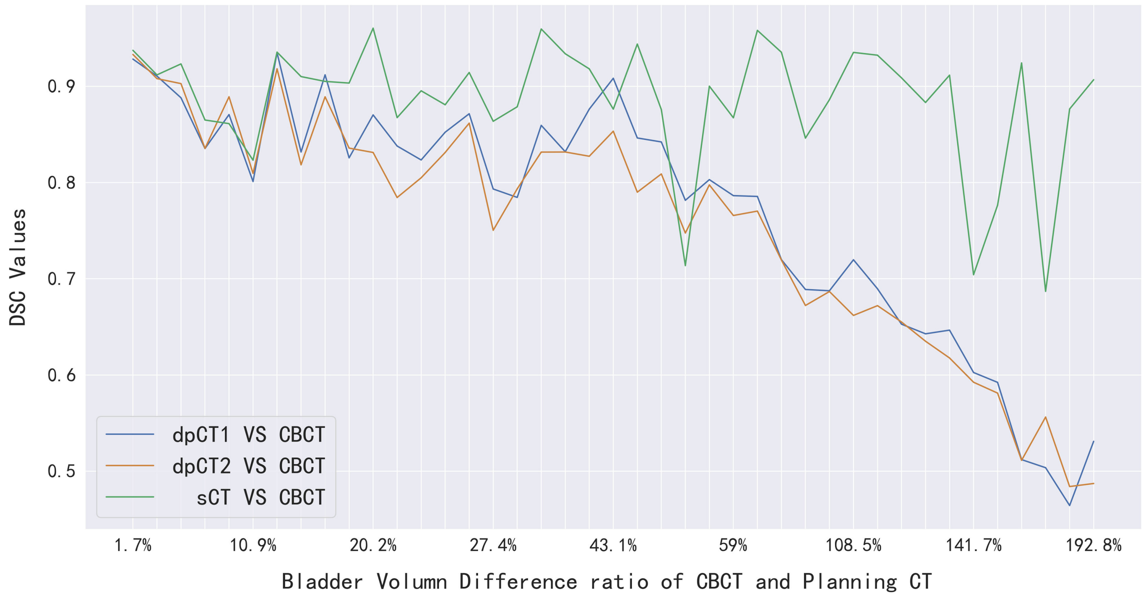


**FIGURE 7** | The effect of bladder volume difference on registration and sCT accuracy. dpCT1, deformed planning CT with Elastix; dpCT2, deformed planning CT with MSnet.

insufficient, which may bring errors in the post-processing correction. In addition, structures such as the bladder and the prostate were close to each other, and the HU values were also very similar, which cannot be identified on the sCT in some instance. Although some studies thought that this situation had little effect on the dose calculation [11], the errors did exist in anatomical structure.

We had studied two methods to improve the image quality of CBCT, and if the two methods could be effectively combined, they may lead to better clinical applications. Note that the difference in bladder volume between pCT images and CBCT images was a major factor affecting the registration accuracy, which could be used as a judgment condition for choosing two methods. We evaluated the accuracy of auto-segmentation on sCT, and the DSC of bladder was 0.874 ± 0.072, which can replace the contours on CBCT approximately. Firstly, we have the pCT images and corresponding contours. When the CBCT images were obtained before radiotherapy, the pCT was registered to the CBCT to obtain the propagated contours, especially the contours of the bladder (dpCT_bladder). Secondly, the CBCT was transformed to sCT, which can be used for auto-segmentation; we can get the contours of bladder on sCT (sCT_bladder). If the DSC of dpCT_bladder and sCT_bladder was above a certain threshold (e.g., DSC > 0.8), the dpCT and corresponding contours would be used. If it was below a certain threshold, the physician would check the auto-segmentation of the sCT for the current radiotherapy, and the generated sCT can be used for dose calculation and evaluation of adaptive radiotherapy. The above process can be done automatically in a short time (less than 1 min), which can be used for more accurate dose tracking.

Several limitations should be noted in this study. First, we selected five OARs to evaluate the accuracy of registration and sCT generation, but the target was the most important concern in clinical practice. It was difficult to delineate the target volume, the small intestine, and rectum on CBCT images due to the existence of artifacts, which were also controversial as the ground truth. In future work, the cases with small differences in anatomical structures can be selected to evaluate the accuracy of target delineation in the sCT images. Second, the focus of this study was to compare the accuracy of registration and sCT generation on structural similarity; the dosimetric differences would be done in our next work.

## CONCLUSION

We proposed two methods to improve the image quality of CBCT in this study. Both registration and sCT generation can effectively improve the image quality of CBCT. When the anatomical structure changes in pCT and CBCT scans were small, the accuracy of the registration and sCT was equivalent, and the anatomical structure of CBCT could be better represented by dpCT. When the anatomical structure changes were large, the accuracy of the sCT was higher than that of the registration, and the anatomical structure of CBCT could be better represented by sCT.

## DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

## ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Medical Ethics Committee of Peking Union Medical College Hospital. The patients/participants provided their written informed consent to participate in this study.

## AUTHOR CONTRIBUTIONS

BY and YC conceived the experiments. YL and ZW collected the clinical dataset. XP, XX, and JQ designed the study and analyzed the result. BY, YC, YL, XP, and JQ participated in writing the manuscript. All authors contributed to the article and approved the submitted version.

## FUNDING

This work was supported by the China National Key R&D Program during the 13th Five-year Plan Period (Grant Nos. 2016YFC0105206 and 2016YFC0105207) and the University of Science and Technology of China (USTC) grants on “New Medicine Team Project: The ROADMAP Medical Physics Platform” (No. YD2140002002).

## REFERENCES

1. Freddie B, Jacques F, Isabelle S, Siegel RL, Torre LA, Jemal A. Global Cancer Statistics 2018: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J Clin* (2018) 68:394–424. doi: 10.3322/caac.21492
2. Cho B. Intensity-Modulated Radiation Therapy: A Review With a Physics Perspective. *Radiat Oncol J* (2018) 36(1):1–10. doi: 10.3857/roj.2018.00122.e1
3. Boylan C, McWilliam A, Johnstone E, Rowbottom C. The Impact of Continuously-Variable Dose Rate VMAT on Beam Stability, MLC Positioning, and Overall Plan Dosimetry. *J Appl Clin Med Phys* (2012) 13:254– 266. doi: 10.1120/jacmp.v13i6.4023
4. Chun M, Joon An H, Kwon O, Oh DH, Park JM, Kim JI., et al. Impact of Plan Parameters and Modulation Indices on Patient-Specific QA Results for Standard and Stereotactic VMAT. *Physica Med* (2019) 62:83– 94. doi: 10.1016/j.ejmp.2019.05.005
5. Maslowski A, Wang A, Sun M, Wareing T, Davis I, Star-Lack J.. Acuros CTS: A Fast, Linear Boltzmann Transport Equation Solver for Computed Tomography Scatter - Part I: Core Algorithms and Validation. *Med Phys* (2018) 45(5):1899–913. doi: 10.1002/mp.12850
6. Wang A, Maslowski A, Messmer P, Lehmann M, Strzelecki A, Yu E, et al. Acuros CTS: A Fast, Linear Boltzmann Transport Equation Solver for Computed Tomography Scatter - Part II: System Modeling, Scatter Correction, and Optimization. *Med Phys* (2018) 45(5):1914–25. doi: 10.1002/mp.12849
7. Wang Z, Chang YK, Peng Z, Lv Y, Shi W, Wang F, et al. Evaluation of Deep Learning-Based Auto-Segmentation Algorithms for Delineating Clinical Target Volume and Organs at Risk Involving Data for 125 Cervical Cancer Patients. *J Appl Clin Med Phys* (2020) 21(12):272–9. doi: 10.1002/acm2.13097
8. Peng Z, Fang X, Yan P, et al. A Method of Rapid Quantification of Patient-Specific Organ Doses for CT Using Deeplearning-Based Multiorgan Segmentation and GPU-Accelerated Monte Carlo Dose Computing. *Med Phys* (2020) 47:2526–36. doi: 10.1002/mp.14131
9. Chang Y, Wang Z, Peng Z, et al. Clinical Application and Improvement of a CNN-Based Autosegmentation Model for Clinical Target Volumes in Cervical Cancer Radiotherapy. *J Appl Clin Med Phy* (2021) 22(11):115–25. doi: 10.1002/acm2.13440
10. Cheng G, He L. Dr. Pecker: A Deep Learning-Based Computer-Aided Diagnosis System in Medical Imaging. *Deep Learning in Healthcare.* (2019) 171(2020):203–16. doi: 10.1007/978-3-030-32606-7_12
11. Hong-meng L, Di Z, Xue-bin C. Deep Learning for Early Diagnosis of Alzheimer’s Disease Based on Intensive AlexNet. *Comput Sci* (2017) 2014:1015–8. doi: 10.1016/B978-0-12-819764-6.00005-3
12. Peng Z, Ni M, Shan HM, Lu Y, Li Y, Zhang Y, et al. Feasibility Evaluation of PET Scan-Time Reduction for Diagnosing Amyloid-β Levels in Alzheimer’s Disease Patients Using a Deep-Learning-Based Denoising Algorithm. *Comput Biol Med* (2021) 138:104919. doi: 10.1016/j.compbiomed.2021.104919
13. Liu Y, Lei Y, Wang T, Fu Y, Tang X, Curran W, et al. CBCT-Based Synthetic CT Generation Using Deep-Attention cycleGAN for Pancreatic Adaptive Radiotherapy. *Med Phys* (2020) 47(6):2472–83. doi: 10.1002/mp.14121
14. Lei Y, Tang X, Higgins K, Lin J, Jeong J, Liu T, et al. Learning-Based CBCT Correction Using Alternating Random Forest Based on Auto-Context Model. *Med Phys* (2019) 46(2):601–18. doi: 10.1002/mp.13295
15. Duan L, Ni X, Liu Q, Gong L, Yuan G, Li M, et al. Unsupervised Learning for Deformable Registration of Thoracic CT and Cone-Beam CT Based on Multiscale Features Matching With Spatially Adaptive Weighting. *Med Phys* (2020) 47(11):5632– 47. doi: 10.1002/mp.14464
16. Han X, Hong J, Reyngold M, Crane C, Cuaron J, Hajj C, et al. Deep-Learning-Based Image Registration and Automatic Segmentation of Organs-at-Risk in Cone-Beam CT Scans From High-Dose Radiation Treatment of Pancreatic Cancer. *Med Phys* (2021) 48(6):3084–95. doi: 10.1002/mp.14906
17. Liang X, Bibault JE, Leroy T, Escande A, Zhao W, Chen Y, et al. Automated Contour Propagation of the Prostate From pCT to CBCT Images *via* Deep Unsupervised Learning. *Med Phys* (2021) 48(4):1764–70. doi: 10.1002/mp.14755
18. Zhao J, Chen Z, Wang J, Xia F, Peng J, Hu Y, et al. MV CBCT-Based Synthetic CT Generation Using a Deep Learning Method for Rectal Cancer Adaptive Radiotherapy. *Front Oncol* (2021) 11:655325. doi: 10.3389/fonc.2021.655325
19. Liang X, Chen L, Nguyen D, Zhou Z, Gu X, Yang M, et al. Generating Synthesized Computed Tomography (CT) From Cone-Beam Computed Tomography (CBCT) Using CycleGAN for Adaptive Radiation Therapy. *Phys Med Biol* (2019) 64(12):125002. doi: 10.1088/1361-6560/ab22f9
20. Zhu J-Y, Park T, Isola P, Efros AA. Unpaired Image-To-Image Translation Using Cycle-Consistent Adversarial Networks. *Proc IEEE Int Conf Comput Vision* (2017) 1:2242–51. doi: 10.1109/ICCV.2017.244
21. Chen L, Liang X, Shen C, Nguyen D, Jiang S, Wang J., et al. Synthetic CT Generation From CBCT Images via Unsupervised Deep Learning. *Phys Med Biol* (2021) 66(11):115019. doi: 10.1088/1361-6560/ac01b6
22. Peroni M, Ciardo D, Spadea MF, Riboldi M, Comi S, Alterio D, et al. Automatic Segmentation and Online virtualCT in Head-and-Neck Adaptive Radiation Therapy. *Int J Radiat Oncol Biol Phys* (2012) 84:e427–33. doi: 10.1016/j.ijrobp.2012.04.003
23. Veiga C, Janssens G, Teng CL, Baudier T, Hotoiu L, McClelland JR, et al. First Clinical Investigation of Cone Beam Computed Tomography and Deformable Registration for Adaptive Proton Therapy for Lung Cancer. *Int J Radiat Oncol Biol Phys* (2016) 95:549–59. doi: 10.1016/j.ijrobp.2016.01.055
24. Kurz C, Kamp F, Park YK, Zöllner C, Rit S, Hansen D, et al. Investigating Deformable Image Registration and Scatter Correction for CBCT-Based Dose Calculation in Adaptive IMPT. *Med Phys* (2016) 43:5635–46. doi: 10.1118/1.4962933
25. McCormick M, Liu X, Jomier J, Marion C, Ibanez L.. ITK: Enabling Reproducible Research and Open Science. *Front Neuroinform* (2014) 8:13. doi: 10.3389/fninf.2014.00013
26. Yoo TS, Ackerman MJ, Lorensen WE, Schroeder W, Chalana V, Aylward S, et al. Engineering and Algorithm Design for an Image Processing API: A Technical Report on ITK – The Insight Toolkit. In: J Westwood, editor. *Proc. Of Medicine Meets Virtual Reality*. Amsterdam: IOS Press (2002). pp 586–592.
27. He K, Zhang X, Ren S, Sun J. (2016). Deep Residual Learning for Image Recognition, in: *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 770–8. doi: 10.1109/CVPR.2016.90

28. Heinrich MP, Jenkinson M, Bhushan M, Matin T, Gleeson FV, Brady SM, et al. MIND: Modality Independent Neighbourhood Descriptor for Multi-Modal Deformable Registration. *Med Image Anal* (2012) 16:1423–35. doi: 10.1016/j.media.2012.05.008
29. Guo Y, Wu X, Wang Z, Pei X, Xu XG., et al. End-To-End Unsupervised Cycle-Consistent Fully Convolutional Network for 3D Pelvic CT-MR Deformable Registration. *J Appl Clin Med Phys* (2020) 21(9):193–200. doi: 10.1002/acm2.12968
30. De Vos BD, Berendsen FF, Viergever MA, Sokooti H, Staring M, Išgum I., et al. A Deep Learning Framework for Unsupervised Affine and Deformable Image Registration. *Med Image Anal* (2019) 52:128–43. doi: 10.1016/j.media.2018.11.010
31. Klein S, Staring M, Murphy K, Viergever M. A., Pluim J. P.W.. "Elastix: A Toolbox for Intensity Based Medical Image Registration,". *IEEE Trans Med Imaging* (2010) 29(1):196–205. doi: 10.1109/TMI.2009.2035616
32. Shamonin DP, Bron EE, Lelieveldt BPF, 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(50):1–15. doi: 10.3389/fninf.2013.00050
33. Buranaporn P, Dankulchai P, Jaikuna T, Prasartseree T.. Relation Between DIR Recalculated Dose Based CBCT and GI and GU Toxicity in Postoperative Prostate Cancer Patients Treated With VMAT. *Radiother Oncol* (2021) 157:8–14. doi: 10.1016/j.radonc.2020.12.036
34. Bobić M, Lalonde A, Sharp GC, Grassberger C, Verburg JM, Winey BA, et al. Comparison of Weekly and Daily Online Adaptation for Head and Neck Intensity-Modulated Proton Therapy. *Phys Med Biol* (2021) 10(1088):1361–6560. doi: 10.1088/1361-6560/abe050
35. Kida S, Nakamoto T, Nakano M, Nawa K, Haga A, Kotoku J, et al. Cone Beam Computed Tomography Image Quality Improvement Using a Deep Convolutional Neural Network. *Cureus* (2018) 10(4):e2548. doi: 10.7759/cureus.2548
36. Dong G, Zhang C, Liang X, Deng L, Zhu Y, Zhu X, et al. A Deep Unsupervised Learning Model for Artifact Correction of Pelvis Cone-Beam Ct. *Front Oncol* (2021) 11:686875. doi: 10.3389/fonc.2021.686875
37. Zhang Y, Yue N, Su MY, Liu B, Ding Y, Zhou Y, et al. Improving CBCT Quality to CT Level Using Deep Learning With Generative Adversarial Network. *Med Phys* (2021) 48(6):2816–26. doi: 10.1002/mp.14624
38. Kurz C, Maspero M, Savenije MHF, Landry G, Kamp F, Pinto M, et al. CBCT Correction Using a Cycle-Consistent Generative Adversarial Network and Unpaired Training to Enable Photon and Proton Dose Calculation. *Phys Med Biol* (2019) 64(22):225004. doi: 10.1088/1361-6560/ab4d8c
39. Brock KK, Mutic S, McNutt TR, Li H., Kessler ML.. Use of Image Registration and Fusion Algorithms and Techniques in Radiotherapy: Report of AAPM Radiation Therapy Committee Task Group No. 132. *Med Phys* (2017) 44:43–76. doi: 10.1002/mp.12256

**Conflict of Interest:** Author XP is employed by Anhui Wisdom Technology Co., Ltd.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

**Publisher's Note:** All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

ORIGINAL RESEARCH
published: 01 June 2022
doi: 10.3389/fonc.2022.908903

# A Feasibility Study of Deep Learning-Based Auto-Segmentation Directly Used in VMAT Planning Design and Optimization for Cervical Cancer

*Along Chen*[1†], *Fei Chen*[2†], *Xiaofang Li*[3], *Yazhi Zhang*[4], *Li Chen*[1], *Lixin Chen*[1*] and *Jinhan Zhu*[1*]

*[1] Department of Radiation Oncology, State Key Laboratory of Oncology in South China, Collaborative Innovation Center for Cancer Medicine, Sun Yat-sen University Cancer Center, Guangzhou, China, [2] School of Biomedical Engineering, Guangzhou Xinhua University, Guangzhou, China, [3] Department of Radiation Oncology, The Second Affiliated Hospital of Zunyi Medical University, Zunyi, China, [4] Department of Oncology and Hematology, The Six People's Hospital of Huizhou City, Huiyang Hospital Affiliated to Southern Medical University, Huizhou, China*

**Edited by:**
*Humberto Rocha,*
*University of Coimbra, Portugal*

**Reviewed by:**
*Kuo Men,*
*Chinese Academy of Medical Sciences and Peking Union Medical College, China*
*Nanna Maria Sijtsema,*
*University of Groningen, Netherlands*
*John Sage,*
*Mirada Medical, United Kingdom*

***Correspondence:**
*Lixin Chen*
*chenlx@sysucc.org.cn*
*Jinhan Zhu*
*zhujh@sysucc.org.cn*

*†These authors have contributed equally to this work and share first authorship*

**Specialty section:**
*This article was submitted to Radiation Oncology, a section of the journal Frontiers in Oncology*

**Received:** *31 March 2022*
**Accepted:** *06 May 2022*
**Published:** *01 June 2022*

**Citation:**
*Chen A, Chen F, Li X, Zhang Y, Chen L, Chen L and Zhu J (2022) A Feasibility Study of Deep Learning-Based Auto-Segmentation Directly Used in VMAT Planning Design and Optimization for Cervical Cancer. Front. Oncol. 12:908903. doi: 10.3389/fonc.2022.908903*

**Purpose:** To investigate the dosimetric impact on target volumes and organs at risk (OARs) when unmodified auto-segmented OAR contours are directly used in the design of treatment plans.

**Materials and Methods:** A total of 127 patients with cervical cancer were collected for retrospective analysis, including 105 patients in the training set and 22 patients in the testing set. The 3D U-net architecture was used for model training and auto-segmentation of nine types of organs at risk. The auto-segmented and manually segmented organ contours were used for treatment plan optimization to obtain the AS-VMAT (automatic segmentations VMAT) plan and the MS-VMAT (manual segmentations VMAT) plan, respectively. Geometric accuracy between the manual and predicted contours were evaluated using the Dice similarity coefficient (DSC), mean distance-to-agreement (MDA), and Hausdorff distance (HD). The dose volume histogram (DVH) and the gamma passing rate were used to identify the dose differences between the AS-VMAT plan and the MS-VMAT plan.

**Results :** Average DSC, MDA and $HD_{95}$ across all OARs were 0.82–0.96, 0.45–3.21 mm, and 2.30–17.31 mm on the testing set, respectively. The $D_{99\%}$ in the rectum and the Dmean in the spinal cord were 6.04 Gy ($P = 0.037$) and 0.54 Gy ($P = 0.026$) higher, respectively, in the AS-VMAT plans than in the MS-VMAT plans. The $V_{20}$, $V_{30}$, and $V_{40}$ in the rectum increased by 1.35% ($P = 0.027$), 1.73% ($P = 0.021$), and 1.96% ($P = 0.008$), respectively, whereas the $V_{10}$ in the spinal cord increased by 1.93% ($P = 0.011$). The differences in other dosimetry parameters were not statistically significant. The gamma passing rates in the clinical target volume (CTV) were 92.72% and 98.77%, respectively, using the 2%/2 mm and 3%/3 mm criteria, which satisfied the clinical requirements.

**Conclusions:** The dose distributions of target volumes were unaffected when auto-segmented organ contours were used in the design of treatment plans, whereas the impact of automated segmentation on the doses to OARs was complicated. We suggest that the auto-segmented contours of tissues in close proximity to the target volume need to be carefully checked and corrected when necessary.

**Keywords: deep learning, automatic segmentation, dosimetric differences, geometric accuracy, cervical cancer**

# 1 INTRODUCTION

In radiotherapy, automatic delineation of normal tissues based on deep learning techniques is an increasingly mature technique, and the automatic delineation of target volumes has been explored in successive multicentre clinical application studies. The convolutional neural network (CNN) is superior to most other algorithms in the segmentation of medical images (1), and, as a result, it is often used for the automatic delineation of normal tissues and target volumes (2–5) on computed tomography (CT) images of the head and neck (6–8), chest (9), abdomen (10, 11), and pelvic cavity (5, 12–15), among others.

Radiotherapy is an effective treatment for cervical cancer (16, 17), and delivery of precision radiotherapy requires accurate contouring of each organ on the patient's CT images. Manual segmentation of normal tissues depends on the experience and ability of the imaging radiologist (18, 19) and has a low efficiency. The poor contrast of pelvic soft tissues on CT images also presents challenges for radiologists. With the rapid development of image segmentation techniques, CNN-based automated organ contouring on CT images has become increasingly popular for patients with cervical cancer. Liu et al. (20) used the improved U-Net model to automatically segment cervical cancer organs at risk (OARs), and the model prediction was highly consistent with the OARs delineated by radiation oncologists. Ju et al. (21) innovatively integrated the Dense Net model with the V-Net model, enabling accurate, efficient, and automatic delineation of six OARs on CT images. Qualitative and quantitative studies conducted by Rhee et al. (5) showed that the auto-contouring tool based on CNN can be used to generate the segmentation of OARs and clinical target volume (CTV) for patients with cervical cancer and achieve clinically acceptable delineation results.

Despite these encouraging results, many challenges remain to be overcome before auto-segmentation methods can be applied in clinical practice. First, patients with cervical cancer are treated in supine or prone positions, and no study has examined whether different patient positions affect automatic delineations of normal tissues. Second, there remains room for improvement in the accuracy of automatic soft tissue segmentation, such as in colons and rectums. More importantly, existing assessments of accuracy in automated normal tissue segmentation are limited to the comparison of geometric accuracy, and few studies have focused on their relevant dosimetric impact. However, a model successfully segments the OARs in geometry is not sufficient to confirm its reliability for clinical application. Fung et al. (22) and Zhu et al. (23) introduced their dosimetric evaluation methods about dose impact between manually and automatically segmented OARs. Vinod et al. (24) believed that it is important to quantify the degree of uncertainty in volume segmentation, but the resulting impact on dosimetry and clinical significance is a more relevant endpoint.

Patients with cervical cancer with different therapeutic positions were included in this study for model training. We then performed automatic delineation of nine types of normal tissues and evaluated its geometric accuracy. On this basis, we discussed the impact of unmodified auto-contouring of tissue structures on the design and optimization of treatment plans. We attempted to use experimental data to investigate the following: 1) whether the dose distribution inside the clinical target volume is affected, and in the case of dose deviations, whether these deviations are within a clinically acceptable range; and 2) whether dose deviations to organs at risk are clinically acceptable.

# 2 MATERIALS AND METHODS

## 2.1 Case Selection

This study included a total of 127 patients with cervical cancer who received radiotherapy at Sun Yat-sen University Cancer Centre between December 2020 and August 2021, including 65 patients in the supine position and 62 patients in the prone position. None of the included patients underwent intestinal tract modification surgery. The images were obtained using a Philips large-aperture CT simulation scanner (Philips Brilliance Big Bore, Netherlands) at 140 keV voltage and a 3-mm slice thickness. The size of images for each slice was 512 × 512 and the number of slices ranged between 140 and 205.

Three clinicians used the Monaco (V5.11) treatment planning system to manually segment bone structures (including the left femoral head, the right femoral head, and the pelvis) as well as tissues and organs (including the spinal cord, the left and right kidneys, the bladder, the rectum, and the colon) from the patient's CT images. Each organ at risk was segmented in strict accordance with the requirements in the radiation therapy oncology group (25) guidelines and the delineation results were reviewed and modified by senior radiation therapists.

## 2.2 Data Pre-Processing

The 105 sets of CT images obtained were used for model training, including 52 sets obtained in the supine position and 53 sets obtained in the prone position. To increase the training sample size, the CT images were cropped into sub-images 100 ×

100 × 100 in size, with random positions selected in the whole body range as the starting points. In addition, 22 sets of CT images were selected for model testing, including 13 sets obtained in the supine position and 9 sets obtained in the prone position.

To highlight the soft tissues, bones, and bladders in the images, we also added 3 types of images processed with different window widths and window levels to the original input sub-images, including soft tissue images: window width = 400, window level = 40; bone images: window width = 1000, window level = 400; bladder images: window width = 250, window level = 50. Hence, the input to the training model was: 4 × 100 × 100 × 100. All input images were normalised to the range of 0–1.

## 2.3 Model Training

The training labels were filled with one-hot images containing ten channels according to the manually segmented structures. The one-hot images were binary class matrices which have zeros everywhere except where the index of channel matches the corresponding value of the class number, in which case it will be 1. The 1st channel represented the undelineated areas; the 2nd channel was marked as the bladder (bladder); the 3rd channel was marked as the left femoral head (femoral_ joint L); the 4th channel was marked as the right femoral head (femoral_ joint R); the 5th channel was marked as the rectum (rectum); the 6th channel was marked as the colon (colon); the 7th channel was marked as the left kidney (Kidney-L); the 8th channel was marked as the right kidney (Kidney-R); the 9th channel was marked as the pelvic bone (PelvicBone); and the 10th channel was marked as the spinal cord (SpinalCord). The dice similarity coefficient (DSC) is commonly used to measure the overlap of two structures (26, 27), and was adopted as the loss function, while the AdamW (28, 29) optimizer was used to train the CNN network. The batch size was set to 2 in the training algorithm and the learning rate was set using the OneCycleLR learning rate scheduler (30), with the maximum learning rate set to 0.01 and the minimum learning rate set to $4e^{-8}$. Cosine annealing was adopted to schedule the learning rate and the step size was set to per sample. The model was trained for a total of 30 epochs and the model parameters were updated based on the minimum loss value of the evaluation set. The 3D U-net architecture (**Figure 1**) used in previous studies (31) was adopted in the model and a 1 × 1 × 1 convolution kernel was utilised in the last layer, with SoftMax as the activation function. The number of image layers with eigenvalues was reduced to 10 before data output.

## 2.4 Assessment Indicators

### 2.4.1 Assessment of Geometrical Differences

The manually segmented organ contours served as "the golden standard" and the auto-segmentation results were compared with the manual delineation results to assess the accuracy. The assessment indicators include the DSC, the Hausdorff Distance (HD), the 95th percentile of the HD, and the Mean Distance to Agreement (MDA) (32). The commercial software MIM (V6.9, MIM Software Inc., Cleveland, OH, USA) and 3D Slicer (V4.8.1) were used to identify and evaluate the geometrical differences between the automated and manual segmentation results.

### 2.4.2 Evaluation of Dose Differences

To evaluate the impact of geometrical differences between automated and manual segmentation on the dosimetric parameters in treatment plans, we selected 22 patients from the testing set and performed optimization procedures with auto-segmented organ contours to obtain new treatment plans (automatic segmentations VMAT, AS-VMAT); this was performed without changing the parameter setting of the cost function for treatment plan optimization and other optimization parameters. These new treatment plans were compared to those optimised using manually segmented organ contours (manual segmentations VMAT, MS-VMAT) to identify the differences in dose to OARs. The dose differences to OARs were evaluated with the following parameters: $D_{1\%}$, $D_{2\%}$, Dmean, $D_{98\%}$, $D_{99\%}$, $V_{10}$, $V_{20}$, $V_{30}$, $V_{40}$, and $V_{50}$. Two assessment methods were adopted, and the specific compared items are shown in **Table 1**.

The SPSS25.0 software was used for statistical analysis, and the data were first tested for conformance with a normal distribution. The paired t-test was performed on the normally distributed data, while the Wilcoxon signed rank test was performed on the data that did not conform to a normal distribution. A P-value < 0.05 was considered to be statistically significant.

To evaluate the sensitivity in the detection of dosimetry differences, the dosimetry results for manually segmented organ contours in the MS-VMAT plans were used to define the 95% confidence intervals and the cut-off values of the parameters evaluating the dosimetry differences in OARs. The number of cases in which the dosimetry results for the auto-segmented organ contours/manually segmented organ contours were outside the confidence interval in the AS-VMAT plan was calculated. The SPSS 25.0 statistical software was used to calculate the 95% confidence interval of the evaluation parameters (Formula 1).

$$CL_{M_C} = mean \pm 1.96\sigma \tag{1}$$

where *mean* represents the mean value of the evaluation parameter; $\sigma$ denotes the corresponding standard deviation; and *CL* denotes the confidence interval.

To evaluate the CTV coverage, the percent coverage of CTV $V_{42.75}$ and CTV $V_{45}$ in the AS-VMAT plans and the MS-VMAT plans was evaluated. The dose distributions in the AS-VMAT plans were compared to those in the MS-VMAT plans to evaluate the differences in CTV gamma passing rates (2%/2 mm and 3%/3 mm criteria). The threshold dose was set at 95% of the prescription dose, because in clinical practice, more attention is paid to tumour control and normal tissue toxicity in high-dose areas (33).

# 3 RESULTS

## 3.1 Results of the Evaluation of Geometrical Differences

**Figure 2** lists the DSC, MDA, HD and $HD_{95}$ between manual segmentation and automated segmentation for each organ at risk in the testing set. The mean DSC (range) between manual segmentation and automated segmentation for all organs at risk was 0.91 (0.82–0.96). The automated segmentation results

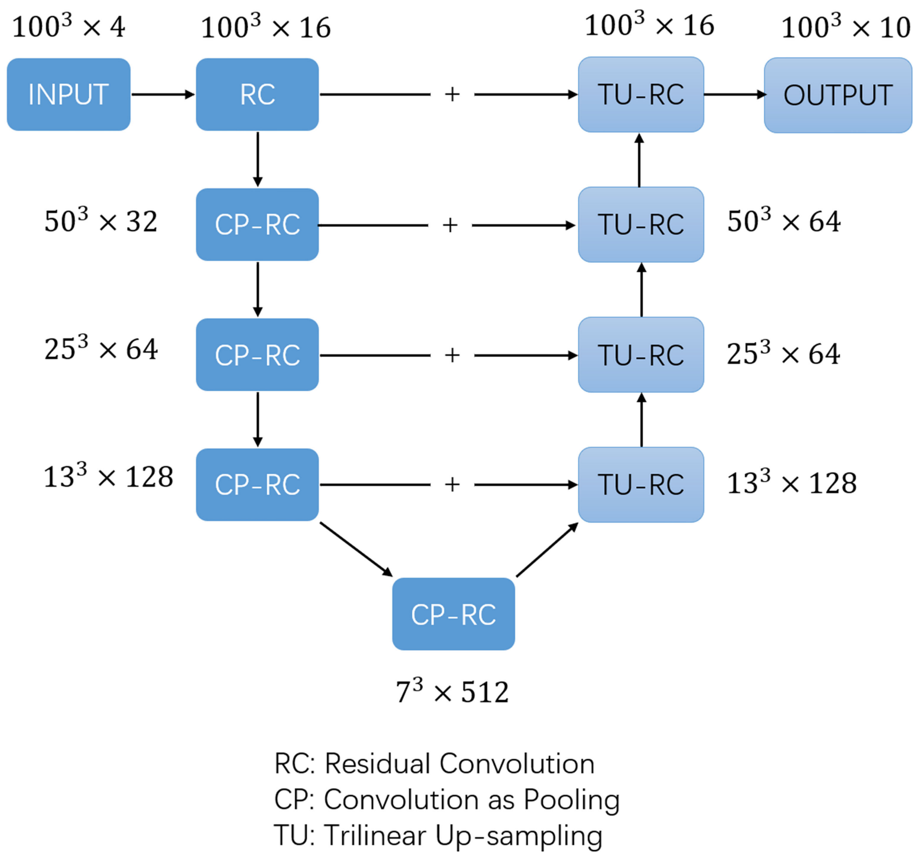


**FIGURE 1** | The structure of 3D U-net network.

were highly similar to those of the manual segmentation results in the bladder, the femoral head, the kidney, and the pelvic bone, with a mean DSC of > 0.94. The mean DSCs in the colon and the rectum were 0.82 and 0.83, respectively. The mean MDA, HD and $HD_{95}$ (range) between manual segmentation and automated segmentation for all organs at risk were 1.17 mm (0.45–3.21 mm), 11.73 mm (4.34–48.72 mm) and 5.32 (2.30–17.31 mm), respectively. The MDA and the $HD_{95}$ were the largest in the colon, with mean values of 3.21 ± 1.26 mm and 48.72 ± 12.60 mm, respectively.

## 3.2 Results of the Evaluation of Dose Differences

**Figure 3**; **Supplementary Table A** shows that compared to the dose distribution within manually segmented organ contours in the MS-VMAT plans of 22 patients, the $D_{99\%}$ within the auto-segmented rectum contours and the Dmean within the auto-segmented spinal cord contours in the AS-VMAT plans were higher by 6.04 Gy ($P = 0.037$) and 0.54 Gy ($P = 0.026$), respectively. The $V_{20}$, $V_{30}$, and $V_{40}$ in the rectum increased by 1.35% ($P = 0.027$), 1.73% ($P = 0.021$), and 1.96% ($P = 0.008$), respectively, whereas the $V_{10}$ in the spinal cord increased by 1.93% ($P = 0.011$). The differences in other dosimetry parameters were not statistically significant.

Based on the dose distribution within the manually segmented organ contours, the dose differences between the AS-VMAT plans and the MS-VMAT plans were relatively small. The $D_{99\%}$ in the rectum was higher by 0.64 Gy ($P = 0.292$), with no significant differences. The Dmean in the spinal cord was higher by 0.53 Gy ($P = 0.044$). The $V_{40}$ in the rectum

**TABLE 1** | Specific compared items in the evaluation of dose differences to organs at risk (two evaluation methods).

| Evaluation methods | ASAP vs. MSMP | | MSAP vs. MSMP | |
|---|---|---|---|---|
| Structures | Automatic segmentations | Manual segmentations | Manual segmentations | Manual segmentations |
| Plans | AS-VMAT plans | MS-VMAT plans | AS-VMAT plans | MS-VMAT plans |

*ASAP, Automatic Segmentation in AS-VMAT Plan; MSAP, Manual Segmentation in AS-VMAT Plan; MSMP, Manual Segmentation in MS-VMAT Plan; AS-VMAT, automatic segmentations VMAT; MS-VMAT, manual segmentations VMAT.*

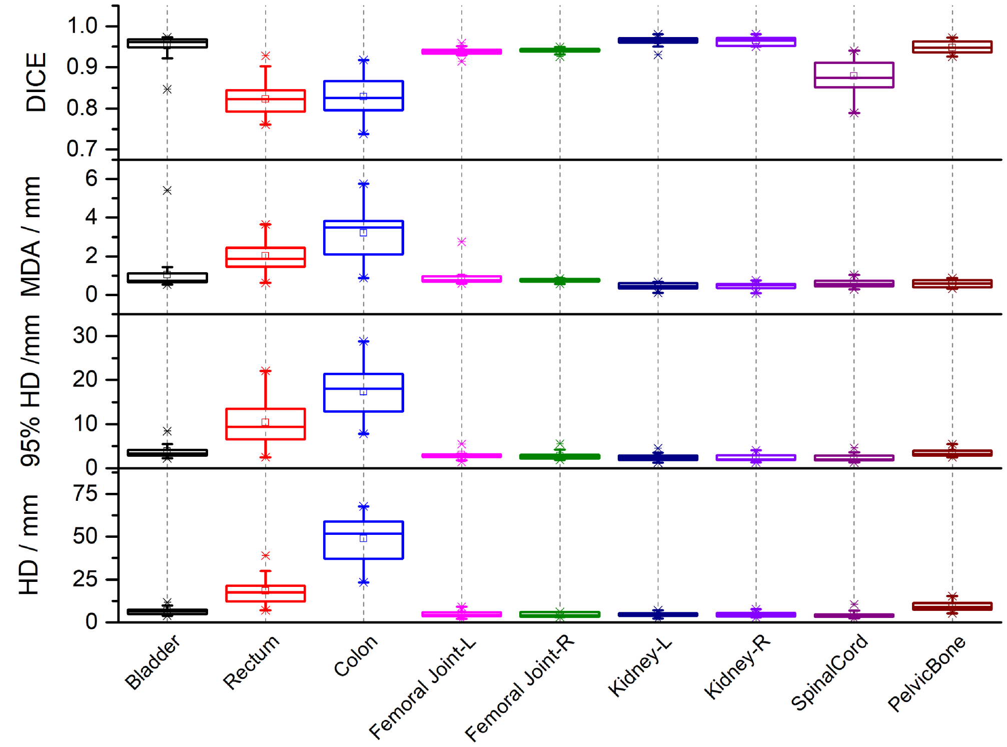


**FIGURE 2** | The Dice similarity coefficients (DSC), Mean Distance to Agreement (MDA), Hausdorff Distance (HD) and 95th-percentile of the HD between automated segmentation and manual segmentation for each organ at risk in the testing set.

increased by 1.00% (P = 0.034), while the $V_{10}$ and $V_{20}$ in the spinal cord increased by 1.76% (P = 0.015) and 1.59% (P = 0.015), respectively. The differences in other dosimetry parameters were not statistically significant.

The AS-VMAT plans of 22 cases were used to evaluate the sensitivity in the detection of dosimetry differences. Among the results for both automatically and manually segmented organ contours, the dosimetry results outside the confidence interval for the bladder ($D_{1\%,}$ $D_{2\%}$ and $V_{40}$) and the rectum ($D_{1\%}$ and $D_{2\%}$) were found in 2 cases each. Among the results for auto-segmented organ contours, the dosimetry results outside the confidence interval for the rectum ($V_{40}$), the colon ($D_{1\%}$ and $D_{2\%}$), the right femoral head ($V_{30}$), the left kidney (Dmean), and the pelvis (Dmean and $V_{30}$) were found in 1 case each. Among the results for manually segmented organ contours, the dosimetry results outside the confidence interval for the colon ($D_{1\%}$ and $D_{2\%}$), the right femoral head ($V_{30}$), the left and right kidneys (Dmean), and the pelvis (Dmean) were found in 1 case each, with a percentage outside the confidence interval of < 10%. No dosimetry results were outside the confidence interval for other evaluation parameters in any of the cases.

Regarding the evaluation of CTV coverage, in the AS-VMAT plans, the percent coverage of CTV V42.75 and CTV V45 was 99.86% ± 0.33% and 99.47% ± 1.67%, respectively, and the corresponding percent coverage in the MS-VMAT plans was 99.77% ± 0.75% and 99.53% ± 0.98%, respectively. The mean percent coverage of CTV V42.75 and the mean percent coverage of CTV V45 were higher by 0.09% (P = 0.453) and lower by 0.06% (P = 0.109), respectively in the AS-VMAT plans compared to the MS-VMAT plans. **Figure 4** shows the correspondence between the AS-VMAT plans and the MS-VMAT plans in terms of gamma passing rates in CTV. The mean gamma passing rates were 92.72% and 98.77%, respectively using the 2%/2 mm and 3%/3 mm criteria, which satisfied the clinical requirements.

**Figure 5** shows a comparison between the AS-VMAT plans and the MS-VMAT plans in terms of CTV and normal tissue DVHs for one patient with cervical cancer. There was no significant difference in the dose to CTV between the VMAT treatment plans optimised with manually segmented organ contours and those optimised with auto-segmented organ contours. There were insignificant dose deviations in normal tissue volume receiving < 30 Gy, such as in the bladder, the rectum, the pelvis, and the femoral head; there were dose deviations in the rectal volume receiving 30Gy–40Gy; and there was basically no dose difference to other normal tissues.

## 4 DISCUSSION

The convolutional neural network algorithm based on multi-layer supervised learning features good fault-tolerance, and strong adaptability and weight-sharing (13, 14, 34, 35). The results

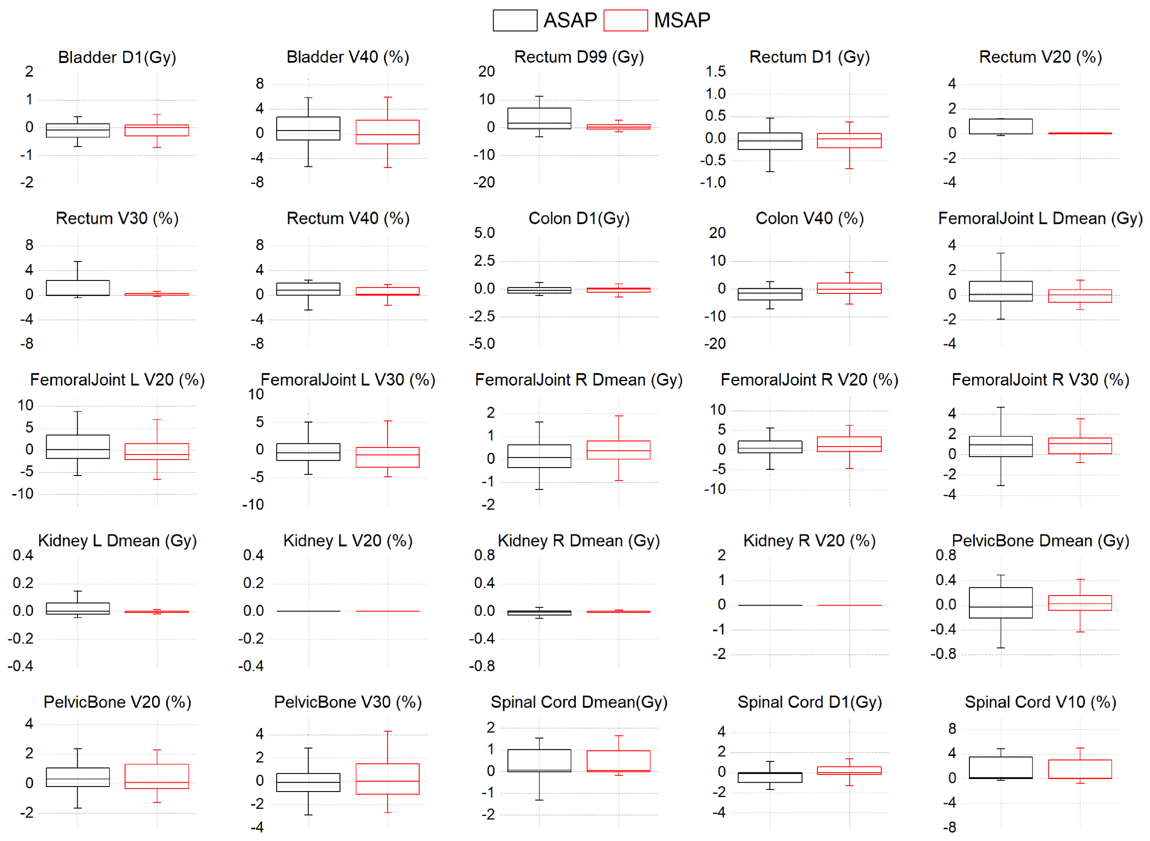


**FIGURE 3** | The dose differences of DVH parameters between the AS-VMAT plans and the MS-VMAT plans of 22 patients.The black box represents the ASAP vs. MSMP results, and the red box represents the MSAP vs. MSMP results. ASAP, Automatic Segmentation in AS-VMAT Plan; MSAP, Manual Segmentation in AS-VMAT Plan; MSMP, Manual Segmentation in MS-VMAT Plan.

generated by the trained model are reliable and applicable in clinical practice. We used the 3D U-net model for the auto-segmentation of nine types of normal tissues. The results suggested high geometric accuracy of automatic segmentation for the bladder, the femoral head, the pelvis, and the kidney, with a Dice value of > 0.94, which is consistent with, or even better than the results reported previously. The main reasons for this include the high density of bone structures (the pelvis and the femoral head) and strong tissue contrast. Indeed, the fluid-filled bladder can be easily distinguished from adjacent soft tissues, while there is a clear-cut anatomical position of the kidneys in the human body.

Relatively speaking, auto-segmentation of intestinal tissues, such as the colon and rectum, has a lower accuracy. Our results showed that auto-segmentation of the rectum and the colon featured a larger HD and a Dice value of 0.82 and 0.83 (< 0.9), respectively. Compared to previous results, Men et al. (14) reported a Dice value of 0.618 for the segmentation of the colon using a deep dilated convolutional neural network (DDCNN), which is lower than our study results; Rhee et al. (5) reported a Dice value of 0.80 for the segmentation of the rectum based on the CNN model, which is roughly equivalent to our study results; and Ju et al. (21) reported a Dice value of 0.87 for the segmentation of the rectum using an innovative fused model Dense V-Network, which is similar to our results. Generally, the Dice value for the segmentation of intestinal tissues can reach approximately 0.8 if proper neural networks and learning models are used (including 3D Unet and Dense-V-Network).

Auto-segmentation of intestinal tissues has a lower accuracy largely because the intestinal tract is a soft tissue with low-contrast image performance in CT images. For example, in terms of the rectum, the lower boundary of the rectum is connected to the anal canal and the boundary between the anal canal and the rectum is unclear on CT images, which makes it challenging to accurately identify the position of the lower boundary. In addition, the upper boundary of the rectum is connected to the sigmoid colon with an anatomical boundary between the rectum and the sigmoid colon, but this boundary is difficult to accurately identify *via* imaging. In terms of the colon, as we included patients treated in both the prone and the supine positions, and given that in some patients in the prone position, the position of the colon was pushed upward, the colon was not well distinguished from the pulmonary cavity and the aerated gastric body during auto-segmentation, resulting in segmentation failure. In addition, the accuracy of auto-segmentation of intestinal tissues is affected by the amount of faeces and gas in the intestines, which is a common problem with other automatic segmentation models when the intestinal organs

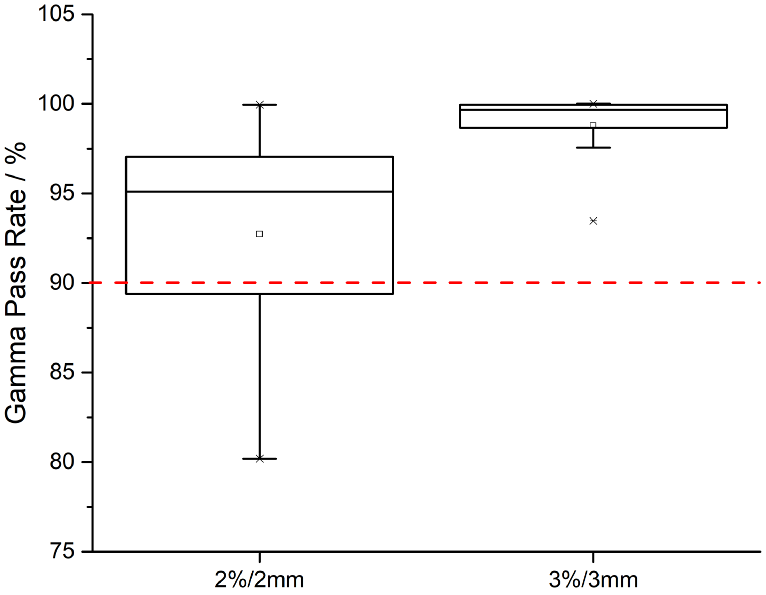


**FIGURE 4** | The correspondence between the AS-VMAT plans and the MS-VMAT plans in terms of gamma passing rates in clinical target volume (CTV). The red dotted line denotes a gamma passing rate of 90%. AS-VMAT, automatic segmentations VMAT; MS-VMAT, manual segmentations VMAT.

are segmented. The single learning model that we used is well suited to patients in different therapeutic positions, and there is no need to construct different learning models for supine positions and prone positions independently, which is why we included patients treated in different body positions.

We sought to determine whether we could directly use the unmodified normal tissue contours in the design of treatment plans given that the auto-segmented normal tissue contours are highly similar to manual segmentation results, and whether the dosimetry results in the optimised treatment plan satisfy the clinical requirements. As can be seen from the study results, irrespective of whether the treatment plans were optimised by auto-segmented or manually segmented normal tissue contours, the dose differences in the target volumes were relatively small (i.e., the doses to CTV were highly consistent). In this study, the gamma passing rate was adopted for quality assurance of treatment plans. Even when using the strict 2%/2 mm criterion, the gamma passing rates were > 90%, indicating that the dosimetry results are acceptable for clinical use.

As for the dose differences of automated segmentation of organs at risk, the situation is more complex and the organs at risk can be divided into three types:

1) The first type of organs, including the left and right femoral heads and the left and right kidneys, were located at a distance from the target volume, and automated segmentation of their contours was accurate. When these auto-segmented normal tissue contours were directly used for the design of treatment plans, the generated dosimetry parameters were not significantly different from those of the MS-VMAT plans. The spinal cord is an exception; although the spinal cord was located at a distance from the target volume and the auto-segmented contours were highly similar to those of manually segmented contours, the differences between the two sets of plans in terms of Dmean and $V_{10}$ in the spinal cord were statistically significant due to the excessively small volume of the spinal cord ($P < 0.05$). A common problem in cord segmentation was the length of cord contoured which adversely affected Dice and Dmean but had no clinical significance. Specifically, the absolute dose difference in Dmean was < 0.54 Gy and the volume difference to the $V_{10}$ was < 2%. Hence, these dosimetry results differences appear to be clinically acceptable, and the spinal cord is still classified as a type I organ at risk.

2) The second type of organs, including the pelvis and the colon, overlapped with the target volume on some CT slices. The volume of the overlap region accounted for a relatively small percentage of the total organ volume, so the geometrical differences in automated segmentation results did not result in large dose deviations and did not affect the dosimetry results in clinical evaluation. The obvious errors in automated organ segmentation need to be addressed and corrected, especially errors in areas close to the target volume. The abovementioned organs are classified as type II organs at risk.

3) The third type of organs, including the rectum and the bladder, were close to the target volume. The differences between the MS-VMAT plans and the AS-VMAT plans in terms of the $D_{98\%}$, $D_{99\%}$, V20, $V_{30}$, and $V_{40}$ in the rectum were statistically significant ($P < 0.05$). In addition, dosimetry results outside the confidence interval for the bladder ($D_{1\%,}$ $D_{2\%}$ and $V_{40}$) and the rectum ($D_{1\%}$ and $D_{2\%}$) were found in 2 cases each. This may be because the rectum and the bladder were close to the CTV, even overlapping in some regions (as shown in **Figure 6**). Hence, the geometrical differences in automated segmentation results had a significant impact on the dose received by high-dose areas. Meanwhile, the dosimetry results were more sensitive to the geometric accuracy of automated contouring due to the relatively small volume of the rectum. Therefore, auto-segmented organ contours need to be carefully checked, with the errors corrected. The abovementioned organs are classified as type III organs at risk.

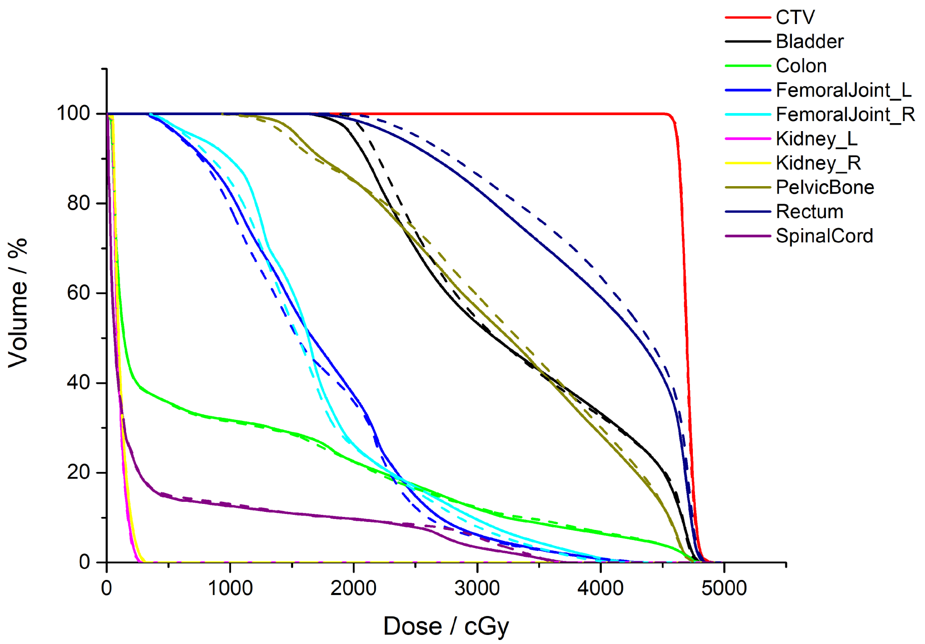


**FIGURE 5** | Comparison of the clinical target volume (CTV) and normal tissue DVHs for one patient with cervical cancer. The dotted lines denote the treatment plans optimised by auto-segmented organ contours (AS-VMAT), whereas the solid lines denote the treatment plans optimised by manually segmented organ contours (MS-VMAT). AS-VMAT, automatic segmentations VMAT; MS-VMAT, manual segmentations VMAT.

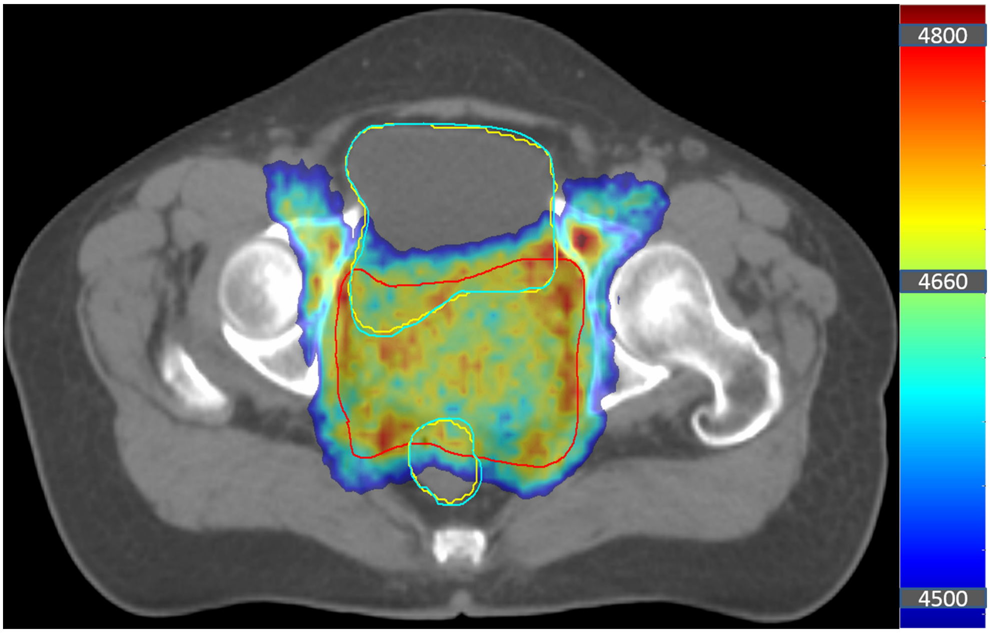


**FIGURE 6** | Diagram showing the position of the target volume, the bladder, and the rectum in a patient with cervical cancer. The coloured area denotes the target volume receiving > 45 Gy. The area marked with the red solid line is the clinical target volume (CTV), the blue solid lines denote the manually segmented contours, and the yellow solid lines denote the auto-segmented contours.

Moreover, another factor contributing to the difference in planned dose lies in the treatment planning system. During the course of the study, we found that after the treatment plan was optimised twice under identical optimization conditions for the same patient (same structures and CT images) in the Monaco system, the generated sequences of the sub-fields and the positions of the leaves were not entirely consistent, which resulted in significant differences in the dose distribution within low-dose areas.

We attempted to segment the normal tissues in patients with cervical cancer using deep learning techniques. In addition, we attempted to analyse which tissues received significantly different doses when automated segmentation results with high geometric accuracy were directly used in the design of treatment plans. Based on the results, we classified the auto-segmented normal tissues into three types. The auto-segmentation results for some tissues need to be carefully checked and corrected, while the auto-segmented contours of other tissues can be almost left unmodified, thereby saving clinicians a significant amount of time (an important objective of this study). A similar finding was reported by Vaasen et al. (36), that most OARs can be left unedited except under certain circumstances where they were close to the planning target volume. However, this study still has its limitations. First, the size of the samples from the testing set was too small to accurately evaluate the dose differences and larger sample sizes will provide more statistically significant results. Second, the analysed patients were collected from the same medical centre and no multicentre comparison was performed. Conclusions based on multicentre studies would be more objective and compelling.

## 5 CONCLUSIONS

The 3D U-net model can be used for accurate, efficient, and automated segmentation of organs at risk in patients with cervical cancer. When auto-segmented organ contours were used in the design of treatment plans, the dose distributions of target volumes were not affected, whereas the impact of automated segmentation on the doses to organs at risk was complicated. We suggest that the auto-segmented contours of tissues in close proximity to the target volume need to be carefully checked and corrected when necessary, while auto-segmented contours of tissues at a distance from the target volume can be left largely unmodified.

## DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://www.researchdata.org.cn/RDDA2022345612.

## AUTHOR CONTRIBUTIONS

AC and FC are responsible for data analysis and paper writing. XL, YZ, and LC are responsible for manually delineating the organs, and for data collection. JZ is responsible for program design and model training. JZ and LXC are responsible for are responsible for research strategy design and paper revision. All authors contributed to the article and approved the submitted version.

## FUNDING

This study was jointly supported by the National Natural Science Foundation of China (No. 12005315) and the National Natural Science Foundation of China (No.12075329).

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fonc.2022.908903/full#supplementary-material

## REFERENCES

1. Cardenas CE, Yang J, Anderson BM, Court LE, Brock KB. Advances in Auto-Segmentation. *Semin Radiat Oncol* (2019) 29(3):185–97. doi: 10.1016/j.semradonc.2019.02.001
2. Shi J, Ding X, Liu X, Li Y, Liang W, Wu J. Automatic Clinical Target Volume Delineation for Cervical Cancer in CT Images Using Deep Learning. *Med Phys* (2021) 48(7):3968–81. doi: 10.1002/mp.14898
3. Ju Z, Guo W, Gu S, Zhou J, Yang W, Cong X, et al. CT Based Automatic Clinical Target Volume Delineation Using a Dense-Fully Connected Convolution Network for Cervical Cancer Radiation Therapy. *BMC Cancer* (2021) 21(1):243. doi: 10.1186/s12885-020-07595-6
4. Jamtheim Gustafsson C, Lempart M, Swärd J, Persson E, Nyholm T, Thellenberg Karlsson C, et al. Deep Learning-Based Classification and Structure Name Standardization for Organ at Risk and Target Delineations in Prostate Cancer Radiotherapy. *J Appl Clin Med Phys* (2021) 22(12):51–63. doi: 10.1002/acm2.13446
5. Rhee DJ, Jhingran A, Rigaud B, Netherton T, Cardenas CE, Zhang L, et al. Automatic Contouring System for Cervical Cancer Using Convolutional Neural Networks. *Med Phys* (2020) 47(11):5648–58. doi: 10.1002/mp.14467
6. Ibragimov B, Xing L. Segmentation of Organs-at-Risks in Head and Neck CT Images Using Convolutional Neural Networks. *Med Phys* (2017) 44(2):547–57. doi: 10.1002/mp.12045
7. Cardenas CE, McCarroll RE, Court LE, Elgohari BA, Elhalawani H, Fuller CD, et al. Deep Learning Algorithm for Auto-Delineation of High-Risk Oropharyngeal Clinical Target Volumes With Built-In Dice Similarity Coefficient Parameter Optimization Function. *Int J Radiat Oncol Biol Phys* (2018) 101(2):468–78. doi: 10.1016/j.ijrobp.2018.01.114
8. Feng X, Qing K, Tustison NJ, Meyer CH, Chen Q. Deep Convolutional Neural Network for Segmentation of Thoracic Organs-at-Risk Using Cropped 3D Images. *Med Phys* (2019) 46(5):2169–80. doi: 10.1002/mp.13466
9. Lustberg T, van Soest J, Gooding M, Peressutti D, Aljabar P, van der Stoep J, et al. Clinical Evaluation of Atlas and Deep Learning Based Automatic Contouring for Lung Cancer. *Radiother. Oncol* (2018) 126(2):312–7. doi: 10.1016/j.radonc.2017.11.012
10. Zhou X, Takayama R, Wang S, Hara T, Fujita H. Deep Learning of the Sectional Appearances of 3D CT Images for Anatomical Structure

Segmentation Based on an FCN Voting Method. *Med Phys* (2017) 44 (10):5221–33. doi: 10.1002/mp.12480

11. Hu P, Wu F, Peng J, Bao Y, Chen F, Kong D. Automatic Abdominal Multi-Organ Segmentation Using Deep Convolutional Neural Network and Time-Implicit Level Sets. *Int J Comput Assist Radiol Surg* (2017) 12(3):399–411. doi: 10.1007/s11548-016-1501-5
12. Rigaud B, Anderson BM, Yu ZH, Gobeli M, Cazoulat G, Söderberg J, et al. Automatic Segmentation Using Deep Learning to Enable Online Dose Optimization During Adaptive Radiation Therapy of Cervical Cancer. *Int J Radiat Oncol Biol Phys* (2021) 109(4):1096–110. doi: 10.1016/j.ijrobp.2020.10.038
13. Balagopal A, Kazemifar S, Nguyen D, Lin MH, Hannan R, Owrangi A, et al. Fully Automated Organ Segmentation in Male Pelvic CT Images. *Phys Med Biol* (2018) 63(24):245015. doi: 10.1088/1361-6560/aaf11c
14. Men K, Dai J, Li Y. Automatic Segmentation of the Clinical Target Volume and Organs at Risk in the Planning CT for Rectal Cancer Using Deep Dilated Convolutional Neural Networks. *Med Phys* (2017) 44(12):6377–89. doi: 10.1002/mp.12602
15. Cheng R, Roth HR, Lay N, Lu L, Turkbey B, Gandler W, et al. Automatic Magnetic Resonance Prostate Segmentation by Deep Learning With Holistically Nested Networks. *J Med Imaging (Bellingham)* (2017) 4 (4):041302. doi: 10.1117/1.JMI.4.4.041302
16. Wang Z, Chang Y, Peng Z, Lv Y, Shi W, Wang F, et al. Evaluation of Deep Learning-Based Auto-Segmentation Algorithms for Delineating Clinical Target Volume and Organs at Risk Involving Data for 125 Cervical Cancer Patients. *J Appl Clin Med Phys* (2020) 21(12):272–9. doi: 10.1002/acm2.13097
17. Berger T, Seppenwoolde Y, Pötter R, Assenholt MS, Lindegaard JC, Nout RA, et al. Importance of Technique, Target Selection, Contouring, Dose Prescription, and Dose-Planning in External Beam Radiation Therapy for Cervical Cancer: Evolution of Practice From EMBRACE-I to II. *Int J Radiat Oncol Biol Phys* (2019) 104(4):885–94. doi: 10.1016/j.ijrobp.2019.03.020
18. Duane FK, Langan B, Gillham C, Walsh L, Rangaswamy G, Lyons C, et al. Impact of Delineation Uncertainties on Dose to Organs at Risk in CT-Guided Intracavitary Brachytherapy. *Brachytherapy.* (2014) 13(2):210–8. doi: 10.1016/j.brachy.2013.08.010
19. Nelms BE, Tomé WA, Robinson G, Wheeler J. Variations in the Contouring of Organs at Risk: Test Case From a Patient With Oropharyngeal Cancer. *Int J Radiat Oncol Biol Phys* (2012) 82(1):368–78. doi: 10.1016/j.ijrobp.2010.10.019
20. Liu Z, Liu X, Xiao B, Wang S, Miao Z, Sun Y, et al. Segmentation of Organs-at-Risk in Cervical Cancer CT Images With a Convolutional Neural Network. *Phys Med* (2020) 69:184–91. doi: 10.1016/j.ejmp.2019.12.008
21. Ju Z, Wu Q, Yang W, Gu S, Guo W, Wang J, et al. Automatic Segmentation of Pelvic Organs-at-Risk Using a Fusion Network Model Based on Limited Training Samples. *Acta Oncol* (2020) 59(8):933–9. doi: 10.1080/0284186X.2020.1775290
22. Fung NTC, Hung WM, Sze CK, Lee MCH, Ng WT. Automatic Segmentation for Adaptive Planning in Nasopharyngeal Carcinoma IMRT: Time, Geometrical, and Dosimetric Analysis. *Med Dosim.* (2020) 45(1):60–5. doi: 10.1016/j.meddos.2019.06.002
23. Zhu J, Chen X, Yang B, Bi N, Zhang T, Men K, et al. Evaluation of Automatic Segmentation Model With Dosimetric Metrics for Radiotherapy of Esophageal Cancer. *Front Oncol* (2020) 10:564737. doi: 10.3389/fonc.2020.564737
24. Vinod SK, Jameson MG, Min M, Holloway LC. Uncertainties in Volume Delineation in Radiation Oncology: A Systematic Review and Recommendations for Future Studies. *Radiother. Oncol* (2016) 121(2):169–79. doi: 10.1016/j.radonc.2016.09.009
25. Gay HA, Barthold HJ, O'Meara E, Bosch WR, El Naqa I, Al-Lozi R, et al. Pelvic Normal Tissue Contouring Guidelines for Radiation Therapy: A Radiation Therapy Oncology Group Consensus Panel Atlas. *Int J Radiat Oncol Biol Phys* (2012) 83(3):e353–62. doi: 10.1016/j.ijrobp.2012.01.023
26. Smistad E, Østvik A, Haugen BO, Lvstakken L. (2017). 2D Left Ventricle Segmentation Using Deep Learning, in: *2017 IEEE International Ultrasonics Symposium (IUS)* Washington, DC, USA: IEEE Press. pp. 1–4.
27. Sørensen TJ. *A Method of Establishing Groups of Equal Amplitude in Plant Sociology Based on Similarity of Species Content and Its Application to Analyses of the Vegetation on Danish Commons*. København: I kommission hos E. Munksgaard (1948).
28. Loshchilov I, Hutter F. Decoupled Weight Decay Regularization. *arXiv [Preprint]* (2017). Available at: https://arxiv.org/abs/1711.05101 (Accessed Jan 4, 2019).
29. Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. *CoRR* (2014) arXiv[Preprint]. Available at: https://arxiv.org/abs/1412.6980 (Accessed Jan 30, 2017). doi: 10.48550/arXiv.1412.6980
30. Smith LN, Topin N. Super-Convergence: Very Fast Training of Neural Networks Using Large Learning Rates. *arXiv[Preprint]* (2019). Available at: https://arxiv.org/abs/1708.07120v3 (Accessed May 17 2018)
31. Zhu J, Zhang J, Qiu B, Liu Y, Liu X, Chen L. Comparison of the Automatic Segmentation of Multiple Organs at Risk in CT Images of Lung Cancer Between Deep Convolutional Neural Network-Based and Atlas-Based Techniques. *Acta Oncologica* (2019) 58(2):257–64. doi: 10.1080/0284186X.2018.1529421
32. Pierfrancesco F, Francesca A, Elisabetta T, Elena G, Stefania M, Carlo IG, et al. Variability of Clinical Target Volume Delineation for Rectal Cancer Patients Planned for Neoadjuvant Radiotherapy With the Aid of the Platform Anatom-E. *Clin Trans Radiat Oncol* (2018) 11:33–9. doi: 10.1016/j.ctro.2018.06.002
33. Jensen N, Ptter R, Spampinato S, Fokdal LU, Tanderup K. Dose-Volume Effects and Risk Factors for Late Diarrhea in Cervix Cancer Patients After Radiochemotherapy With Image Guided Adaptive Brachytherapy in the EMBRACE I Study. *Int J Radiat Oncol Biol Phys* (2020) 109(3):688–700. doi: 10.1016/j.ijrobp.2020.10.006
34. Gibson E, Giganti F, Hu Y, Bonmati E, Bandula S, Gurusamy K, et al. Automatic Multi-Organ Segmentation on Abdominal CT With Dense V-Networks. *IEEE Trans Med Imaging* (2018) 37(8):1822–34. doi: 10.1109/TMI.2018.2806309
35. Kazemilar S, Balagopal A, Nguyen D, McGuire, Hannan R. Segmentation of the Prostate and Organs at Risk in Male Pelvic CT Images Using Deep Learning. *Biomed Physics Eng Express* (2018) 4(5):055003. doi: 10.1088/2057-1976/aad100
36. Vaassen F, Hazelaar C, Canters R, Peeters S, Petit S, van Elmpt W. The Impact of Organ-at-Risk Contour Variations on Automatically Generated Treatment Plans for NSCLC. *Radiother Oncol* (2021) 163:136–42. doi: 10.1016/j.radonc.2021.08.014

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

**Publisher's Note:** All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Oncology

ORIGINAL RESEARCH
published: 08 June 2022
doi: 10.3389/fonc.2022.910792

# Impact of Using Unedited CT-Based DIR-Propagated Autocontours on Online ART for Pancreatic SBRT

*Alba Magallon-Baro**, *Maaike T. W. Milder*, *Patrick V. Granton*, *Wilhelm den Toom*, *Joost J. Nuyttens* and *Mischa S. Hoogeman*

*Department of Radiotherapy, Erasmus MC Cancer Institute, University Medical Center Rotterdam, Rotterdam, Netherlands*

**Purpose:** To determine the dosimetric impact of using unedited autocontours in daily plan adaptation of patients with locally advanced pancreatic cancer (LAPC) treated with stereotactic body radiotherapy using tumor tracking.

**Materials and Methods:** The study included 98 daily CT scans of 35 LAPC patients. All scans were manually contoured (MAN), and included the PTV and main organs-at-risk (OAR): stomach, duodenum and bowel. Precision and MIM deformable image registration (DIR) methods followed by contour propagation were used to generate autocontour sets on the daily CT scans. Autocontours remained unedited, and were compared to MAN on the whole organs and at 3, 1 and 0.5 cm from the PTV. Manual and autocontoured OAR were used to generate daily plans using the VOLO$^{\text{TM}}$ optimizer, and were compared to non-adapted plans. Resulting planned doses were compared based on PTV coverage and OAR dose-constraints.

**Results:** Overall, both algorithms reported a high agreement between unclipped MAN and autocontours, but showed worse results when being evaluated on the clipped structures at 1 cm and 0.5 cm from the PTV. Replanning with unedited autocontours resulted in better OAR sparing than non-adapted plans for 95% and 84% plans optimized using Precision and MIM autocontours, respectively, and obeyed OAR constraints in 64% and 56% of replans.

**Conclusion:** For the majority of fractions, manual correction of autocontours could be avoided or be limited to the region closest to the PTV. This practice could further reduce the overall timings of adaptive radiotherapy workflows for patients with LAPC.

**Keywords: pancreas, SBRT, adaptive, replanning, autocontouring**

OPEN ACCESS

***Edited by:***
*Nesrin Dogan,*
*University of Miami, United States*

***Reviewed by:***
*Rodrigo Delgadillo,*
*University of Miami Health System,*
*United States*
*Yang Kyun Park,*
*University of Texas Southwestern*
*Medical Center, United States*

****Correspondence:***
*Alba Magallon-Baro*
*a.magallonbaro@erasmusmc.nl*

***Specialty section:***
*This article was submitted to*
*Radiation Oncology,*
*a section of the journal*
*Frontiers in Oncology*

***Received:*** *01 April 2022*
***Accepted:*** *19 May 2022*
***Published:*** *08 June 2022*

***Citation:***
*Magallon-Baro A, Milder MTW, Granton PV, Toom Wd, Nuyttens JJ and Hoogeman MS (2022) Impact of Using Unedited CT-Based DIR-Propagated Autocontours on Online ART for Pancreatic SBRT.*
*Front. Oncol. 12:910792.*
*doi: 10.3389/fonc.2022.910792*

## INTRODUCTION

Adaptive radiotherapy (ART) is a desired paradigm in radiation therapy. Its goal is to adjust the treatment plan to the patient anatomy of the day to compensate for anatomical changes (1, 2). An online ART workflow has to be time efficient as the patient awaits treatment (1, 3). In recent years, efforts have been focused on speeding up the ART process through fast treatment plan reoptimization techniques and through automatically segmenting anatomical structures in medical images (3–10). The latter aims to reduce delineation times, which in ART remains a crucial point since contouring has been traditionally performed manually by dedicated and trained staff (11).

Carcinomas located close to radiosensitive and mobile organs-at-risk (OAR), such as unresectable locally advanced pancreatic cancer (LAPC), are excellent candidates for ART (4, 8, 9, 12). LAPC is a dose-limited tumor type, whose dosage is often compromised to protect surrounding organs. To manage this limitation, stereotactic body radiotherapy (SBRT) has become a standard of care for LAPC, owing to its capability to deliver highly conformal doses with steep dose gradients (13–17). Nonetheless, due to day-to-day OAR mobility, unintended doses are received by OAR close to the tumor (3, 18). For that reason, ART is recently being explored for LAPC patients using systems such as the MRIdian (ViewRay, Oakwook Village, OH) (8, 9, 12, 19, 20), the Elekta Unity (Elekta AB, Stockholm, Sweden) (7, 9, 21), or the Ethos (Varian Medical Systems Inc, Palo Alto, CA) (22, 23).

In our clinic, LAPC patients are treated on the CyberKnife (CK) (Accuray Inc, Sunnyvale, USA) using real-time tracking (24, 25). The CK does not have an integrated 3D imaging system, but our institute has a unique CT-on-rails in the treatment room that allows daily imaging (26). Our previous work investigated the potential trade-offs of applying different fast and quasi-automated plan adaptation methods on the CK (6). Nonetheless, a major challenge remains in laborious daily organ delineation, i.e. contouring.

Automatic contouring methods may offer a solution and are often based on the propagation of contours from the planning (pCT) to the fraction CT (FxCT) through deformable image registration (DIR) (2–4, 7). The use of automatic algorithms not only speeds up this task, but could also offer consistency to limit intra- and inter-observed variations. However, due to poor soft tissue contrast in the abdominal area, autosegmented organ contours (i.e. autocontours) generally require further manually editing before being used for daily replanning purposes (3, 27). Within an ART framework, manual delineation is one of the most time-consuming steps, but is thought to be essential to guarantee the quality of the adapted treatment plan. The time required for delineation delays the start of radiation delivery, and allows for additional intra-fraction OAR motion to occur, which can devaluate further the adapted plan. For this reason, in this study we have explored if manual editing of daily contours can be avoided while replanning. We have investigated the impact of using unedited autocontours generated with two commercially DIR algorithms available in Precision$^{TP}$ (Accuray Inc, Sunnyvale, USA) and in MIM (MIM Software Inc, Cleveland, USA). The value of replanning directly on unedited autocontours has been established by: (a) comparing resulting plans to replans obtained using manual contours in the optimization, and (b) comparing them to conventional non-adapted SBRT plans. In addition, we also quantified the geometric accuracy of both DIR algorithms, especially close to the target volume.

# MATERIALS AND METHODS

## Patient Data

A total of 35 patients with pancreatic cancer were included in this study. All patients were diagnosed with inoperable nonmetastatic LAPC, and presented a stable disease after receiving 8 cycles of chemotherapy (FOLFIRINOX). They received subsequent hypofractionated SBRT treatment of 40 Gy in 5 fractions, prescribed to the 80% isodose line. Patients gave informed consent to be included in the LAPC-1 Phase II study, which was approved by the Institutional Review Board (ID: NL49643.078.14) in accordance with the recommendations of the Declaration of Helsinki.

The study protocol indicated that each patient received a planning CT (pCT) and 3 contrast-enhanced in-room daily scans under instructed end-expiration breath-hold prior to treatment delivery (FxCT). All scans were acquired after manually injecting intravenous contrast agent, and by immobilizing patients using a vacuum bag on the treatment couch. Patients were recommended to avoid food and drink intake 2 h before the treatment fraction. In total, 98 FxCT were collected in this cohort, since only 2 daily CTs were available for 7 out of 35 patients.

The pCTs were delineated by a radiation oncologist (with 10+ years of experience) following the RTOG guidelines on the abdominal region (28). The gross tumor volume (GTV) was expanded by 5 mm to generate the clinical target volume (CTV), which was subsequently expanded by 2 mm to create the planning target volume (PTV). Additionally, the main organs-at-risk (stomach, duodenum, bowel, kidneys and liver) were also manually contoured.

Patients were treated using the CyberKnife M6 system with synchrony respiratory motion tracking on pre-implanted gold fiducial markers (24, 25, 29). Each patient had a median of 3 fiducials in or around the pancreatic tumor. The clinical protocol stated that 95% of the PTV should receive 95% of the prescribed dose (i.e., 40 Gy/5 fx), although PTV underdosage was allowed to fulfill OAR constraints. The stomach, duodenum and bowel had a near-maximum dose constraint of V35 Gy < 0.5 cc. For the liver, dose-constraint was V20 Gy < 700 cc, for the kidneys, mean dose < 15 Gy and V15 Gy < 30%, and for the spinal cord, allowed max dose was < 27.5 Gy.

## Delineations on the Daily Scans

### Baseline of Manual Contour Set

FxCTs were delineated by the same radiation oncologist that delineated the pCT scans. The GTV and PTV were rigidly transferred to FxCTs after applying a fiducial pre-match. Additional details regarding OAR delineations can be seen in (30).

### Autocontour Sets

Contours from the pCT were propagated to FxCTs using the deformable image registration (DIR) algorithm available in both Precision$^{TP}$ (version 2.0.1.1) and MIM (version 6.9.3). A summary of each DIR method is available in **Supplementary Materials (A)**, as well as the procedure followed for parameter selection in MIM DIR. Whereas MIM DIR settings could be tuned to optimize the resulting contours for our dataset, Precision DIR settings are fixed and cannot be modified. The autocontours (AUTO) obtained using Precision DIR (asPREC) and MIM DIR (asMIM) remained unedited.

### Contour Sets Geometrical Comparison

Both autocontours sets (asPREC and asMIM) were geometrically compared to MAN through the Dice coefficient (DC) (which describes the overlapping ratio between two volumes), mean surface distance (MSD), Hausdorff distance (HD) (which describes the maximum distance between two contour surfaces) and volumetric difference (VOL_DIFF) between the automatic vs. manual contours. These 4 accuracy metrics complement each other by giving an indication of the volumetric error and the distance between the structures boundaries, as recommended in Sharp et al. (2) and AAPM TG-132 (31). All metrics were collected using an in-house algorithm. Most of these metrics present a skewed distribution, and hence, median and interquartile range (IQR) parameters describing the data spread between quartile 1 (Q1) and 3 (Q3) (i.e. the 25% and 75% percentiles in which the distribution lies), were collected for the subsequent comparison analysis.

MAN, asPREC and asMIM stomach, duodenum and bowel structures (the closest OAR to the target and mostly located within the high dose region), were clipped at 3, 1 and 0.5 cm from the PTV for geometrical comparison (4–6). The resulting asPREC and asMIM clipped organs were compared to MAN clipped structures by means of DC, MSD, HD and VOL_DIFF metrics.

Since the three gastrointestinal (GIO) organs (i.e. stomach, duodenum and bowel) have the same dose-constraints in the clinical protocol, a structure combining the three was created at each different scenario (whole and clipped GIO at 3, 1 and 0.5 cm). GIO structures were also compared using DC, MSD, HD and VOL_DIFF. No recommendations on a combined GIO structure are included in the clinical protocol. The GIO structure was only created to evaluate the geometrical similarity of the combined organs, while minimizing the effect of registration errors in the transition between organs (e.g. stomach to duodenum).

The minimum distance (MIN_DIST) from GTV and PTV to OARs and the overlapping volume (OVLP) of the expanded PTV (with 0.5 and 1 cm) with the OAR was also retrieved for MAN, asPREC and asMIM.

## Replanning on MAN, asPREC and asMIM Contours

Treatment plans were optimized using the VOLO$^{\text{TM}}$ optimizer in Precision$^{\text{TP}}$ (v2.0.1.1). As detailed in (6), a fast patient-specific template, including all clinically optimal cost functions used in the pCT, was generated. These fast templates reproduced the delivered clinical plans, while using a reduced number of nodes and OAR clipped at 3 cm from the PTV. These parameter combinations significantly reduced plan optimization times (6).

The patient-specific templates were used to perform an automated full inverse planning on the pCT. These planning doses were rigidly transferred to FxCTs to evaluate non-adapted (NoAd) doses. We transferred the dose to the FxCT rather than recalculating it, as in our previous work (6) we saw clinically irrelevant dose differences in the OAR and in the target volumes when comparing transferred and recalculated plans. Next, the template was used to perform a new automated full inverse planning on the FxCT to generate adapted plans using the clipped MAN, asPREC and asMIM at 3 cm. The resulting adapted plans are referenced hereafter as MAN_Rp, asPREC_Rp and asMIM_Rp, respectively. **Figure 1** shows an example patient with the 4 planned doses that were created and evaluated on the FxCT scan, as well as the contours used to optimize each different plan.

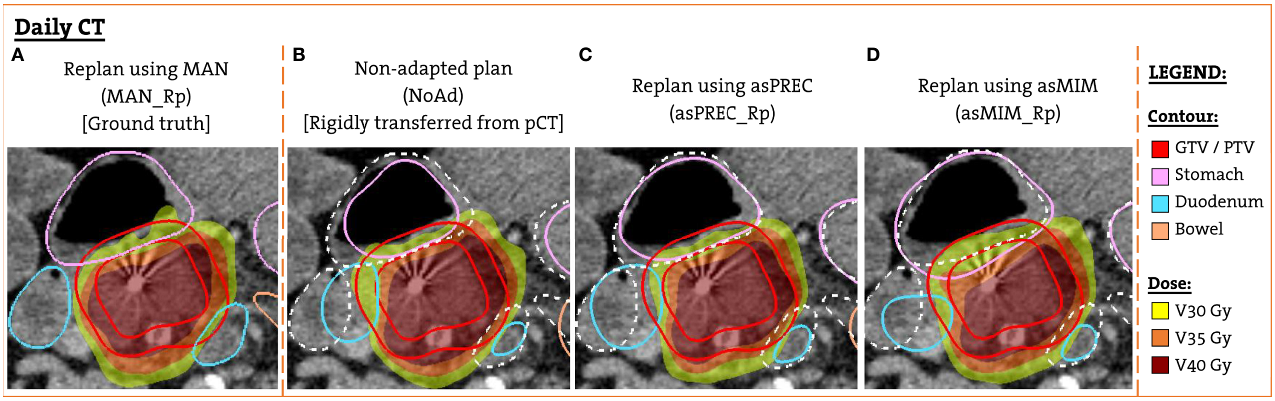


**FIGURE 1** | Example patient FxCT scan with the different structure set and dose distribution used for the dosimetric evaluation. **(A)** Replanned dose optimized using manual contours (ground truth). **(B)** Non-adapted dose with planning anatomy rigidly transferred from the pCT (solid lines). **(C)** Replanned dose optimized using contours obtained with Precision DIR (solid lines). **(D)** Replanned dose using contours from MIM DIR (solid lines). For **(B–D)** manual contours are also overlaid (dashed white lines).

## Dosimetric Plan Comparison

The four resulting doses in the FxCT scans (NoAd, MAN_Rp, asPREC_Rp and asMIM_Rp) were compared based on coverage, mean and minimum doses of the GTV and PTV, and near-maximum dose constraints (V35 < 0.5 cc) and mean doses of the OAR. All four doses were evaluated on the daily MAN contours during the subsequent dosimetric analysis, although plan optimization had been done using the planning contours (as in NoAd) or autocontours (as in asPREC_Rp and asMIM_Rp). Median and interquartile range (IQR) of these parameters were abstracted, and were compared using a two-sided Wilcoxon signed rank test, with a statistically significance defined by a p-value of < 0.05.

The following plan comparisons were performed. Firstly, replanned doses (MAN_Rp, asPREC_Rp and asMIM_Rp) were compared to non-adapted doses (NoAd) to determine the value of daily plan adaptation with respect to conventional planning. Secondly, replanned doses optimized using unedited autosegmented contours (asPREC_Rp and asMIM_Rp) were compared to replanned doses optimized using MAN, to determine the impact of inaccuracies in organ delineation on the replans.

To determine if autocontouring inaccuracies could be correlated with OAR constraints violations after replanning, the volumetric differences of auto vs. manual contours (i.e. VOL_DIFF) were compared between the fractions exceeding and the fractions not exceeding dose-constraints after replanning. VOL_DIFF was compared within different isotropic rings sets at different distances from the PTV: 0-1 vs 1-3 cm, 0-1.5 vs 1.5-3 cm, and 0-2 vs 2-3 cm. A Mann-Whitney test was performed to assess the differences between rings results. Statistical significance was set by a p-value < 0.05.

# RESULTS

## Contour Sets Geometrical Comparison

MAN, asPREC and asMIM contours were compared by means of DC, MSD, HD and VOL_DIFF on the whole (**Table B1**) and clipped OAR (**Figure 2** and **Table B2**), and by means of MIN_DIST and OVLP between target and OAR volumes (**Table 1**).

When evaluating the structures as a whole (**Table B1**), both algorithms reported high agreements between AUTO and MAN structures. A median (IQR: Q1, Q3) DC of 0.9 (0.9, 0.9), MSD of 2 (2, 3) mm, HD of 18 (15, 23) mm and VOL_DIFF of -1 (-16, 12) cc was observed for the combined GIO for asPREC, and a median DC of 0.9 (0.8, 0.9), MSD of 2 (2, 3) mm, HD of 19 (16, 23) mm and VOL_DIFF of 13 (-6, 27) cc for asMIM. The liver and kidneys were the organs reporting best results in both methods, and the bowel the worst, followed by the stomach and the duodenum.

When evaluating the clipped OAR at different distances from the PTV (**Figure 2** and **Table B2**), only the stomach, duodenum, bowel, and the combined GIO structure were considered. AUTO bowel contours were the structures showing less agreement with MAN bowels, followed by the duodenum and finally the stomach. Bowel contours reported the lowest DC, and larger MSD, HD and VOL_DIFF. The GIO structure generally outperformed individual organ measurements.

The DC in the 4 structures (i.e. stomach, duodenum, bowel and GIO) decreased closer to the PTV. Depending on the structure and method, DC ranged from 0.7 to 0.9 at 3 cm, and reduced to 0.5 to 0.8 at 0.5 cm distance from the PTV. The MSD showed little change at the 3 distances from the PTV, oscillating between 1 to 2 mm depending on the structure. The HD decreased for all structures when evaluated at 3 and 1 cm away of the PTV, reducing from a median of 18 to 13 mm in the GIO, but remained similar between 1 and 0.5 cm. Finally, the VOL_DIFF of AUTO vs. MAN reported similar volumes between MAN and asPREC. Conversely, asMIM showed positive differences compared to MAN ranging from 17 to 2 cc between 3 to 0.5 cm.

Generally, asPREC reported higher agreement with MAN than asMIM. As observed in **Figure 2** and **Table B2**, stomachs and bowels segmented with MIM were overestimated (i.e., positive VOL_DIFF), whereas with Precision both organs were slightly underestimated (i.e., negative VOL_DIFF). Both algorithms slightly underestimated the duodenum. Similar tendencies are observed in **Table 1**, in which asMIM reported smaller MIN_DIST to both GTV and PTV compared to MAN and asPREC, and also reported higher OVLP with the expanded PTV structure with autosegmented OAR.

## Dosimetric Comparison After Replanning

**Table 2** summarizes the dosimetric measurements performed in the non-adapted and adapted plans according to the different daily contours. After evaluating planned doses (NoAd) on MAN, 71% (70/98) of the plans resulted in OAR dose-constraint violations.

Replanning based on MAN, asPREC and asMIM using a patient template resulted in plans satisfying OAR constraints (evaluated using MAN) for 93% (91/98), 64% (63/98) and 56% (55/98) of the fractions. Nonetheless, the V35Gy in unedited AUTO OARs was significantly lower in all organs compared to non-adapted plans for both asPREC and asMIM. Compared to NoAd plans, replanned doses on daily adapted contours (MAN, asPREC or asMIM) improved V35Gy in all OAR for 100% (98/98), 95% (93/98) and 84% (82/98) of the fractions. Using asPREC, the 5 fractions performing worse than NoAd occurred in 4 patients. Similarly, using asMIM, the 16 fractions performing worse than NoAD occurred in 14 patients. Median PTV coverage reduced by 2%, 2.7% and 5.1% compared to NoAD plans after replanning with MAN, asPREC and asMIM, respectively.

**Table 3** summarizes the differences between replanning using MAN vs. replanning using AUTO. V35Gy is significantly higher for the stomach and duodenum in plans based on autocontours compared to those based on MAN contours. This effect does not occur in the case of the bowel. **Table 3** also shows that the PTV coverage decreased when using AUTO. This result was not significant when replanning using asPREC, but was significant when using asMIM.

**Figure 3** shows the dosimetric parameters of adapted plans based on MAN, asPREC or asMIM vs. non-adapted plans. Dots located under the unity line (in diagonal) represent the dose

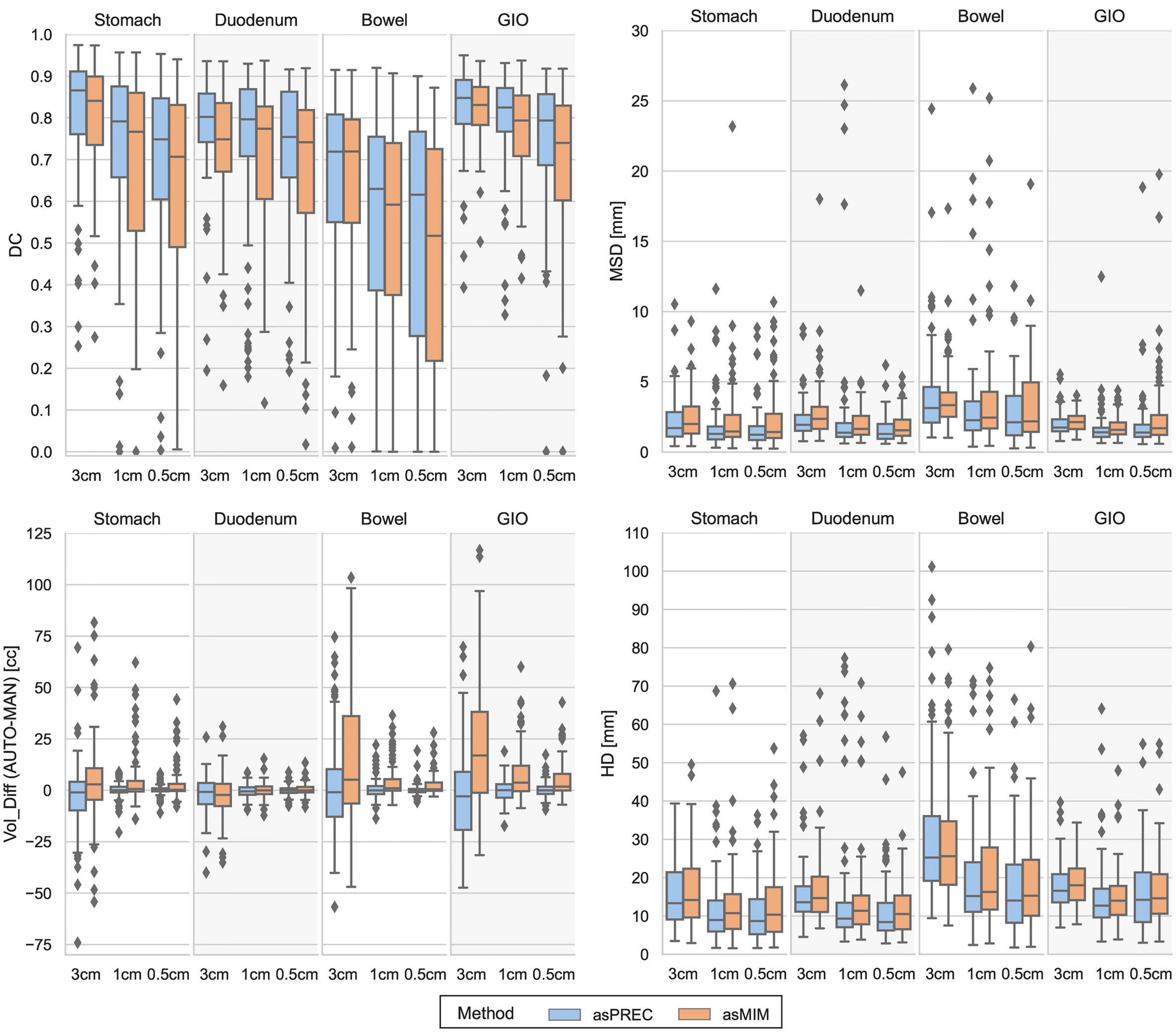


**FIGURE 2** | Boxplots showing the differences between Dice coefficient (DC) [top left], mean surface distance (MSD) [top right], volumetric difference between auto vs. manual contours (VOL_DIFF) [bottom left] and Hausdorff distance (HD) [bottom right] for structures autosegmented with Precision (asPREC in blue) and MIM (asMIM in orange). Each column of each subfigure distinguishes the boxplots on each structure (stomach, duodenum, bowel and GIO) and for each organ, distributions are separated for the clipped structures at 3, 1 and 0.5 cm from the PTV.

distributions that improved compared to non-adapted plans. Similarly, dots located under the horizontal dashed red line at 0.5 cc on the y-axis represent the amount of adapted dose distributions that fulfilled the dose-constraints after adapting the plans using the three different contours sets. **Figure 3** visually presents the results from **Tables 2**, **3**: most plans fulfill the dose-constraints for the three organs after replanning at the cost of PTV coverage.

The correlation between autocontontour geometrical errors (assessed using VOL_DIFF of AUTO vs. MAN contours) and OAR violations (i.e., V35 Gy > 0.5 cc) were reported to be significant on all OAR within the ring of 0 to 1.5 cm from the PTV and not significant within the ring from 1.5 to 3 cm (**Table 4**). Other ring combinations results can be found in **Supl.Mat** (**Table B3**), but reported similar tendencies to **Table 4**. In short, large OAR autosegmentation inaccuracies (i.e., showing negative VOL_DIFF) occurring close to the PTV, appeared to be correlated with OAR violations after replanning. This correlation disappeared for large geometrical differences occurring at larger distances (i.e., within 1.5–3 cm ring from the PTV). **Tables 4** and **B3** suggest that recontouring efforts should primarily be addressed to OAR volumes close to the PTV, as this effort already solves most dose-constraint violations when replanning while minimizing the editing time involved.

## DISCUSSION

Treatments using ART, especially online adaptive replanning, heavily rely on autosegmentation for a speedy and efficient workflow. However, current autosegmentation methods generally lack accuracy in the abdominal region and need to be followed by time and labor-intensive manual contour correction.

**TABLE 1 |** Median and interquartile range (Q1, Q3) of the minimum distance (MIN_DIST) from GTV and PTV to OARs (stomach, duodenum and bowel), and the overlapping volume (OVLP) of the expanded PTV (at 0.5 and 1 cm) and OAR.

| Metric | Method | Stomach | Duodenum | Bowel |
|---|---|---|---|---|
| **MIN_DIST**<br>**GTV – OAR [mm]** | **MAN** | 2.1 (-0.3, 6.9) | -0.3 (-0.8, 4.3) | 9.4 (3.4, 15.0) |
| | **asPREC** | 2.3 (-0.6, 7.1) | 0.0 (-1.7, 5.4) | 9.7 (3.1, 20.8) |
| | **asMIM** | 1.2 (-1.5, 6.4) | -0.2 (-2.2, 4.6) | 8.5 (0.5, 16.9) |
| | **(asPREC – MAN)** | -0.3 (-1.3, 1.3) | -0.5 (-1.4, 1.2) | 0.4 (-1.5, 4.0) |
| | **(asMIM – MAN)** | -0.9 (-2.9, 0.4) | -0.9 (-2.1, 0.4) | -0.8 (-3.2, 2.0) |
| **MIN_DIST**<br>**PTV – OAR [mm]** | **MAN** | -4.2 (-6.5, 0.4) | -6.4 (-7.4, -1.5) | 3.1 (-2.6, 8.8) |
| | **asPREC** | -4.0 (-6.9, 1.1) | -6.0 (-8.1, -1.0) | 3.3 (-2.4, 14.2) |
| | **asMIM** | -5.1 (-7.9, -0.1) | -6.5 (-8.6, -1.8) | 2.3 (-5.5, 10.8) |
| | **(asPREC – MAN)** | -0.1 (-1.2, 1.3) | -0.4 (-1.2, 1.2) | 0.6 (-1.3, 3.6) |
| | **(asMIM – MAN)** | -0.8 (-3.1, 0.6) | -0.8 (-2.2, 0.4) | -0.6 (-3.3, 1.9) |
| **OVLP**<br>**PTV_0.5cm - OAR [cc]** | **MAN** | 3.4 (0.6, 8.2) | 5.8 (1.5, 14.6) | 0.0 (0.0, 1.6) |
| | **asPREC** | 3.0 (0.5, 9.1) | 5.6 (1.2, 16.5) | 0.0 (0.0, 1.9) |
| | **asMIM** | 4.3 (0.8, 12.1) | 6.6 (1.8, 17.3) | 0.3 (0.0, 3.4) |
| | **(asPREC – MAN)** | 0.0 (-0.4, 0.8) | 0.0 (-1.2, 0.9) | 0.0 (-0.4, 0.2) |
| | **(asMIM – MAN)** | 0.2 (-0.3, 2.7) | 0.0 (-0.9, 1.5) | 0.0 (0.0, 1.4) |
| **OVLP**<br>**PTV_1cm - OAR [cc]** | **MAN** | 9.5 (4.2, 18.7) | 13.3 (4.4, 27.7) | 1.7 (0.0, 6.7) |
| | **asPREC** | 9.2 (2.9, 19.2) | 12.2 (4.7, 29.6) | 2.4 (0.0, 8.1) |
| | **asMIM** | 10.1 (4.4, 23.9) | 13.4 (5.4, 29.7) | 2.8 (0.0, 11.9) |
| | **(asPREC – MAN)** | 0.0 (-1.0, 1.5) | -0.4 (-2.1, 1.5) | 0.0 (-1.2, 1.7) |
| | **(asMIM – MAN)** | 0.3 (-0.5, 3.8) | 0.0 (-1.8, 1.9) | 0.2 (-0.1, 4.4) |

*Results are presented for both manual (MAN), and autosegmented contours using Precision (asPREC) and MIM (asMIM), as well as the difference between auto and manual contours.*

**TABLE 2 |** Median and interquartile range (Q1, Q3) plan parameters of the replanned doses based on manual (MAN), and autosegmented contours using precision (asPREC) and MIM (asMIM) vs. non-adapted planned doses (NoAd).

| Structure | Parameters | No adaptation (NoAd) | Replanning | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | MAN_Rp – NoAd | ρ | asPREC_Rp – NoAd | ρ | asMIM_Rp – NoAd | ρ |
| **PTV** | **Coverage (%)** | 83.8 (78.0, 90.7) | -2.0 (-4.6, 0.1) | <.001 | -2.7 (-4.5, -0.6) | <.001 | -5.1 (-8.4, -2.6) | <.001 |
| | **Dmean (Gy)** | 43.1 (42.2, 44.1) | -0.5 (-1.0, 0.0) | <.001 | -0.3 (-0.7, 0.0) | <.001 | -0.7 (-1.2, -0.1) | <.001 |
| | **Dmin (Gy)** | 26.7 (25.5, 28.2) | -0.5 (-1.6, 0.3) | <.001 | -0.7 (-1.4, 0.1) | <.001 | -0.5 (-1.5, 0.2) | <.001 |
| **GTV** | **Coverage (%)** | 95.7 (91.1, 99.0) | -0.1 (-2.1, 0.6) | 0.02 | -0.4 (-1.9, 0.1) | <.001 | -1.6 (-5.2, 0.0) | <.001 |
| | **Dmean (Gy)** | 45.8 (45.0, 46.5) | -0.5 (-1.0, 0.0) | <.001 | -0.1 (-0.7, 0.2) | <.001 | -0.5 (-1.2, 0.1) | <.001 |
| **Stomach** | **V35 Gy (cc)** | 0.2 (0.0, 0.8) | -0.2 (-0.8, 0.0) | <.001 | -0.1 (-0.7, 0.0) | <.001 | -0.1 (-0.4, 0.0) | <.001 |
| | **Dmean (Gy)** | 5.4 (3.3, 7.4) | -0.1 (-0.5, 0.5) | NS | -0.0 (-0.5, 0.5) | NS | -0.1 (-0.5, 0.4) | NS |
| **Duodenum** | **V35 Gy (cc)** | 0.5 (0.1, 1.2) | -0.4 (-1.0, 0.0) | <.001 | -0.2 (-0.7, 0.0) | <.001 | -0.2 (-0.5, 0.0) | <.001 |
| | **Dmean (Gy)** | 9.7 (5.7, 12.7) | -0.3 (-1.1, 0.3) | <.001 | -0.4 (-1.0, -0.1) | <.001 | -0.4 (-0.9, 0.1) | <.001 |
| **Bowel** | **V35 Gy (cc)** | 0.0 (0.0, 0.3) | 0.0 (-0.3, 0.0) | <.001 | 0.0 (-0.1, 0.0) | <.001 | 0.0 (-0.1, 0.0) | <.001 |
| | **Dmean (Gy)** | 1.9 (1.3, 2.6) | -0.1 (-0.3, 0.1) | <.001 | -0.2 (-0.3, 0.0) | <.001 | -0.2 (-0.3, 0.0) | <.001 |

*Statistically not significant (NS) for p > 0.05.*

**TABLE 3 |** Median and interquartile range (Q1, Q3) plan parameters of the replanned doses based on autosegmented contours using precision (asPREC) and MIM (asMIM) vs. replanned doses based on manual contours (MAN).

| Structure | Parameters | Replanning | | | | |
|---|---|---|---|---|---|---|
| | | MAN_Rp | asPREC_Rp – MAN_Rp | ρ | asMIM_Rp – MAN_Rp | ρ |
| **PTV** | **Coverage (%)** | 82.5 (75.1, 88.7) | -0.5 (-3.5, 1.6) | NS | -2.7 (-7.4, 0.2) | <.001 |
| | **Dmean (Gy)** | 42.5 (41.5, 43.7) | 0.0 (-0.3, 0.6) | NS | -0.2 (-1.0, 0.4) | 0.04 |
| | **Dmin (Gy)** | 26.4 (24.6, 28.0) | 0.0 (-0.8, 0.8) | NS | 0.0 (-1.1, 1.0) | NS |
| **GTV** | **Coverage (%)** | 95.6 (90.7, 98.9) | -0.1 (-1.9, 1.1) | NS | -1.6 (-4.0, 0.0) | <.001 |
| | **Dmean (Gy)** | 45.3 (44.2, 46.1) | 0.3 (-0.2, 0.9) | .001 | 0.0 (-0.8, 0.6) | NS |
| **Stomach** | **V35 Gy (cc)** | 0.0 (0.0, 0.0) | 0.0 (0.0, 0.1) | <.001 | 0.0 (0.0, 0.3) | <.001 |
| | **Dmean (Gy)** | 5.2 (3.2, 7.5) | 0.0 (-0.3, 0.4) | NS | 0.0 (-0.4, 0.5) | NS |
| **Duodenum** | **V35 Gy (cc)** | 0.0 (0.0, 0.0) | 0.1 (0.0, 0.4) | <.001 | 0.0 (0.0, 0.4) | <.001 |
| | **Dmean (Gy)** | 9.5 (4.9, 12.1) | -0.2 (-0.6, 0.2) | .02 | -0.1 (-0.4, 0.3) | NS |
| **Bowel** | **V35 Gy (cc)** | 0.0 (0.0, 0.0) | 0.0 (0.0, 0.0) | NS | 0.0 (0.0, 0.0) | NS |
| | **Dmean (Gy)** | 1.8 (1.0, 2.6) | 0.0 (-0.1, 0.1) | NS | 0.0 (-0.2, 0.1) | .01 |

*Statistically not significant (NS) for p > 0.05.*

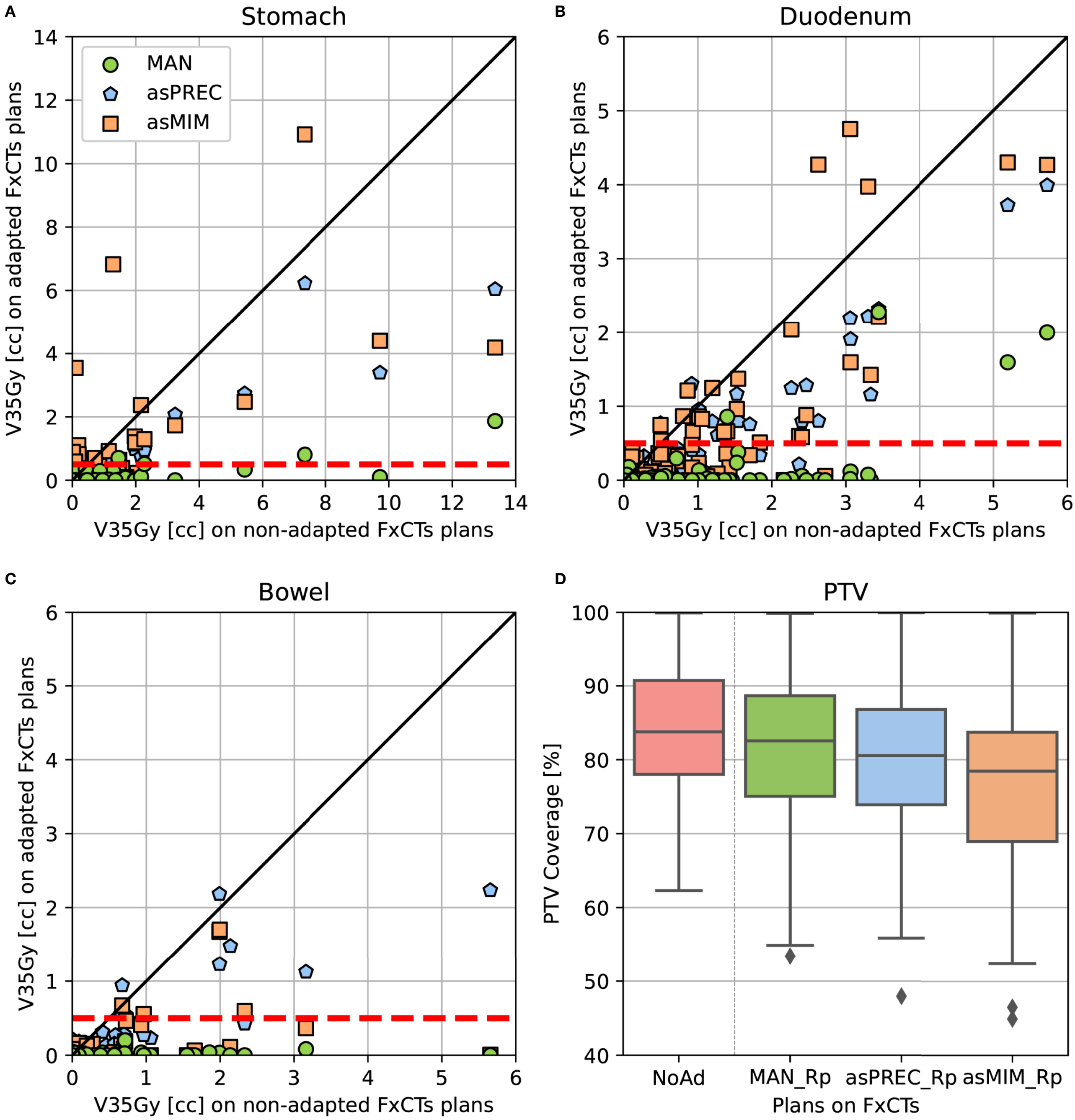


**FIGURE 3** | Pair-point comparison of OAR V35Gy parameter on non-adapted vs. adapted plans using manual and autosegmented contours with Precision (asPREC) and MIM (asMIM) on the stomach **(A)**, duodenum **(B)**, bowel **(C)**. Dashed lines depict OAR dose-constraints (V35Gy < 0.5 cc). In **(D)**, PTV coverage boxplot comparison of non-adapted (NoAd – red) vs. replanned doses: MAN_Rp (green), asPREC_Rp (blue) and asMIM_Rp (orange).

In this study, we have quantified autocontouring quality of two commercially available software tools in the upper abdomen, and assessed the use of the resulting contours without further editing in daily replanning. Replanning with unedited contours resulted in better OAR sparing than non-adapted plans in 95% and 84% of plans optimized using Precision and MIM autocontours, respectively. For a large proportion of these fractions, resulting replanned doses stayed within OAR constraints (64% of plans when using Precision DIR, and 56% when using MIM DIR). Although autosegmentation inaccuracies can be located all over the OARs, the errors located closer to the PTV structure have the largest impact on OAR doses when replanning. These results

**TABLE 4 |** Median and interquartile range (Q1, Q3) of the volumetric difference of auto and manual contours in fractions violating and non-violating dose-constraints (V35Gy > 0.5cc) in the stomach, duodenum and bowel after replanning using precision (asPREC) and MIM (asMIM) autocontours.

| Structure | Method | Distance to PTV | VOL_DIFF (AUTO – MAN) [cc] | | |
|---|---|---|---|---|---|
| | | | Do not violate(V35 < 0.5 cc) | Violate(V35 > 0.5 cc) | *p* |
| **Stomach** | **asPREC** | **Ring 0 – 1.5 cm** | 0.3 (-1.6, 2.0) | -10.9 (-13.6, -3.1) | .002 |
| | | **Ring 1.5 – 3 cm** | -1.9 (-7.2, 2.3) | -17.9 (-25.6, 3.4) | NS |
| | **asMIM** | **Ring 0 – 1.5 cm** | 1.7 (-0.0, 6.2) | -6.2 (-10.2, 1.1) | <.001 |
| | | **Ring 1.5 – 3 cm** | 1.0 (-4.2, 4.6) | -1.0 (-11.7, 8.1) | NS |
| **Duodenum** | **asPREC** | **Ring 0 – 1.5 cm** | 0.2 (-2.3, 2.8) | -2.9 (-6.1, -1.1) | .001 |
| | | **Ring 1.5 – 3 cm** | -0.1 (-5.6, 1.7) | 0.2 (-2.2, 4.8) | NS |
| | **asMIM** | **Ring 0 – 1.5 cm** | 0.5 (-2.2, 2.7) | -3.0 (-7.4, 0.5) | .007 |
| | | **Ring 1.5 – 3 cm** | -0.4 (-7.1, 2.1) | 0.3 (-6.7, 2.3) | NS |
| **Bowel** | **asPREC** | **Ring 0 – 1.5 cm** | 0.5 (-1.5, 6.4) | -7.8 (-11.9, -4.9) | <.001 |
| | | **Ring 1.5 – 3 cm** | 0.5 (-10.7, 8.2) | -6.0 (-10.1, -2.8) | NS |
| | **asMIM** | **Ring 0 – 1.5 cm** | 1.0 (-0.8, 12.0) | -6.2 (-6.6, -3.4) | .017 |
| | | **Ring 1.5 – 3 cm** | 6.6 (-3.7, 27.9) | 2.5 (1.3, 4.5) | NS |

*Results are presented for the contour evaluated in the ring from 0 to 1.5 cm from the PTV vs. the ring from 1.5 to 3 cm from the PTV. Statistically not significant (NS) for p > 0.05.*

suggest that manual editing of autosegmented OAR can be avoided in many fractions, but if applied, it can be limited to the region closest to the PTV to reduce the overall time of the ART workflow when treating patients with LAPC. Our research suggests that a cut-off limit of 1.5 cm could be sufficient, but an exact cut-off point requires further research and will be treatment protocol dependent.

A similar study was recently published using unedited contours for daily online ART in prostate patients using the Ethos system (32). In this study, the authors evaluated the gain of adapted plans with unedited contours vs. non-adapted plans. They report that 96% of their fractions would have required manual editing of the generated contours, but that 100% of the fractions achieved higher CTV coverage based on autocontours than using non-adapted plans. Similar to our work, the authors show that autocontouring methods are still inaccurate and require manual editing, but they also show that replanning on unedited contours is already beneficial compared to treating patient with non-adapted plans.

The added value of our work is that we also evaluated the dosimetric differences between adapted plans using manually corrected contours vs. using autocontours, hence, we also measured the potential gain in plan quality if autocontours are edited before replanning.

Regarding the geometrical analysis performed in our data, as expected, there were differences between manual and autocontours in the low and high dose region (within 3 cm from the PTV). Dice coefficient degraded when getting closer to the PTV. This is in part a natural expectation from this metric, as reports the overlapping ratio between 2 structures. The smaller the evaluated volumes, the more impact segmentation inaccuracies have. The Hausdorff distance measurement, reporting the maximum distance between 2 volumes, remains constant at different distances from the PTV, what reassures that there are relevant inaccuracies occurring close to the tumor.

Generally, contours propagated by Precision DIR showed a slightly higher agreement with manual contours than with MIM DIR, which tended to overestimate OARs (**Figure 2**, **Table B2**), and get closer to the tumor (**Table 1**). Consequently, asMIM_Rp dose distributions more often exceeded dose-constraints and lost more PTV coverage than asPREC_Rp. This difference between autocontour quality might be because Precision DIR optimizes the deformation vector field using localized patches within the image instead of the global image as done by MIM DIR (33–35) (see **Supl.Mat-A**).

Daily recontouring has traditionally relied on intra-patient contour propagation (as in this study) or atlas-based methods also using DIR (2, 3). Alternative autosegmentation methods are described in the literature, including artificial intelligence (AI). AI-based methods have shown improved accuracy and efficiency compared to traditional methods while being computationally very fast (36, 37). Several studies have shown improvements in different treatment sites (e.g. head-and-neck (38–40), prostate (39, 41), rectum (42), whole body (43)). However, abdominal organs present additional challenges including strong interpatient variability, bowel loop displacements and hollow organs, which causes AI studies still report similar results to those achieved in our current study (10, 44–46). Additionally, all studies focus on reporting autosegmentation accuracy on whole organ structures, whereas our results suggest mainly the accuracy close to the target influences plan quality.

Regarding replanning, manually corrected contours achieved the best results in OAR sparing compared to non-adapted plans (100% FxCT). However, replanning directly on unedited structures also improved OAR sparing for the large majority of fractions: 95% (93/98 FxCT) for Precision, and 84% (82/98 FxCT) for MIM. The corresponding 5 and 16 fractions in which plans based on autocontours increased OAR dose compared to non-adapted plans belonged to 4 and 14 patients, respectively. When looking further into the cases in which this phenomenon occurred (see two example cases in **Figure C1 in Supl. Mat.**), we noticed that manual contours were closer to the PTV than autosegmented contours, resulting in large inaccuracies close to the PTV for AUTO. Replanning on the autosegmented contours results in large dose violations, as the manual contours lie in the high dose area. Nonetheless, this poses a relatively small dosimetric risk for the patient especially taking into account that we analyzed single fractions rather than the total treatment dose, in which the effect of dose violations occurring in one single fraction, as in the case for the majority of our reported violations, is likely to be reduced.

Although OAR dose decreased when using unedited contours, the number of fractions obeying OAR constraints reduced compared to plans based on corrected contours. Also, PTV coverage generally decreased in fractions needing replanning. This similarly occurred when using MAN or asPREC, and slightly more often when using asMIM. Mostly, this was explained due to daily OAR moving closer to the high dose region or an increased OAR overlap with the PTV.

Our proposed implementation of ART is based on CT images and uses commercially available software. Although we are still in process of clinically implementing online adaptive replanning, we have performed end-to-end tests to mimic a clinical workflow. A complete adaptive procedure can be completed within 45 min, excluding treatment delivery, with room for improvement in delineation time. Similar to other publications, depending on the treatment site, editing of the contours on the FxCT – even when limited to a distance of 3 cm from the PTV - can take up a considerable amount of time in the entire procedure (around 10 min (4, 22, 27, 47)). The time of our total procedure is however in line with procedures performed on the MR-Linac (9, 12, 27), but is considerably longer than an online workflow on the Ethos system (22, 23). An inherent advantage of CyberKnife treatments is the excellent intra-fraction, both respiratory and non-respiratory, motion tracking. Currently this is lacking in the MR-Unity and Ethos systems leading to a possible increase in target size. The MRIdian is compensating for intra-fraction respiratory motion by means of gating.

Another limitation of our work is that we have a relatively small cohort group for this study. A validation involving an independent dataset potentially from other institutes should be performed to verify the relevance of our findings in pancreatic cancer. Although MR-Linacs and the Ethos systems rely on different imaging modalities, we believe our results could be transferred to other systems. For instance, similar trends were already observed in the work of Moazzezi et al. about online ART using unedited contours in prostate patients using the Ethos system (32). However, the complexity of the procedure might increase as the amount of elements involved also increases, e.g. generating correct Hounsfield Units.

Finally, intrafraction OAR motion has not been accounted for in this study. In our clinic, we use Synchrony respiratory motion tracking to mitigate the effect of intrafraction motion of the target, of which the accuracy has been reported elsewhere (25). Generally, intrafraction OAR variations while tracking are expected to be smaller than interfraction variations. Replans based on unedited contours already correct for interfraction OAR variations and generally outperform non-adapted plans in this study. We believe intrafraction OAR variations will have a smaller impact on the replans.

In conclusion, autosegmentation methods applying contour propagation after DIR in the abdominal region result in contours requiring manual correction. However, replanning on the unedited daily contours generally resulted in higher organ sparing than treating with a conventional SBRT scheme. In the majority of fractions, it even resulted in plans obeying the tight OAR dose constraints of our clinical protocol. In a large number of fractions, manual editing of automatic contours could, therefore, be avoided or at least restricted to contour sections in close proximity to the PTV, reducing the time required for online adaptive treatments for pancreatic cancer patients.

## DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**. Further inquiries can be directed to the corresponding author.

## ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Institutional Review Board (ID: NL49643.078.14). The patients/participants provided their written informed consent to participate in this study.

## AUTHOR CONTRIBUTIONS

AM-B: conducting main research, data interpretation and analysis, and writing manuscript. MM, PG, and MH: supervision of research, contributed to the study design and writing manuscript. WT: contributed to data collection. JN: PI of data collection. All authors contributed to the article and approved the submitted version.

## FUNDING

This work was in part funded by a research grant of Accuray Inc., Sunnyvale, USA. Erasmus MC Cancer Institute also has a research collaboration with Elekta AB, Stockholm, Sweden and Varian, Paolo Alto, USA. The funder was not involved in the study design, collection, analysis, interpretation of data, the writing of this article or the decision to submit it for publication.

## ACKNOWLEDGMENTS

The authors wish to thank Warren Kilby and Mark Foskey from Accuray research and development team, as well as Michael Duchateau from MIM Software for providing valuable insight and carefully reviewing this manuscript.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fonc.2022.910792/full#supplementary-material

## REFERENCES

1. Lim-Reinders S, Keller BM, Al-Ward S, Sahgal A, Kim A. Online Adaptive Radiation Therapy. *Int J Radiat Oncol Biol Phys* (2017) 99(4):994–1003. doi: 10.1016/j.ijrobp.2017.04.023
2. Sharp G, Fritscher KD, Pekar V, Peroni M, Shusharina N, Veeraraghavan H, et al. Vision 20/20: Perspectives on Automated Image Segmentation for Radiotherapy. *Med Phys* (2014) 41(5):050902. doi: 10.1118/1.4871620
3. Glide-Hurst CK, Lee P, Yock AD, Olsen JR, Cao M, Siddiqui F, 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* (2021) 109(4):1054–75. doi: 10.1016/j.ijrobp.2020.10.021
4. Bohoudi O, Bruynzeel AME, Senan S, Cuijpers JP, Slotman BJ, Lagerwaard FJ, et al. Fast and Robust Online Adaptive Planning in Stereotactic MR-Guided Adaptive Radiation Therapy (SMART) for Pancreatic Cancer. *Radiother Oncol* (2017) 125(3):439–44. doi: 10.1016/j.radonc.2017.07.028
5. Olberg S, Green O, Cai B, Yang D, Rodriguez V, Zhang H, et al. Optimization of Treatment Planning Workflow and Tumor Coverage During Daily Adaptive Magnetic Resonance Image Guided Radiation Therapy (MR-IGRT) of Pancreatic Cancer. *Radiat Oncol* (2018) 13(1):1–8. doi: 10.1186/s13014-018-1000-7
6. Magallon-Baro A, Milder MTW, Granton PV, Nuyttens JJ, Hoogeman MS. Comparison of Daily Online Plan Adaptation Strategies for a Cohort of Pancreatic Cancer Patients Treated With SBRT. *Int J Radiat Oncol Biol Phys* (2021) 134:127–134. doi: 10.1016/j.ijrobp.2021.03.050
7. Winkel D, Bol GH, Kroon PS, van Asselen B, Hackett SS, Werensteijn-Honingh AM, et al. Adaptive Radiotherapy: The Elekta Unity MR-Linac Concept. *Clin Transl Radiat Oncol* (2019) 18:54–9. doi: 10.1016/j.ctro.2019.04.001
8. Corradini S, Alongi F, Andratschke N, Belka C, Boldrini L, Cellini F, et al. MR-Guidance in Clinical Reality: Current Treatment Challenges and Future Perspectives. *Radiat Oncol* (2019) 14(1):1–12. doi: 10.1186/s13014-019-1308-y
9. Boldrini L, Cusumano D, Cellini F, Azario L, Mattiucci GC, Valentini V. Online Adaptive Magnetic Resonance Guided Radiotherapy for Pancreatic Cancer: State of the Art, Pearls and Pitfalls. *Radiat Oncol* (2019) 14(1):1–6. doi: 10.1186/s13014-019-1275-3
10. Boldrini L, Bibault JE, Masciocchi C, Shen Y, Bittner MI. Deep Learning: A Review for the Radiation Oncologist. *Front Oncol* (2019) 9. doi: 10.3389/fonc.2019.00977
11. Liang F, Qian P, Su KH, Baydoun A, Leisser A, Van Hedent S, et al. Abdominal, Multi-Organ, Auto-Contouring Method for Online Adaptive Magnetic Resonance Guided Radiotherapy: An Intelligent, Multi-Level Fusion Approach. *Artif Intell Med* (2018) 90:34–41. doi: 10.1016/j.artmed.2018.07.001
12. Henke L, Kashani R, Robinson C, Curcuru A, DeWees T, Bradley J, et al. Phase I Trial of Stereotactic MR-Guided Online Adaptive Radiation Therapy (SMART) for the Treatment of Oligometastatic or Unresectable Primary Malignancies of the Abdomen. *Radiother Oncol* (2018) 126(3):519–26. doi: 10.1016/j.radonc.2017.11.032
13. Petrelli F, Comito T, Ghidini A, Torri V, Scorsetti M, Barni S. Stereotactic Body Radiation Therapy for Locally Advanced Pancreatic Cancer: A Systematic Review and Pooled Analysis of 19 Trials. *Int J Radiat Oncol Biol Phys* (2017) 97(2):313–22. doi: 10.1016/j.ijrobp.2016.10.030
14. Goyal K, Einstein D, Ibarra RA, Yao M, Kunos C, Ellis R, et al. Stereotactic Body Radiation Therapy for Nonresectable Tumors of the Pancreas. *J Surg Res* (2012) 174(2):319–25. doi: 10.1016/j.jss.2011.07.044
15. Chuong MD, Springett GM, Freilich JM, Park CK, Weber JM, Mellon EA, et al. Stereotactic Body Radiation Therapy for Locally Advanced and Borderline Resectable Pancreatic Cancer is Effective and Well Tolerated. *Int J Radiat Oncol Biol Phys* (2013) 86(3):516–22. doi: 10.1016/j.ijrobp.2013.02.022
16. Buwenge M, Cellini F, Silvestris N, Cilla S, Deodato F, Macchia G, et al. Robotic Radiosurgery in Pancreatic Cancer: A Systematic Review. *World J Gastroenterol* (2015) 21(31):9420–9. doi: 10.3748/wjg.v21.i31.9420
17. Niedzielski JS, Liu Y, Ng SSW, Martin RM, Perles LA, Beddar S, et al. Dosimetric Uncertainties Resulting From Interfractional Anatomic Variations for Patients Receiving Pancreas Stereotactic Body Radiation Therapy and Cone Beam Computed Tomography Image Guidance. *Int J Radiat Oncol Biol Phys* (2021) 111(5):1298–309. doi: 10.1016/j.ijrobp.2021.08.002
18. Loi M, Magallon-Baro A, Suker M, van Eijck C, Sharma A, Hoogeman M, et al. Pancreatic Cancer Treated With SBRT: Effect of Anatomical Interfraction Variations on Dose to Organs at Risk. *Radiother Oncol* (2019) 134:67–73. doi: 10.1016/j.radonc.2019.01.020
19. Acharya S, Fischer-Valuck BW, Kashani R, Parikh P, Yang D, Zhao T, et al. Online Magnetic Resonance Image Guided Adaptive Radiation Therapy: First Clinical Applications. *Int J Radiat Oncol Biol Phys* (2016) 94(2):394–403. doi: 10.1016/j.ijrobp.2015.10.015
20. Henke LE, Contreras JA, Green OL, Cai B, Kim H, Roach MC, et al. Magnetic Resonance Image-Guided Radiotherapy (MRIgRT): A 4.5-Year Clinical Experience. *Clin Oncol (R Coll Radiol)* (2019) 30(11):720–7. doi: 10.1016/j.clon.2018.08.010
21. Winkel D, Bol GH, Werensteijn-Honingh AM, Kiekebosch IH, van Asselen B, Intven MPW, et al. Evaluation of Plan Adaptation Strategies for Stereotactic Radiotherapy of Lymph Node Oligometastases Using Online Magnetic Resonance Image Guidance. *Phys Imaging Radiat Oncol* (2019) 9:58–64. doi: 10.1016/j.phro.2019.02.003
22. Sibolt P, Andersson LM, Calmels L, Sjöström D, Bjelkengren U, Geertsen P, et al. Clinical Implementation of Artificial Intelligence-Driven Cone-Beam Computed Tomography-Guided Online Adaptive Radiotherapy in the Pelvic Region. *Phys Imaging Radiat Oncol* (2021) 17:1–7. doi: 10.1016/j.phro.2020.12.004
23. Archambault Y, Boylan C, Bullock D, Morgas T, Peltola J, Ruokokoski E, et al. Making on-Line Adaptive Radiotherapy Possible Using Artificial Intelligence and Machine Learning for Efficient Daily Re-Planning. *Med Phys Int J* (2020) 8(2):77–86.
24. Kilby W, Dooley JR, Kuduvalli G, Sayeh S, Maurer CRJr. The CyberKnife ® Robotic Radiosurgery System in 2010. *Technol Cancer Res Treat* (2010) 9 (5):433–52. doi: 10.1177/153303461000900502
25. Hoogeman M, Prévost J-B, Nuyttens J, Pöll J, Levendag P, Heijmen B. Clinical Accuracy of the Respiratory Tumor Tracking System of the CyberKnife: Assessment by Analysis of Log Files. *Int J Radiat Oncol* (2009) 74(1):297–303. doi: 10.1016/j.ijrobp.2008.12.041
26. Papalazarou C, Klop GJ, Milder MTW, Marijnissen JPA, Gupta V, Heijmen BJM, et al. CyberKnife With Integrated CT-On-Rails: System Description and First Clinical Application for Pancreas SBRT. *Med Phys* (2017) 44(9):4816–27. doi: 10.1002/mp.12432
27. Lamb J, Cao M, Kishan A, Agazaryan N, Thomas DH, Shaverdian N, et al. Online Adaptive Radiation Therapy: Implementation of a New Process of Care. *Cureus* (2017) 9(8):e16118. doi: 10.7759/cureus.1618
28. Jabbour SK, Hashem SA, Bosch W, Kim TK, Finkelstein SE, Anderson BM, et al. Upper Abdominal Normal Organ Contouring Guidelines and Atlas: A Radiation Therapy Oncology Group Consensus. *Pract Radiat Oncol* (2014) 4 (2):82–9. doi: 10.1016/j.prro.2013.06.004
29. Baker S, Sharma A, Antonisse I, Cornelissen R, Moelker A, Nuyttens JJ. Endovascular Coils as Lung Tumor Fiducial Markers for Real-Time Tumor Tracking in Stereotactic Body Radiotherapy: Comparison of Complication Rates With Transthoracic Fiducial Marker Placement. *J Vasc Interv Radiol* (2019) 30(12):1901–7. doi: 10.1016/j.jvir.2019.04.025
30. Magallon-Baro A, Loi M, Milder MTW, Granton PV, Zolnay AG, Nuyttens JJ, et al. Modeling Daily Changes in Organ-at-Risk Anatomy in a Cohort of Pancreatic Cancer Patients. *Radiother Oncol* (2019) 134:127–34. doi: 10.1016/j.radonc.2019.01.030
31. 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(7): e43–76. doi: 10.1002/mp.12256
32. Moazzezi M, Rose B, Kisling K, Moore KL, Ray X. Prospects for Daily Online Adaptive Radiotherapy *via* Ethos for Prostate Cancer Patients Without Nodal Involvement Using Unedited CBCT Auto-Segmentation. *J Appl Clin Med Phys* (2021) 22(10):82–93. doi: 10.1002/acm2.13399
33. Gupta V, Wang Y, Méndez Romero A, Myronenko A, Jordan P, Maurer C, et al. Fast and Robust Adaptation of Organs-at-Risk Delineations From Planning Scans to Match Daily Anatomy in Pre-Treatment Scans for Online-Adaptive Radiotherapy of Abdominal Tumors. *Radiother Oncol* (2018) 127(2):332–8. doi: 10.1016/j.radonc.2018.02.014

34. Piper JW, Richmond JH, Nelson AS. *VoxAlign Deformation Engine ® Deformable Algorithms* (2018). Available at: www.mimsoftware.com.
35. Jordan P, Myronenko A, Gorczowski K, Foskey M, Holloway R, Maurer CR. Accuray Deformable Image Registration: Description and Evaluation. In: *White Pap Accuray Software, Accuray Precis* (2017). p. 1–8.
36. Brock KK. Adaptive Radiotherapy: Moving Into the Future. *Semin Radiat Oncol* (2019) 29(3):181–4. doi: 10.1016/j.semradonc.2019.02.011
37. Cardenas CE, Yang J, Anderson BM, Court LE, Brock KB. Advances in Auto-Segmentation. *Semin Radiat Oncol* (2019) 29(3):185–97. doi: 10.1016/j.semradonc.2019.02.001
38. van Dijk LV, Van den Bosch L, Aljabar P, Peressutti D, Both S, Steenbakkers Roel JHM, et al. Improving Automatic Delineation for Head and Neck Organs at Risk by Deep Learning Contouring. *Radiother Oncol* (2020) 142:115–23. doi: 10.1016/j.radonc.2019.09.022
39. Wong J, Huang V, Wells D, Giambattista J, Giambattista J, Kolbeck C, et al. Implementation of Deep Learning-Based Auto-Segmentation for Radiotherapy Planning Structures: A Workflow Study at Two Cancer Centers. *Radiat Oncol* (2021) 16(1):1–10. doi: 10.1186/s13014-021-01831-4
40. Chen W, Li Y, Dyer BA, Feng X, Rao S, Benedict SH, et al. Deep Learning vs. Atlas-Based Models for Fast Auto-Segmentation of the Masticatory Muscles on Head and Neck CT Images. *Radiat Oncol* (2020) 15(1):1–10. doi: 10.1186/s13014-020-01617-0
41. Elguindi S, Zelefsky MJ, Jiang J, Veeraraghavan H, Deasy JO, Hunt MA, et al. Deep Learning-Based Auto-Segmentation of Targets and Organs-at-Risk for Magnetic Resonance Imaging Only Planning of Prostate Radiotherapy. *Phys Imaging Radiat Oncol* (2019) 12:80–6. doi: 10.1016/j.phro.2019.11.006
42. Song Y, Hu J, Wu Q, Xu F, Nie S, Zhao Y, et al. Automatic Delineation of the Clinical Target Volume and Organs at Risk by Deep Learning for Rectal Cancer Postoperative Radiotherapy. *Radiother Oncol* (2020) 145:186–92. doi: 10.1016/j.radonc.2020.01.020
43. Chen X, Sun S, Bai N, Han K, Liu Q, Yao S, et al. A Deep Learning-Based Auto-Segmentation System for Organs-at-Risk on Whole-Body Computed Tomography Images for Radiation Therapy. *Radiother Oncol* (2021) 160:175–84. doi: 10.1016/j.radonc.2021.04.019
44. Ahn SH, Yeo AU, Kim KH, Kim C, Goh Y, Cho S, et al. Comparative Clinical Evaluation of Atlas and Deep-Learning-Based Auto-Segmentation of Organ Structures in Liver Cancer. *Radiat Oncol* (2019) 14(1):1–13. doi: 10.1186/s13014-019-1392-z
45. Kim H, Jung J, Kim J, Cho B, Kwak J, Jang JY, et al. Abdominal Multi-Organ Auto-Segmentation Using 3D-Patch-Based Deep Convolutional Neural Network. *Sci Rep* (2020) 10(1):1–9. doi: 10.1038/s41598-020-63285-0
46. Fu Y, Mazur TR, Wu X, Liu S, Chang X, Lu Y, et al. A Novel MRI Segmentation Method Using CNN-Based Correction Network for MRI-Guided Adaptive Radiotherapy. *Med Phys* (2018) 45(11):5129–37. doi: 10.1002/mp.13221
47. Bertelsen AS, Schytte T, Møller PK, Mahmood F, Riis HL, Gottlieb KL, et al. First Clinical Experiences With a High Field 1.5 T MR Linac. *Acta Oncol (Madr)* (2019) 58(10):1352–7. doi: 10.1080/0284186X.2019.1627417

**Conflict of Interest:** All authors are employed by the Erasmus MC. MM and MH report serving as an advisory board member for Accuray during the conduct of the study.

**Publisher's Note:** All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Oncology

TYPE Original Research
PUBLISHED 28 July 2022
DOI 10.3389/fonc.2022.930432

OPEN ACCESS

EDITED BY
Puneeth Iyengar,
University of Texas Southwestern Medical Center, United States

REVIEWED BY
Remco De Bree,
University Medical Center Utrecht, Netherlands
Bianca Muresan,
Hospital General Universitario De Valencia, Spain

*CORRESPONDENCE
Clifton D. Fuller
cdfuller@mdanderson.org

SPECIALTY SECTION
This article was submitted to Radiation Oncology, a section of the journal Frontiers in Oncology

RECEIVED 27 April 2022
ACCEPTED 29 June 2022
PUBLISHED 28 July 2022

CITATION
Naser MA, Wahid KA, Grossberg AJ, Olson B, Jain R, El-Habashy D, Dede C, Salama V, Abobakr M, Mohamed ASR, He R, Jaskari J, Sahlsten J, Kaski K and Fuller CD (2022) Deep learning auto-segmentation of cervical skeletal muscle for sarcopenia analysis in patients with head and neck cancer. *Front. Oncol.* 12:930432. doi: 10.3389/fonc.2022.930432

# Deep learning auto-segmentation of cervical skeletal muscle for sarcopenia analysis in patients with head and neck cancer

Mohamed A. Naser[1], Kareem A. Wahid[1], Aaron J. Grossberg[2], Brennan Olson[3], Rishab Jain[2], Dina El-Habashy[1,4], Cem Dede[1], Vivian Salama[1], Moamen Abobakr[1], Abdallah S. R. Mohamed[1], Renjie He[1], Joel Jaskari[5], Jaakko Sahlsten[5], Kimmo Kaski[5] and Clifton D. Fuller[1]*

[1]Department of Radiation Oncology, The University of Texas MD Anderson Cancer Center, Houston, TX, United States, [2]Department of Radiation Medicine, Oregon Health & Science University, Portland, OR, United States, [3]Medical Scientist Training Program, Oregon Health & Science University, Portland, OR, United States, [4]Department of Clinical Oncology, Menoufia University Shibin El Kom, Shibin El Kom, Egypt, [5]Department of Computer Science, Aalto University School of Science, Espoo, Finland

**Background/Purpose:** Sarcopenia is a prognostic factor in patients with head and neck cancer (HNC). Sarcopenia can be determined using the skeletal muscle index (SMI) calculated from cervical neck skeletal muscle (SM) segmentations. However, SM segmentation requires manual input, which is time-consuming and variable. Therefore, we developed a fully-automated approach to segment cervical vertebra SM.

**Materials/Methods:** 390 HNC patients with contrast-enhanced CT scans were utilized (300-training, 90-testing). Ground-truth single-slice SM segmentations at the C3 vertebra were manually generated. A multi-stage deep learning pipeline was developed, where a 3D ResUNet auto-segmented the C3 section (33 mm window), the middle slice of the section was auto-selected, and a 2D ResUNet auto-segmented the auto-selected slice. Both the 3D and 2D approaches trained five sub-models (5-fold cross-validation) and combined sub-model predictions on the test set using majority vote ensembling. Model performance was primarily determined using the Dice similarity coefficient (DSC). Predicted SMI was calculated using the auto-segmented SM cross-sectional area. Finally, using established SMI cutoffs, we performed a Kaplan-Meier analysis to determine associations with overall survival.

**Results:** Mean test set DSC of the 3D and 2D models were 0.96 and 0.95, respectively. Predicted SMI had high correlation to the ground-truth SMI in males and females (r>0.96). Predicted SMI stratified patients for overall survival in males (log-rank p = 0.01) but not females (log-rank p = 0.07), consistent with ground-truth SMI.

**Conclusion:** We developed a high-performance, multi-stage, fully-automated approach to segment cervical vertebra SM. Our study is an essential step towards fully-automated sarcopenia-related decision-making in patients with HNC.

KEYWORDS

auto-segmentation, deep learning, skeletal muscle index, head and neck cancer, sarcopenia

## Introduction

Sarcopenia – the excessive loss of skeletal muscle (SM) mass and function – is a common and debilitating phenomenon in head and neck cancer (HNC) patients (1). Weight loss is frequent in HNC due to nutritional deficiencies induced by tumor geometry affecting normal tissues (2) and/or side effects caused by therapeutic interventions (3). Although the link between treatment-associated weight loss and survival in HNC is unclear (4), sarcopenia has been strongly associated with oncologic outcomes and late radiation-induced toxicities (5–7). Notably, in a recent meta-analysis of HNC patients by Surov et al. (5), sarcopenia was significantly associated with lower overall survival (hazard ratio = 1.64, $p < 0.00001$) and disease-free survival (hazard ratio = 2.00, $p < 0.00001$). Therefore, sarcopenia prediction is of paramount importance in patients with HNC.

Sarcopenia can be identified using different diagnostic criteria (8). One quantitative method investigated in various studies is using a threshold based on the skeletal muscle index (SMI), the cross-sectional area of skeletal muscle measured on axial imaging normalized to the square of the patient's height (9). The SMI is most commonly calculated and referenced using CT imaging of abdominal musculature (10–14). However, abdominal imaging is not available for all HNC patients. Importantly, Swartz et al. (15), van Rijn-Dekker et al. (6), and Olson et al. (16) have recently suggested the C3 cervical vertebra musculature cross-sectional area may also be used to quantify sarcopenia accurately.

Current approaches to generate C3 musculature segmentations needed for SMI calculation rely on either semi-automated or completely manual segmentation (6), which can be time-consuming, introduce unnecessary errors, and suffer from interobserver variability. A fully-automated approach would be an attractive alternative to the current manual/semi-automated standard. Deep learning, which has found success in medical image segmentation (17–20), may be an ideal choice for fully-automated segmentation of SM. Several recent studies have utilized deep learning methods for automated SM measurement based on abdominal CT scans with reasonable performance (21–26). However, to date, no studies have attempted to automate the SMI calculation workflow based on head and neck imaging.

The primary objective of this study was to develop a fully-automated approach to segment skeletal muscle at the C3 vertebral level for use in SMI calculations. These calculations could be directly used to determine sarcopenia status for predicting prognostic outcomes. To achieve this goal, we developed and implemented a two-stage deep learning system that utilizes 3D and 2D ResUNets to detect the C3 vertebra and segment the corresponding C3 musculature, respectively. We show that our approach can faithfully generate segmentations comparable to ground-truth human-generated segmentations. By fully automating the sarcopenia determination workflow, we can ensure rapid, reproducible, and accurate measurements for use in clinical decision-making.

## Materials and methods

### Patient and imaging data

495 patients from the head and neck squamous cell carcinoma (HNSCC) publicly available dataset collection on The Cancer Imaging Archive (TCIA) (27–29) were retrospectively collected in 2021. All patients had a histopathologically-proven diagnosis of squamous cell carcinoma of the oropharynx and were treated with curative-intent intensity-modulated radiotherapy. DICOM-formatted contrast-enhanced CT scans were acquired from the TCIA databases (27–29). Of the 495 patients available in the HNSCC collection, 396 were selected due to their inclusion of the C3 vertebrae on imaging. Subsequently, 6 patients were removed due to image reconstruction errors (n=1), image processing errors (n=1), or oblique image orientations (n=4), leading to a final set of 390 patients used in this analysis. The clinical and demographic characteristics of these patients are shown in **Table 1**. The majority of patients were male (86.6%) with base of tongue tumors (51.6%). SM (paraspinal and sternocleidomastoid muscles) was manually segmented for each CT image in one slice (2D image) at the level of the C3 vertebra. The segmentations were performed using sliceOmatic, version 5.0 (Tomovision) using previously published Hounsfield unit thresholds to define muscle and fat (12, 30);

**Abbreviations:** DSC, Dice similarity coefficient; HNC, head and neck cancer; ROI, region of interest; SM, skeletal muscle; SMI, skeletal muscle index.

TABLE 1 Clinical demographics of patients whose data were used in this study.

| Characteristic | Count |
|---|---|
| Age (median, range) | 57 (28–87) |
| Sex | |
| Male | 337 |
| Female | 52 |
| Tumor subsite | |
| Base of tongue | 201 |
| Glossopharyngeal sulcus | 9 |
| Soft palate | 6 |
| Tonsil | 157 |
| Not otherwise specified | 16 |
| HPV status | |
| Negative | 36 |
| Positive | 215 |
| Unknown | 138 |
| T-category | |
| T1 | 77 |
| T2 | 166 |
| T3 | 91 |
| T4 | 55 |
| N-category | |
| N0 | 36 |
| N1 | 44 |
| N2 | 301 |
| N3 | 8 |
| AJCC stage (7th ed) | |
| I | 3 |
| II | 12 |
| III | 57 |
| IV | 317 |

AJCC, American Joint Committee on Cancer. One patient did not have clinical information from The Cancer Imaging Archive so was not included in this table.

specifically, a range of -29 to +150 Hounsfield units was used to initially define SM followed by manual corrections. No pathological tissue was located in the segmented SM. The single-slice 2D CT images selected for segmentation and the corresponding SM segmentation masks were exported as DICOM files and tag files, respectively. Segmentations are made publicly available on Figshare (doi: 10.6084/m9.figshare.18480917); additional information on the dataset used in this analysis can be found in the corresponding data descriptor (31).

## Image processing

The DICOM 3D volumetric and single-slice 2D CT images were converted to Neuroimaging Informatics Technology Initiative (NIfTI) format using the DICOM processing toolkit DICOMRTTool v. 0.3.21 (32). The SM segmentation. tag files were converted to NIfTI format using an in-house Python script. The NIfTI files for the single-slice 2D CT images and SM segmentation were used to train the 2D segmentation model (described below). The 2D CT slice location in the C3 vertebra was extracted from the DICOM file, which was then used to generate the ground-truth segmentation mask for the C3 section, defined as a volume 33 mm in thickness centered at the location of the 2D CT slice. The tissue regions in the 3D CT images were distinguished from the background by thresholding the images using a value of greater than -500 Hounsfield units with any air gaps within the tissue region filled to generate a binary mask for the external boundaries. The generated external boundary masks and the locations of the 2D CT slices were used to create the ground-truth C3 section segmentations to train the 3D model (described below). As we have described elsewhere (33), all the images and masks were resampled to a fixed image resolution of 1 mm across all dimensions. The CT intensities were truncated in the range of [−250, 250] Hounsfield units to increase soft tissue contrast and then normalized to the range of [-1, 1] scale (**Figures 1A**, **B**). We used the Medical Open Network for AI (MONAI) (34) software transformation functions to rescale and normalize images.

## Segmentation model

We used a multi-stage deep learning convolutional neural network approach for SM segmentation. Our approach was based on the UNet architecture with residual connections (ResUNet) included in the MONAI software package, as we have described in previous publications (33, 35). In the first stage of our approach (**Figure 1C**), a 3D ResUNet model auto-segmented the C3 vertebra section (33 mm), which was then followed by auto-selection of the middle slice of the section. In the second stage of our approach (**Figure 1D**), a 2D ResUNet model auto-segmented the SM on the auto-selected slice of the C3 section. Additional details of our architecture are described in **Appendix A**.

## Model implementation

We randomly split the data into 300 patients for training and 90 patients for testing. For training, we used a 5-fold cross-validation approach where the 300 patients from the training data were divided into five non-overlapping sets. Each set (60 patients) was used for model validation while the 240 patients in the remaining sets were used for training, i.e., each set was used once for testing and four times for training, leading to five sub-models. The processed CT and corresponding masks for 3D ResUNet model and 2D ResUNet models (C3 section and SM, respectively) were randomly cropped to four random fixed-sized regions (patches) of size (96, 96, 96) and (96, 96) per patch per patient, respectively. Additional details on the model

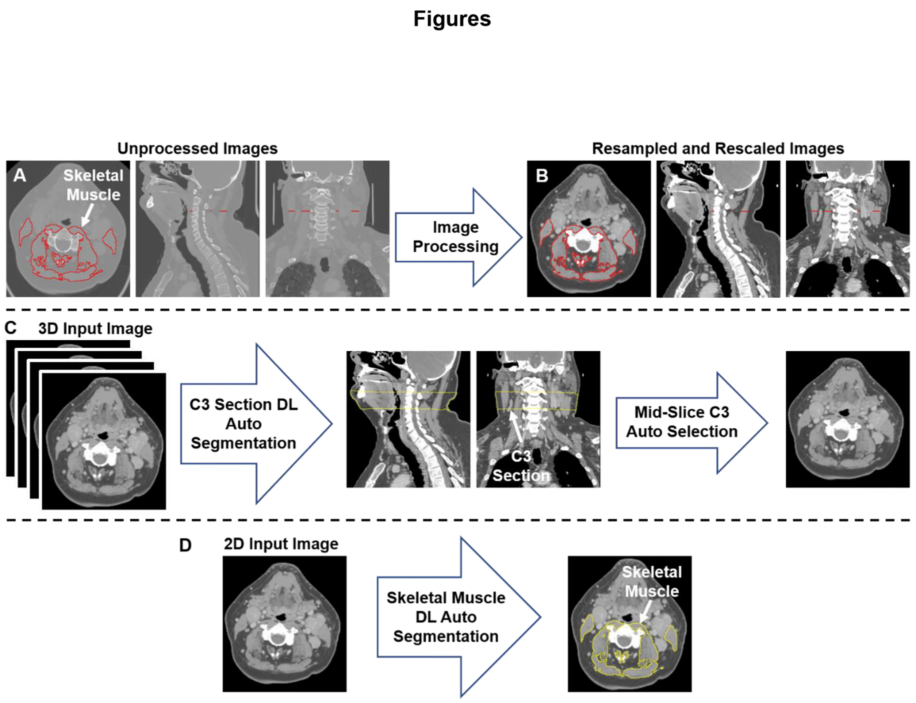

Figures
Unprocessed Images
Resampled and Rescaled Images
A
Skeletal Muscle
Image Processing
B
C
3D Input Image
C3 Section DL Auto Segmentation
C3 Section
Mid-Slice C3 Auto Selection
D
2D Input Image
Skeletal Muscle DL Auto Segmentation
Skeletal Muscle

FIGURE 1
An illustration of the workflow used for skeletal muscle (SM) auto-segmentation at the C3 vertebra. **(A)** Overlays of the ground-truth SM segmentation and the original CT images. **(B)** Overlays of the ground-truth SM segmentation and the processed CT images. **(C)** An illustration of the workflow used to auto-select a single CT slice at the C3 vertebra for SM auto-segmentation. The auto-selected slice is the middle slice of the auto-segmented C3 section (33 mm in height) using a 3D ResUNet applied to the 3D volumetric CT image. **(D)** Auto-segmentation of SM using a selected C3 vertebra CT image using a 2D ResUNet model.

implementation are described in Appendix A. We implemented additional data augmentation to both image and mask patches to minimize overfitting, including random horizontal flips of 50% and random affine transformations with an axial rotation range of 12 degrees and a scale range of 10%. We used the Adam optimizer for computing the parameter updates and the soft Dice loss function. The models were trained for 300 iterations with a learning rate of $2\times10^{-4}$ for the first 250 iterations and $1\times10^{-4}$ for the remaining 50 iterations. The values for the Adam optimizer coefficients $\beta1$ and $\beta2$ were 0.9 and 0.999, respectively. Data augmentation and loss functions were provided by the MONAI framework (34). The final segmentations on the test set for both models were obtained by a majority vote on a pixel-by-pixel basis for all predicted segmentation masks by the 5-fold cross-validation sub-models (model ensemble), as described in a previous study (33).

## Model validation

For both the 3D ResUNet and 2D ResUNet models, we evaluated the performance on the corresponding cross-validation sets as well as the final ensemble segmentation on the test set using the Dice similarity coefficient (DSC) (36). Specific to the 3D model, we also evaluated the accuracy of the C3 section segmentation by quantifying the absolute difference between the slice locations of the mid-section of the C3 section predicted by the 3D model and the 2D CT ground-truth image (in mm). Specific to the 2D model, we compared the SM cross-sectional areas obtained using the SM ground-truth segmentation with 1. the 2D model predicted SM segmentations on the same ground-truth CT image (Pred_GT) and 2. the 2D model predicted SM segmentations on the slices auto-selected by the 3D model (Pred_C3). We evaluated the correlation between the SM cross-sectional areas using the Pearson correlation coefficient; we also used a two-sided Wilcoxon signed-rank test to determine if these SM values were significantly different. Additionally, to derive the SMI, we normalized the SM cross-sectional areas (in $cm^2$) with the patients' heights (in $m^2$). We then examined the correlation between the SMI values produced by the ground-truth and deep learning segmentations using the Pearson correlation coefficient; we also used a two-sided Wilcoxon signed-rank test to determine if these SMI values were significantly different. Based on previous work by Swartz et al. (15) and van Rijn-Dekker et al. (6), we used **Equation 1** to calculate the cross-sectional area

(CSA) at the L3 lumbar level based on the CSA at the C3 cervical level and subsequently **Equation 2** to calculate the lumbar SMI:

$$\begin{aligned} & CSA \;\; at \;\; L3 \;\; \left(cm^2\right) \\ & = \;\; 27.304 + 1.363 * CSA \;\; at \;\; C3 \;\; \left(cm^2\right) - 0.671 * age \;\; (years) + 0.640 \\ & * weight \;\; (kg) + 26.422 * sex(sex = 1 \;\; for \;\;\; female, \;\;\;\; 2 \;\;\; for \;\;\; male) \end{aligned} \qquad (Eq.1)$$

$$Lumbar \quad SMI \quad \left(\frac{cm^2}{m^2}\right) = \frac{CSA \;\; at \;\; L3 \;\; \left(cm^2\right)}{height^2 \left(m^2\right)} \qquad (Eq.2)$$

Based on previous work by Prado et al. (30), SMI thresholds of 52.4 $cm^2/m^2$ (males) and 38.5 $cm^2/m^2$ (females) were applied to lumbar SMI derived from SM ground-truth and deep learning segmentations to stratify patients by sarcopenia status ('normal' and 'depleted' muscle); body composition related measurements in the training and testing sets are shown in **Appendix B**. These stratifications were then used for Kaplan-Meier analysis to determine associations between sarcopenia status and overall survival probabilities. To determine the sarcopenia status for the whole data set (i.e., 390 patients), we implemented Kaplan-Meier analysis on the 5-fold cross-validation data and the test data. We aggregated the SMI estimated for each cross-validation fold (i.e., 60 patients per fold) using the corresponding trained 3D and 2D models in addition to the SMI for the test data using the average predictions of the five cross-validation models.

## Results

### 3D ResUNet *model performance: C3 section auto-segmentation*

The performance of the 3D ResUNet model for segmenting the C3 section of the neck is summarized in **Figure 2A**. When assessing the performance of each individual sub-model from our 5-fold cross-validation, the DSCs calculated between the predicted region segmentations and the ground-truth region segmentations were high and consistent between all training folds, with a mean (± standard deviation) DSC of 0.95 ± 0.01. When combining the cross-validation fold predictions using our ensemble approach, the performance on the test set increased to 0.96 ± 0.06. The middle slices of the predicted 3D regional segmentations for the test set were mostly within 4 mm of the ground-truth segmentation slice locations, with the greatest number of patients being within 1 mm (**Figure 2B**); the maximum outlier was at a distance of 10 mm.

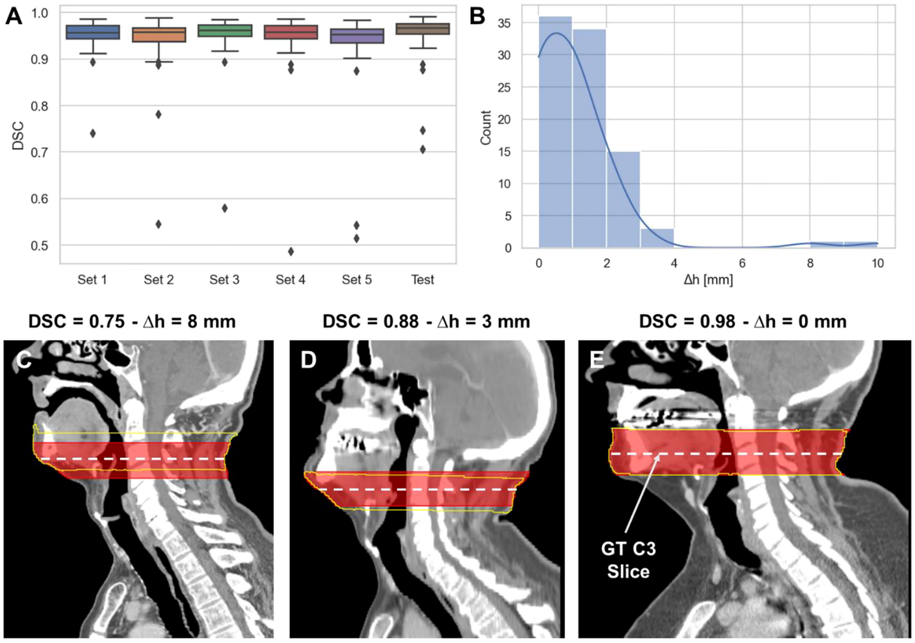


FIGURE 2
3D ResUNet model performance for segmentation of C3 vertebra section. **(A)** Boxplots of the Dice similarity coefficient (DSC) distributions for the 5-fold cross-validation data sets (Set 1 to Set 5 – 60 patients each) and the test data (90 patients). **(B)** Histogram of the absolute difference (in mm) of the C3 slice location at the middle slice of the auto-segmented C3 section and the location of the ground-truth manually segmented CT slice. Illustrative examples overlaying the C3 ground-truth segmentations (red) (33 mm centered at the ground-truth manually segmented CT slice) and predicted segmentations (yellow) on the CT images with different DSC values (low – 0.75 **(C)**, medium – 0.88 **(D)**, and high – 0.98 **(E)** performance compared to the mean DSC value of 0.95). The middle slice at the center of mass of the segmented C3 region was auto-selected for further skeletal muscle auto-segmentation by the 2D ResUNet model.

Examples of test set predictions for cases with low, medium, and high performance compared to the mean DSC are shown in Figures 2C–E. As can be visually confirmed, the low-performance case still generated a segmentation such that the middle slice was contained in the C3 region.

## 2D ResUNet model performance: SM auto-segmentation

The performance of our 2D ResUNet model for segmenting the C3 vertebra SM is summarized in Figure 3A. The DSCs calculated between the model-predicted segmentations and the ground-truth segmentations were high and consistent between all training folds, with a mean DSC of 0.95 ± 0.002. When combining the cross-validation fold predictions using our ensemble approach, the mean DSC performance on the test set remained consistent at 0.95 ± 0.02.

The cross-sectional areas derived from the 2D model predictions using both the ground-truth slice locations and auto-selected slice locations from the 3D ResUNet model were highly correlated to the cross-sectional areas derived from the ground-truth segmentations (Figure 3B). The predicted areas using the ground-truth slice locations had a Pearson r=0.98 ($p < 0.0001$) and nonsignificant Wilcoxon test (p=0.43). Similarly, the predicted areas using the auto-selected slice locations had a Pearson r=0.98 ($p < 0.0001$) and nonsignificant Wilcoxon test (p=0.22). Examples of test set predictions for cases with low, medium, and high performance compared to the mean DSC for predictions using ground-truth slice location are shown in Figures 3C–E. As can be visually confirmed, the low-performance case successfully generated a segmentation for musculature that was not included in the ground-truth segmentation. Moreover, the predictions using the auto-selected slice location from the 3D ResUNet model yielded virtually indistinguishable results for the low-performance and medium-

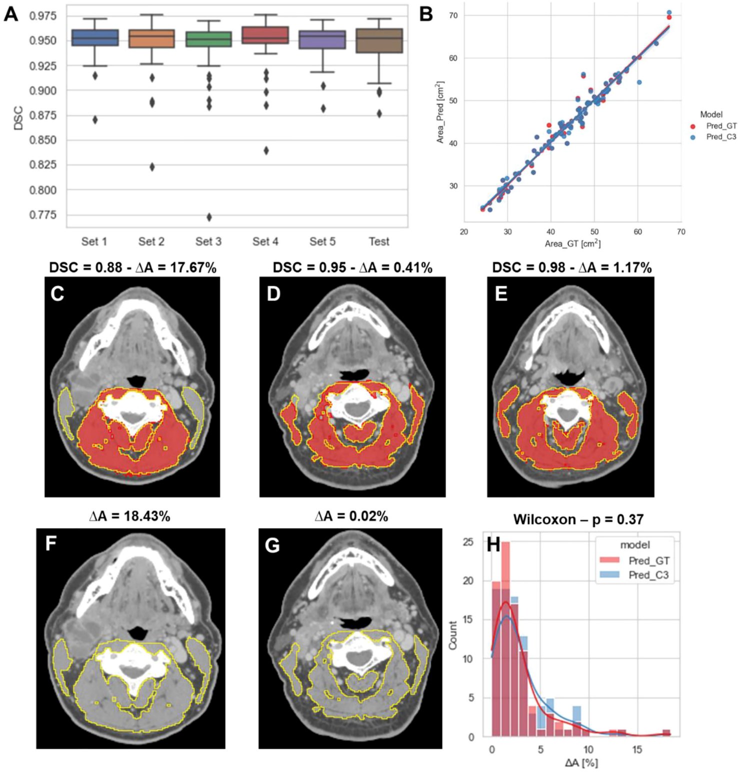


FIGURE 3
2D ResUNet model performance for segmentation of C3 skeletal muscle (SM). **(A)** Boxplots of the Dice similarity coefficient (DSC) distributions for the 5-fold cross-validation datasets (Set 1 to Set 5 – 60 patients each) and the test data (90 patients). **(B)** A scatter plot of the SM cross-sectional area using the ground-truth manual segmentation (x-axis) and the SM cross-sectional areas (y-axis) using predicted segmentations of the 2D ResUNet applied to the ground-truth CT image slice (Pred_GT) and the auto-selected CT image slice using the C3 section auto-segmentation (Pred_C3). Illustrative examples overlaying the skeletal muscle (SM) ground-truth segmentations (red) and predicted segmentations (yellow) on the same ground-truth CT images **(C-E)** and auto-selected CT images **(F, G)** with different DSC values (low – 0.88, medium - 0.95, and high – 0.98 compared to the mean estimated DSC value of 0.95). The auto-selected CT image for the high-performance example was identical to the ground-truth image and therefore provided the same segmentation as shown in panel C **(H)** Histogram of percentage difference of SM cross-sectional areas between ground-truth segmentations compared to the predicted SM cross-sectional areas (ΔA%) corresponding to the model using ground-truth slice location (red) or auto-selected slice location (blue).

performance cases (Figures 3F, G) and identical results for the high-performance case (Figure 3E). Finally, when investigating the percentage difference in cross-sectional areas between the model-generated and ground-truth segmentations, there was no significant difference when using the ground-truth slice location or the auto-selected slice location (p=0.37) (Figure 3H).

## SMI measurement comparisons

We compared SMI values for test set patients calculated using ground-truth SM segmentations with predicted SMI values calculated using SM segmentations generated from our 2D ResUNet models using the ground-truth slice location (Figure 4A) or auto-selected slice location (Figure 4B). Both model SM segmentations led to predicted SMI values that were highly correlated to the ground-truth SMI values. The predicted SMI values using the ground-truth slice location for males and females both had a Pearson r=0.98 ($p < 0.0001$) and nonsignificant Wilcoxon signed-rank tests (p=0.17 and p=0.43, respectively) compared to ground-truth SMI values. Similarly, the predicted SMI values using the auto-selected slice location for males and females had Pearson r values of 0.97 and 0.96, respectively (both $p < 0.0001$) and nonsignificant Wilcoxon signed-rank tests (p=0.19 and p=0.98, respectively) compared to the ground-truth SMI values.

## Survival analysis

The results of the overall survival analysis based on sarcopenia thresholds are shown in Figure 5. Independent of the method of SMI calculation (GT, Pred_GT, or Pred_C3), there were significant differences in overall survival of males between those with normal and depleted muscle tissue (Figures 5A–C), while females exhibited no significant differences (Figures 5D–F). Hazard ratios (95% confidence intervals) in males for GT, Pred_GT, and Pred_C3 were 1.82 (1.1-3.0), 1.95 (1.18-3.22), and 1.97 (1.19-3.25), respectively. Hazard ratios (95% confidence intervals) in females for GT, Pred_GT, and Pred_C3 were 2.76 (0.59-13.02), 3.4 (0.73-15.83), and 3.72 (0.8-17.31), respectively.

# Discussion

In this study, we utilized a multi-stage deep learning approach to segment the C3 region of the head and neck, auto-select a single representative slice, and auto-segment the corresponding SM. Our approach determined slice location and segmented SM with high accuracy when compared to ground-truth segmentations. By fully automating this workflow, we have enabled more rapid testing and application of sarcopenia-related clinical decision-making. To our knowledge, this is the first study to fully automate sarcopenia prediction based on non-abdominal HNC imaging.

We utilized both 2D and 3D ResUNet models in our approach. By decomposing the C3 detection and SM segmentation problem into two separate tasks, we ensure that accurate representations of patient anatomy are identified by the models (C3 region) and subsequently maximize performance for SM segmentation. While previous SM auto-segmentation studies often required specific slices as model inputs (21, 26) or utilized separate pre-processing software (23, 25), multi-stage deep learning methods have recently been adapted in this domain as well (22, 24). Both the 2D and 3D ResUNet models that make up our segmentation pipeline had high performance, with mean DSC values in the test set above 0.95. Importantly, the performance of our C3 SM segmentation model is comparable to that of previous L3 SM deep learning segmentation models, which also demonstrate test set DSCs of ~0.95 (21–26). Moreover, for cases with relatively low performance, we visually confirmed that results were reasonable, i.e., the auto-selected slice was still contained within the C3 region for the 3D model, and the correct musculature was

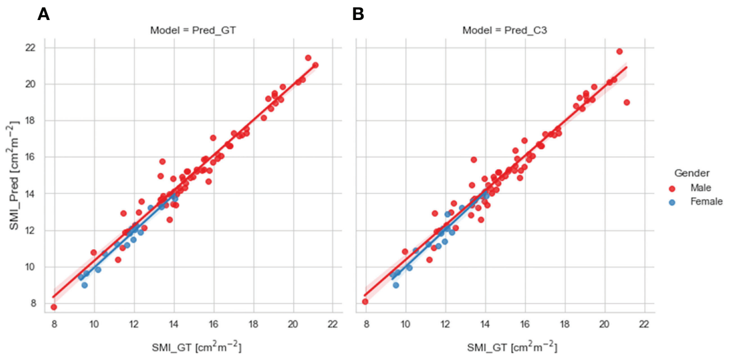


FIGURE 4
Scatter plots of the skeletal muscle index (SMI) values determined for test set patients (stratified by gender) using the ground-truth manual segmentation (x-axis) and those determined using predicted segmentations of the 2D ResUNet (y-axis) using **(A)** the ground-truth CT image slice (Pred_GT) and **(B)** the auto-selected CT image slice using the C3 section auto-segmentation (Pred_C3).

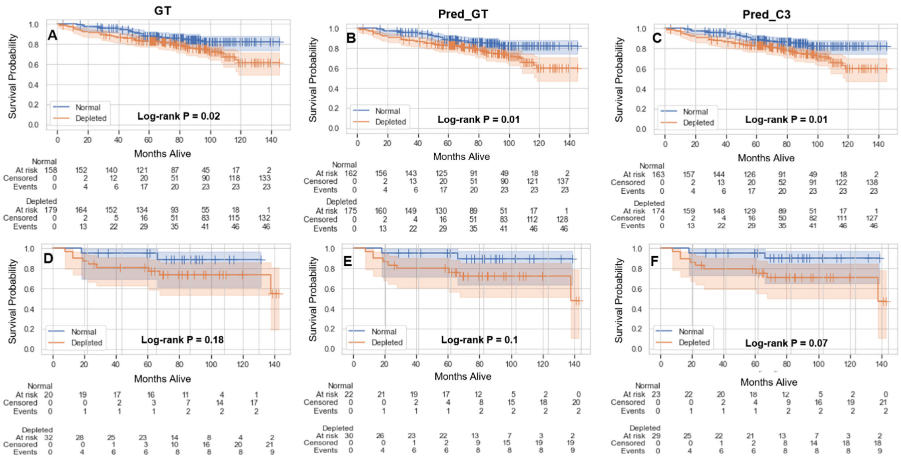


FIGURE 5
Kaplan-Meier plots showing overall survival probabilities (test and validation set combined, 390 patients) as a function of time in days for estimated skeletal muscle (SM) index (normal vs depleted) in male **(A-C)** and female **(D-F)** patients using the ground-truth SM segmentation (GT) **(A, D)**, auto-segmented SM using the ground-truth slice location (Pred_GT) **(B, E)**, and auto-segmented SM using the auto-selected slice location (Pred_C3) **(C, F)**.

segmented on the 2D model. Importantly, we also showed minimal differences in the 2D SM segmentation model regardless of how the slice location was determined, indicating the model is robust to the specific C3 slice location. Consistent with quantitative measures of segmentation performance, using our deep learning segmentations to calculate SMI demonstrated a high correlation with ground-truth SMI independent of gender stratification.

A recent meta-analysis by Surov et al. calculated the cumulative prevalence of sarcopenia in HNC patients at 42% (5), highlighting the clinical need for accurate quantification of sarcopenia. Several previous studies have demonstrated that SMI values can be used to stratify patients into sarcopenia subgroups that are strongly associated with prognostic outcomes (5–7). Using lumbar SMI conversion equations previously derived by Swartz et al. (15) and van Rijn-Dekker et al. (6) combined with validated SMI thresholds (12), we demonstrated that calculations based on our deep learning segmentations predict similar overall survival outcomes as calculations based on ground-truth segmentations. Moreover, p-values for all methods were significant for males but not females. These results are consistent with recent literature by Olson et al. (16) which emphasized that sarcopenia was associated with poor survival outcomes in males but not in females. Our results suggest that our automated methods are dependable for use in prognostic outcome prediction.

While our study presents encouraging results towards full automation of sarcopenia-related clinical decision-making for HNC patients, there were some limitations. First, we only tested our method on pre-therapy images. Importantly, some studies have suggested prognostic evidence for sarcopenia measurements based on alternative or additional timepoints (e.g., body composition changes) (7, 37). Therefore, further confirmatory work is needed to ensure our methods can be used accurately and reproducibly for intra-therapy and post-therapy imaging. Additionally, when defining sarcopenia using SMI cutoffs, we have relied on historically accepted thresholds in literature, but several recent developments in standardizing SMI values, e.g., through body mass index (38), have been proposed that warrant further exploration. We must also note that while no universal consensus on sarcopenia definitions currently exists, European consensus guidelines (39) emphasize the importance of evaluating muscle performance and strength in addition to muscle mass; therefore, by European consensus guidelines we have only investigated “presarcopenia” in this analysis. Moreover, we have limited our analysis to CT images as CT is the most common imaging modality for HNC radiotherapy treatment planning. However, the use of additional modalities for SM segmentation, i.e., MRI, as has been utilized in other studies (40), may warrant additional auto-segmentation investigations. Finally, while we believe current model performance is satisfactory for clinical applications as demonstrated by comparisons with ground-truth segmentations and SMI measures, different architectural choices or ensemble approaches could be further explored to improve performance.

## Conclusions

In summary, using open-source toolkits and public data, we applied 3D and 2D deep learning approaches to head and neck CT images to develop an end-to-end automated workflow for SM segmentation at the C3 vertebral level. When evaluated on independent test data, our fully-automated approach yielded mean DSCs of up to 0.96 for segmenting the C3 vertebra region and 0.95 for segmenting the corresponding SM. Cross-sectional areas and calculated SMI values derived from our approach were highly correlated to ground-truth (r>0.95), indicating their potential clinical acceptability. Moreover, our methods can be reliably combined with validated SMI thresholds for use in prognostic stratification. Our study is an essential first step towards fully-automated workflows for sarcopenia-related clinical-decision making. Future studies should consider incorporating additional imaging timepoints and modalities for automated sarcopenia prediction.

## Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: Segmentations generated in this project are available on Figshare, doi: 10.6084/m9.figshare.18480917. The original unprocessed images used in this project can be found on The Cancer Imaging Archive: https://wiki.cancerimagingarchive.net/display/DOI/Radiomics+outcome+prediction+in+Oropharyngeal+cancer.

## Ethics statement

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

## Author contributions

Study concepts: MN, KW, AG, BO, RJ, AM, KK, and CF; Study design: MN, KW, AG, BO, RJ, and AM; Data acquisition: AG, BO, RJ, DE-H, CD, VS, and MA; Quality control of data and algorithms: MN, KW, RH, JJ, JS, and KK; Data analysis and interpretation: MN, KW, AG, BO, RH, JJ, and JS; Manuscript editing: MN, KW, AG, BO, RJ, AM, and KK. All authors contributed to the article and approved the submitted version.

## Funding

This work was supported by the National Institutes of Health (NIH)/National Cancer Institute (NCI) through a Cancer Center Support Grant (CCSG; P30CA016672-44). MN is supported by an NIH grant (R01DE028290-01). KW is supported by the Dr. John J. Kopchick Fellowship through The University of Texas MD Anderson UTHealth Graduate School of Biomedical Sciences, the American Legion Auxiliary Fellowship in Cancer Research, and an NIH/National Institute for Dental and Craniofacial Research (NIDCR) F31 fellowship (1 F31DE031502-01). AG received funding from the National Cancer Institute (K08 245188, R01 CA264133) and the American Association for Cancer Research/Mark Foundation "Science of the Patient" Award (20-60-51-MARK). BO received funding from the Radiologic Society of North America Research Medical Student Grant (RMS2026). VS received funding from The University of Texas, Graduate School of Biomedical Sciences Graduate research assistantship. CF received funding from the NIH/NIDCR (1R01DE025248-01/R56DE025248); an NIH/NIDCR Academic-Industrial Partnership Award (R01DE028290); the National Science Foundation (NSF), Division of Mathematical Sciences, Joint NIH/NSF Initiative on Quantitative Approaches to Biomedical Big Data (QuBBD) Grant (NSF 1557679); the NIH Big Data to Knowledge (BD2K) Program of the NCI Early Stage Development of Technologies in Biomedical Computing, Informatics, and Big Data Science Award (1R01CA214825); the NCI Early Phase Clinical Trials in Imaging and Image-Guided Interventions Program (1R01CA218148); an NIH/NCI Pilot Research Program Award from the UT MD Anderson CCSG Radiation Oncology and Cancer Imaging Program (P30CA016672); an NIH/NCI Head and Neck Specialized Programs of Research Excellence (SPORE) Developmental Research Program Award (P50CA097007); and the National Institute of Biomedical Imaging and Bioengineering (NIBIB) Research Education Program (R25EB025787).

## Acknowledgments

The authors acknowledge Sunita Patterson (Research Medical Library, MD Anderson Cancer Center) for editorial assistance.

## Conflict of interest

CF has received direct industry grant support, speaking honoraria, and travel funding from Elekta AB.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fonc.2022.930432/full#supplementary-material

## References

1. Anjanappa M, Corden M, Green A, Roberts D, Hoskin P, McWilliam A, et al. Sarcopenia in cancer: Risking more than muscle loss. *Tech Innov Patient Support Radiat Oncol Elsevier* (2020) 16:50–7. doi: 10.1016/j.tipsro.2020.10.001

2. Zhao J-Z, Zheng H, Li L-Y, Zhang L-Y, Zhao Y, Jiang N. Predictors for weight loss in head and neck cancer patients undergoing radiotherapy: A systematic review. *Cancer Nurs LWW* (2015) 38(6):E37–45. doi: 10.1097/NCC.0000000000000231

3. Powrózek T, Dziwota J, Małecka-Massalska T. Nutritional deficiencies in radiotherapy-treated head and neck cancer patients. *J Clin Med Multidiscip Digital Publ Inst* (2021) 10(4):574. doi: 10.3390/jcm10040574

4. Ghadjar P, Hayoz S, Zimmermann F, Bodis S, Kaul D, Badakhshi H, et al. Impact of weight loss on survival after chemoradiation for locally advanced head and neck cancer: secondary results of a randomized phase III trial (SAKK 10/94). *Radiat Oncol Springer* (2015) 10(1):1–7. doi: 10.1186/s13014-014-0319-y

5. Surov A, Wienke A. Low skeletal muscle mass predicts relevant clinical outcomes in head and neck squamous cell carcinoma. a meta analysis. *Ther Adv Med Oncol SAGE Publ SAGE UK: London England* (2021) 13:17588359211008844. doi: 10.1177/17588359211008844

6. van Rijn-Dekker MI, van den Bosch L, van den Hoek JGM, Bijl HP, van Aken ESM, van der Hoorn A, et al. Impact of sarcopenia on survival and late toxicity in head and neck cancer patients treated with radiotherapy. *Radiother Oncol Elsevier* (2020) 147:103–10. doi: 10.1016/j.radonc.2020.03.014

7. Findlay M, White K, Stapleton N, Bauer J. Is sarcopenia a predictor of prognosis for patients undergoing radiotherapy for head and neck cancer? A meta-analysis. *Clin Nutr Elsevier* (2021) 40(4):1711–8. doi: 10.1016/j.clnu.2020.09.017

8. Han A, Bokshan SL, Marcaccio SE, DePasse JM, Daniels AH. Diagnostic criteria and clinical outcomes in sarcopenia research: A literature review. *J Clin Med Multidiscip Digital Publ Inst* (2018) 7(4):70. doi: 10.3390/jcm7040070

9. Hua X, Liu S, Liao J-F, Wen W, Long Z-Q, Lu Z-J, et al. When the loss costs too much: A systematic review and meta-analysis of sarcopenia in head and neck cancer. *Front Oncol Front* (2020) 9:1561. doi: 10.3389/fonc.2019.01561

10. Cho Y, Kim JW, Keum KC, Lee CG, Jeung HC, Lee IJ. Prognostic significance of sarcopenia with inflammation in patients with head and neck cancer who underwent definitive chemoradiotherapy. *Front Oncol Front* (2018) 8:457. doi: 10.3389/fonc.2018.00457

11. Stone L, Olson B, Mowery A, Krasnow S, Jiang A, Li R, et al. Association between sarcopenia and mortality in patients undergoing surgical excision of head and neck cancer. *JAMA Otolaryngol Neck Surg Am Med Assoc* (2019) 145(7):647–54. doi: 10.1001/jamaoto.2019.1185

12. Grossberg AJ, Chamchod S, Fuller CD, Mohamed AS, Heukelom J, Eichelberger H, et al. Association of body composition with survival and locoregional control of radiotherapy-treated head and neck squamous cell carcinoma. *JAMA Oncol Am Med Assoc* (2016) 2(6):782–9. doi: 10.1001/jamaoncol.2015.6339

13. Fattouh M, Chang GY, Ow TJ, Shifteh K, Rosenblatt G, Patel VM, et al. Association between pretreatment obesity, sarcopenia, and survival in patients with head and neck cancer. *Head Neck Wiley Online Library* (2019) 41(3):707–14. doi: 10.1002/hed.25420

14. Chamchod S, Fuller CD, Mohamed ASR, Grossberg A, Messer JA, Heukelom J, et al. Quantitative body mass characterization before and after head and neck cancer radiotherapy: A challenge of height-weight formulae using computed tomography measurement. *Oral Oncol Elsevier* (2016) 61:62–9. doi: 10.1016/j.oraloncology.2016.08.012

15. Swartz JE, Pothen AJ, Wegner I, Smid EJ, Swart KM, de Bree R, et al. Feasibility of using head and neck CT imaging to assess skeletal muscle mass in head and neck cancer patients. *Oral Oncol* (2016) 62:28–33. doi: 10.1016/j.oraloncology.2016.09.006

16. Olson B, Edwards J, Degnin C, Santucci N, Buncke M, Hu J, et al. Establishment and validation of pre-therapy cervical vertebrae muscle quantification as a prognostic marker of sarcopenia in patients with head and neck cancer. *Front Oncol* (2022) 12. doi: 10.3389/fonc.2022.812159

17. Tajbakhsh N, Jeyaseelan L, Li Q, Chiang JN, Wu Z, Ding X. Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation. *Med Image Anal Elsevier* (2020) 63:101693. doi: 10.1016/j.media.2020.101693

18. Zhou T, Ruan S, Canu S. A review: Deep learning for medical image segmentation using multi-modality fusion. *Array Elsevier* (2019) 3:100004. doi: 10.1016/j.array.2019.100004

19. Bakator M, Radosav D. Deep learning and medical diagnosis: A review of literature. *Multimodal Technol Interact Multidiscip Digital Publ Inst* (2018) 2(3):47. doi: 10.3390/mti2030047

20. Naser MA, Deen MJ. Brain tumor segmentation and grading of lower-grade glioma using deep learning in MRI images. *Comput Biol Med Elsevier Ltd* (2020) 121:103758. doi: 10.1016/j.compbiomed.2020.103758

21. Amarasinghe KC, Lopes J, Beraldo J, Kiss N, Bucknell N, Everitt S, et al. A deep learning model to automate skeletal muscle area measurement on computed tomography images. *Front Oncol Front Media SA* (2021) 11. doi: 10.3389/fonc.2021.580806

22. Kanavati F, Islam S, Arain Z, Aboagye EO, Rockall A. Fully-automated deep learning slice-based muscle estimation from CT images for sarcopenia assessment. *arXiv Prepr arXiv* (2020) 218:200606432. doi: 10.48550/arXiv.2006.06432

23. Pickhardt PJ, Perez AA, Garrett JW, Graffy PM, Zea R, Summers RM. Fully automated deep learning tool for sarcopenia assessment on CT: L1 versus L3 vertebral level muscle measurements for opportunistic prediction of adverse clinical outcomes. *Am J Roentgenol Am Roentgen Ray Soc* (2021) 1–8. doi: 10.2214/AJR.21.26486

24. Burns JE, Yao J, Chalhoub D, Chen JJ, Summers RM. A machine learning algorithm to estimate sarcopenia on abdominal CT. *Acad Radiol Elsevier* (2020) 27 (3):311–20. doi: 10.1016/j.acra.2019.03.011

25. Graffy PM, Liu J, Pickhardt PJ, Burns JE, Yao J, Summers RM. Deep learning-based muscle segmentation and quantification at abdominal CT: Application to a longitudinal adult screening cohort for sarcopenia assessment. *Br J Radiol Br Inst Radiol* (2019) 92(1100):20190327. doi: 10.1259/bjr.20190327

26. Paris MT, Tandon P, Heyland DK, Furberg H, Premji T, Low G, et al. Automated body composition analysis of clinically acquired computed tomography scans using neural networks. *Clin Nutr Elsevier* (2020) 39(10):3049–55. doi: 10.1016/j.clnu.2020.01.008

27. Elhalawani H, Mohamed ASR, White AL, Zafereo J, Wong AJ, Berends JE, et al. Matched computed tomography segmentation and demographic data for oropharyngeal cancer radiomics challenges. *Sci Data Nat Publ Group* (2017) 4:170077. doi: 10.1038/sdata.2017.77

28. Grossberg A, Elhalawani H, Mohamed A, Mulder S, Williams B, White AL, et al. HNSCC [ dataset ]. *Cancer Imaging Arch* (2020). doi: 10.7937/k9/tcia.2020.a8sh-7363

29. 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 Springer* (2013) 26(6):1045–57. doi: 10.1007/s10278-013-9622-7

30. Prado CMM, Lieffers JR, McCargar LJ, Reiman T, Sawyer MB, Martin L, et al. Prevalence and clinical implications of sarcopenic obesity in patients with solid tumours of the respiratory and gastrointestinal tracts: A population-based study. *Lancet Oncol Elsevier* (2008) 9(7):629–35. doi: 10.1016/S1470-2045(08)70153-0

31. Wahid KA, Olson B, Jain R, Grossberg AJ, El-Habashy D, Dede C, et al. Muscle and adipose tissue segmentations at the C3 vertebral level for sarcopenia-related clinical decision-making in patients with head and neck cancer. *medRxiv* (2022) 2022:1. doi: 10.1101/2022.01.23.22269674

32. Anderson BM, Wahid KA, Brock KK. Simple python module for conversions between dicom images and radiation therapy structures, masks, and prediction arrays. *Pract Radiat Oncol Elsevier* (2021) 11(3):226–9. doi: 10.1016/j.prro.2021.02.003

33. Naser MA, Wahid KA, van Dijk LV, He R, Abdelaal MA, Dede C, et al. Head and neck cancer primary tumor auto segmentation using model ensembling of deep learning in pet-ct images. *3D Head Neck Tumor Segmentation PET/CT Challenge Lect Notes Comput Sci Springer Cham* (2022) 13209:121–33. doi: 10.1007/978-3-030-98253-9_11

34. doi: 10.5281/zenodo.4323059

35. Wahid KA, Ahmed S, He R, van Dijk LV, Teuwen J, McDonald BA, et al. Evaluation of deep learning-based multiparametric MRI oropharyngeal primary tumor auto-segmentation and investigation of input channel effects: Results from a prospective imaging registry. *Clin Transl Radiat Oncol* (2022) 32:6–14. doi: 10.1016/j.ctro.2021.10.003

36. Taha AA, Hanbury A. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. *BMC Med Imaging BioMed Central* (2015) 15(1):1–28. doi: 10.1186/s12880-015-0068-x

37. Ferrão B, Neves PM, Santos T, Capelas ML, Mäkitie A, Ravasco P. Body composition changes in patients with head and neck cancer under active treatment: A scoping review. *Support Care Cancer* (2020) 28(10):4613–25. doi: 10.1007/s00520-020-05487-w

38. Martin L, Birdsell L, Macdonald N, Reiman T, Clandinin MT, McCargar LJ, et al. Cancer cachexia in the age of obesity: skeletal muscle depletion is a powerful prognostic factor, independent of body mass index. *J Clin Oncol Off J Am Soc Clin Oncol United States* (2013) 31(12):1539–47. doi: 10.1200/JCO.2012.45.2722

39. Cruz-Jentoft AJ, Baeyens JP, Bauer JM, Boirie Y, Cederholm T, Landi F, et al. Sarcopenia: European consensus on definition and diagnosis: Report of the European working group on sarcopenia in older people. *Age Ageing* (2010) 39 (4):412–23. doi: 10.1093/ageing/afq034

40. Zwart AT, Becker J-N, Lamers MJ, Dierckx RA, de Bock GH, Halmos GB, et al. Skeletal muscle mass and sarcopenia can be determined with 1.5-T and 3-T neck MRI scans, in the event that no neck CT scan is performed. *Eur Radiol Springer* (2021) 31(6):4053–62. doi: 10.1007/s00330-020-07440-1

frontiers | Frontiers in Oncology

TYPE Original Research
PUBLISHED 27 September 2022
DOI 10.3389/fonc.2022.973067

OPEN ACCESS

EDITED BY
David Pasquier,
Oscar Lambret Centre, France

REVIEWED BY
Francesco Ricchetti,
Sacro Cuore Don Calabria Hospital
(IRCCS), Italy
Frederik Crop,
Centre Oscar Lambret, France

*CORRESPONDENCE
Emil Fredén
emil.freden@gmail.com

SPECIALTY SECTION
This article was submitted to
Radiation Oncology,
a section of the journal
Frontiers in Oncology

RECEIVED 19 June 2022
ACCEPTED 08 September 2022
PUBLISHED 27 September 2022

CITATION
Fredén E, Tilly D and Ahnesjö A (2022)
Adaptive dose painting for
prostate cancer.
*Front. Oncol.* 12:973067.
doi: 10.3389/fonc.2022.973067

# Adaptive dose painting for prostate cancer

Emil Fredén[1]*, David Tilly[2,3] and Anders Ahnesjö[2]

[1]Department of Oncology, Södersjukhuset, Stockholm, Sweden, [2]Department of Genetics, Immunology and Pathology, Medical Radiation Sciences, Uppsala University, Uppsala, Sweden, [3]Department of Medical Physics, Uppsala University Hospital, Uppsala, Sweden

**Purpose:** Dose painting (DP) is a radiation therapy (RT) strategy for patients with heterogeneous tumors delivering higher dose to radiation resistant regions and less to sensitive ones, thus aiming to maximize tumor control with limited side effects. The success of DP treatments is influenced by the spatial accuracy in dose delivery. Adaptive RT (ART) workflows can reduce the overall geometric dose delivery uncertainty. The purpose of this study is to dosimetrically compare ART and non-adaptive conventional RT workflows for delivery of DP prescriptions in the treatment of prostate cancer (PCa).

**Materials and methods:** We performed a planning and treatment simulation study of four study arms. Adaptive and conventional workflows were tested in combination with DP and Homogeneous dose. We used image data from 5 PCa patients that had been treated on the Elekta Unity MR linac; the patients had been imaged in treatment position before each treatment fraction (7 in total). The local radiation sensitivity from apparent diffusion coefficient maps of 15 high-risk PCa patients was modelled in a previous study. these maps were used as input for optimization of DP plans aiming for maximization of tumor control probability (TCP) under rectum dose constraints. A range of prostate doses were planned for the homogeneous arms. Adaptive plans were replanned based on the anatomy-of-the-day, whereas conventional plans were planned using a pre-treatment image and subsequently recalculated on the anatomy-of-the-day. The dose from 7 fractions was accumulated using dose mapping. The endpoints studied were the TCP and dose-volume histogram metrics for organs at risk.

**Results:** Accumulated DP doses (adaptive and conventional) resulted in high TCP, between 96-99%. The largest difference between adaptive and conventional DP was 2.6 percentage points (in favor of adaptive DP). An analysis of the dose per fraction revealed substantial target misses for one patient in the conventional workflow that—if systematic—could jeopardize the TCP. Compared to homogeneous prescriptions with equal mean prostate dose, DP resulted in slightly higher TCP.

**Conclusion:** Compared to homogeneous dose, DP maintains or marginally increases the TCP. Adaptive DP workflows could avoid target misses compared to conventional workflows.

KEYWORDS

dose painting, dose-response modeling, adaptive radiation therapy, prostate cancer, MR-linac

# 1 Introduction

When a tumor's radiation sensitivity is heterogeneous, radiation therapy (RT) with conventional homogeneous dose prescriptions will not maximize the tumor's response per delivered radiant energy ("integral dose"). Under the assumption that individual cancer cells respond independently to each other, several authors have shown theoretically that stronger curative effects per delivered dose is achieved by prescribing higher dose to radiation resistant tumor sub-volumes and less to sensitive ones (1, 2). With an overall reduction of dose, it is assumed that the overall risk for side effects can be reduced. The approach of differentiating the tumor dose over its sub-volumes requires that pretreatment functional imaging (3) can be used to spatially map a quantity that correlates with radiation sensitivity. The spatial differentiation of dose prescriptions on a per voxel basis has been referred to as 'dose painting by numbers' (DPBN) (4). By building upon the work of Vogelius et al. (5), Grönlund et al. (6) developed a *failure-driven* DPBN formalism incorporating clinical endpoint data and information from functional imaging. Their formalism was applied to prostate cancer (PCa) in a simulation study based on imaging of the apparent diffusion coefficient (ADC) with MRI (7) for which they used a correlation between ADC and an assumed Gleason score (GS) (8) as an intermediate step for scoring and modelling dose-response variations. The tumor dose-response was then modelled based on treatment failure frequencies versus biopsied GS from a retrospective study of patients treated with homogeneous dose (9, 10). The feasibility of delivering spatially differentiated DPBN plans for PCa patients was investigated in a follow up simulation study considering dose delivery uncertainties (11). They also concluded that the potential of dose painting increases as the geometric uncertainties of treatment delivery decrease.

The MR-linac (MRL) enables an adaptive workflow taking advantage of the soft-tissue contrast of magnetic resonance imaging (12). A key feature of the MRL is the ability to perform plan adaptation prior to each given fraction based on imaging of the patient in treatment position. The present work aims to investigate if the reduced geometric uncertainties obtained by adaptive RT (ART) can increase the potential of dose painting compared to conventional, non-adaptive, treatment workflows. To this end, we present a treatment simulation study where the DPBN formalism for PCa by Grönlund et al. (7, 11) is combined with adaptive workflow features provided by the Elekta Unity MRL system (13). For reference we also included study arms with homogeneous dose escalation. The more peaked dose distributions used in prostate SBRT (14) could be interesting as reference as well. However, to avoid bias caused by the arbitrariness of SBRT dose max locations we preferred homogeneous dose arms as reference. As primary endpoint we used the calculated tumor control probability (TCP), and as secondary endpoints dose-volume histogram (DVH) based metrics for the dose to organs at risk (OARs). Previous studies have investigated adaptive dose painting strategies for head and neck cancers (15), but to our knowledge the present simulation study is the first to combine DPBN with daily replanning for PCa.

# 2 Materials and methods

## 2.1 Overview of study design

In the present treatment simulation study, we investigate if an adaptive RT workflow can increase the potential of dose painting in terms of TCP and/or reduced dose to risk organs. For planning and treatment simulation we used a research version of a commercial treatment planning system (TPS) (RayStation 10.1.130.16, RaySearch Laboratories, Stockholm, Sweden) together with purpose designed scripts. We simulated two different dose prescription strategies (homogeneous dose vs. DPBN) combined with two different treatment delivery workflows (conventional vs. adaptive), i.e., in total four study arms labelled: Homo-conv for homogeneous dose with conventional delivery, Homo-adap for homogenous dose with adaptive delivery, DPBN-conv for dose painting by numbers with conventional delivery, and DPBN-adap for dose painting by numbers with adaptive delivery. The DPBN-conv and DPBN-adap plans were *constrained by rectum dose-volume criteria*, but no upper limit was set on the dose to individual voxels of the prostate, i.e. these plans had the planning aim *'treat-*

*to-tolerance'*. The homogeneous dose plans were optimized towards fixed target dose, aiming at high target coverage while keeping the dose to the rectum as low as possible. For the homogeneous arms we thus implemented target coverage constraints together with rectum objectives. We optimized a set of homogeneous plans with a range of target dose levels to study the relationship between target dose, rectum load, and tumor control, and to set the DPBN plans in a clinical context. Our study design differs from that of Grönlund et al. (11) who kept the mean dose to the prostate equal between the homogeneous and dose painted plans; moreover, they only simulated conventional treatment flows. The resulting TCP and rectum DVH metrics were calculated based on the accumulated dose from simulated full treatment courses; these metrics were used to compare the four study arms. Table 1 summarizes the key parameters for the study arms. A more detailed description is given in the following sctions.

## 2.2 Patient data and case generation

For this project we had access to two sets of patient data from which we constructed 75 fictive PCa cases by fusing image data sets from the two groups. The first set consisted of images for 5 intermediate-risk PCa patients that had been treated to 42.7 Gy with hypofractionation (6.1 Gy×7 fx) on the Elekta Unity MRL at Akademiska sjukhuset ethical approval reference number: 2019–03050 (Uppsala, Sweden). For each of these 5 patients (A1-A5) we had access to one reference T2w MRI (acquired prior to treatment) and 7 fractions of T2w MRI, as well as the corresponding structure sets including the prostate, seminal vesicles (SV), rectum, bladder, anal canal, penile bulb, and femoral heads. The intrapatient anatomical variations captured in these 8 image sets allowed us to simulate a full hypofractionated treatment course of 7 fractions. We selected 15 patients included in the PARAPLY phase 2 trial with Umeå board ethical approval reference numbers 2013/154-31 and 2015/ 75-32 from the high-risk PCa patient group included in Grönlund's previous works (7, 11) and used the ADC maps of their prostates. These 15 ADC maps were registered and fused to each of the 5 patient reference geometries from the MRL patient group through deformable image registration (DIR), resulting in planning reference ADC maps for 15×5 = 75 fictive PCa cases. In this work, a *case* is defined as the *combination of a specific patient anatomy and a single realistic prostate ADC (spatial) distribution for a high-risk PCa*. The thus fitted ADC maps were then through a subsequent intrapatient DIR operation assigned to each fraction's geometry, thus assuming that the pre-treatment ADC values were invariantly determining the Gleason values over the full treatment course. All ADC distributions were visually checked after the transformations to minimize the risk for artifacts entering into the image data flow. In addition, the mean and spread of the ADC distributions were evaluated both before and after a deformation for validation. A schematic overview of the process is shown in Figure 1.

## 2.3 Treatment simulation of a case

An overview of the simulation flow for the four study arms is shown in Figure 2. The main operations include generation of a treatment plan, modelling of the geometric uncertainties, and finally dose accumulation over all treatment fractions for endpoint calculation of the TCP and DVH metrics.

### 2.3.1 Setup of treatment plans

Optimized treatment plans were created for all four study arms and for each of the 75 cases. All were planned for hypofractionation (7 fractions), with 7-field IMRT (static MLC, 70 segments in total). The gantry angles were set equal to those clinically used for the MRL treatments. The Homo-adap and DPBN-adap plans were optimized on the anatomy-of-the-

TABLE 1 The four study arms simulated.

| | Homo-conv homogeneous doseconventional delivery | Homo-adap homogeneous doseadaptive delivery |
|---|---|---|
| Plan generation | Reference plan | Plan of the day |
| Margin | 6 mm | 3 mm |
| Target *constraints* | $D_{98\%} > 0.95D_p$, $D_{2\%} < 1.05D_{p,}$ $D_p$ ϵ{43.89,44.89,…,60.89}Gy | |
| Rectum *objectives* | $V_{33Gy}$<30%, $V_{38Gy}$<15%, $V_{41Gy}$<10% | |
| | **DPBN-conv dose paintingconventional delivery** | **DPBN-adap dose paintingadaptive delivery** |
| Plan generation | Reference plan | Plan of the day |
| Minimax optimization | 6 mm | 3 mm |
| Target goal | Maximize TCP | |
| Rectum *constraints* | $V_{33Gy}$<30%, $V_{38Gy}$<15%, $V_{41Gy}$<10% | |

For the two arms with homogeneous dose, the plans were designed based on rectum dose objectives that can be violated with a penalty in favor of covering the target with the homogeneous prescription dose $D_p$, while for the dose painting by numbers plans, we used rectum dose constraints that cannot be violated.

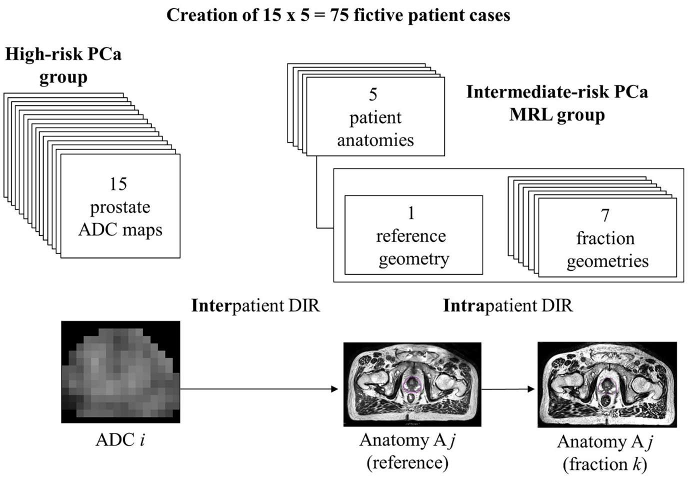


FIGURE 1
Data from two different patient groups were combined to generate images to represent 75 test cases. For the 5 patients from the group consisting of intermediate-risk PCa patients treated on Elekta Unity MRL (i.e., 'MRL group'), we had 1 reference geometry and 7 fraction geometries. From the second group, we had ADC maps of 15 high-risk PCa patients. The bottom part of the figure shows the process of deforming the ADC maps for patients of the high-risk PCa group to the reference geometry for a patient from the MRL group. The (deformed) reference ADC maps were subsequently deformed to fit the 7 fraction geometries. Patients in the MRL group are labelled A1-A5.

day, whereas the conventional Homo-conv and DPBN-conv plans were planned on each patient's reference geometry. The conventional plans were subsequently recalculated on each fraction image based on rigid CTV-to-CTV translations to simulate treatments with the field setup translated based on gold marker registrations. For the conventional plans, the isocenter was set at the volumetric center of the prostate CTV. For the adaptive plans, we extracted the isocenters from the

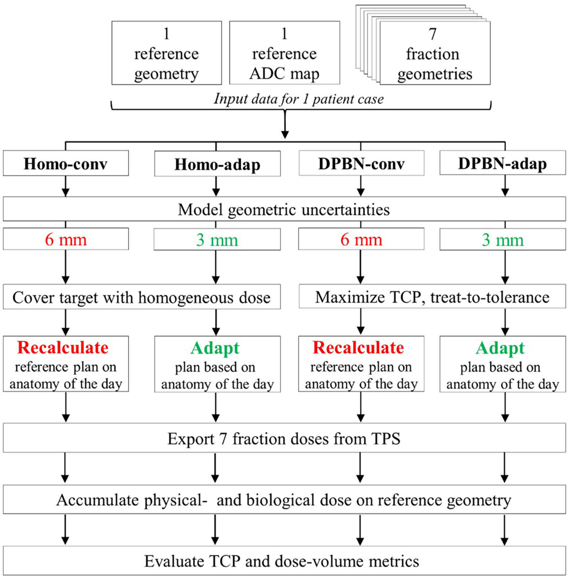

1 reference geometry
1 reference ADC map
7 fraction geometries
Input data for 1 patient case
Homo-conv
Homo-adap
DPBN-conv
DPBN-adap
Model geometric uncertainties
6 mm
3 mm
6 mm
3 mm
Cover target with homogeneous dose
Maximize TCP, treat-to-tolerance
Recalculate reference plan on anatomy of the day
Adapt plan based on anatomy of the day
Recalculate reference plan on anatomy of the day
Adapt plan based on anatomy of the day
Export 7 fraction doses from TPS
Accumulate physical- and biological dose on reference geometry
Evaluate TCP and dose-volume metrics

FIGURE 2
Process chart for treatment simulation of a patient case showing the key differences between the four simulated study arms. The CTV to PTV margins for the two treatment workflows were 3 mm vs 6 mm, respectively. Conventional plans (Homo-conv, DPBN-conv) were optimized using the reference geometry and were subsequently recalculated on the anatomy-of-the-day. Adaptive plans (Homo-adap, DPBN-adap) were optimized using the anatomy-of-the-day. The 7 fraction doses were exported and finally accumulated on the reference geometry to facilitate evaluation of primary TCP- and secondary DVH endpoints.

clinical MRL plans as the fixed isocenter cannot in general be placed in the target volume. Dose calculations were performed with Monte Carlo (1% uncertainty, 3 mm voxel size). Appropriate tissue compositions and densities were assigned to the body and bony anatomy to facilitate the calculation of dose. As the patient couch of Elekta Unity is highly attenuating, the bed was included in the dose calculations for the MRL workflow, and the angle intervals with the highest change in attenuation were avoided as per clinical practice. Magnetic field effects modelled as described by Malkov and Rogers (16) had been added to the Monte Carlo engine of the TPS (17).

### 2.3.2 Modelling of the geometric uncertainties

A major advantage of adaptive workflows is the ability to compensate at plan generation for interfraction organ deformations. The increased accuracy for adaptive workflow patients has two benefits: 1) the margins added to form the planning target volume (PTV) for generation of homogeneous plans can be reduced to lower the risk of inducing normal tissue toxicities, and 2) it has been shown that larger TCP increases can be obtained through dose painting when geometric dose delivery uncertainties are small (11).

We divided the overall geometric uncertainty of the two treatment workflows into subcomponents and assigned to each an estimate of the standard deviation (SD) based on published data. For both workflows, these subcomponents included the residual positioning errors resulting from intrafraction motion (SD 1 mm) (18), interobserver target delineation variations (SD 1 mm) (19), and an estimate of the finite precision of the treatment machine (1 mm). The residual effect of interfraction prostate deformation was only considered for the conventional workflow for which it was assigned an SD of 1 mm (20); in the adaptive workflow, prostate deformation was taken into account *via* redelineation of the prostate on the anatomy-of-the-day. For the conventional workflows, we also added the combined effect of image registration- and table translation uncertainties, which was estimated to have an SD of 1 mm (19). The SD of different components were added in quadrature and used as input to margin calculations based on van Herk's margin formula (21) in which systematic uncertainties (preparation errors) are denoted by Σ and random uncertainties (daily patient setup variations) are denoted by sigma Since we planned for hypofractionation (7 fractions), we used van Herk's formula

$$M(\ \Sigma_{\text{eff}}, \sigma_{\text{eff}})\ = 2.5\ \ \Sigma_{\text{eff}} + 0.7\sigma_{\text{eff}} \tag{1}$$

with *effective* uncertainty components (22, 23), adjusted to consider the finite number of treatment fractions $N$:

$$\Sigma^2_{\mathbf{eff}} = \Sigma^2 + \frac{\sigma^2}{N}, \qquad \sigma^2_{\mathbf{eff}} = \ \sigma^2\left(1 - \frac{1}{N}\right). \tag{2}$$

The resulting (isotropic) prostate margins calculated with equations (1, 2) for the adaptive and conventional workflows were 3 mm and 6 mm, respectively. For the SV, we used the lower (LR: 5, AP: 7, SI: 7 mm) and upper limit (LR: 6, AP: 9, SI: 9 mm) of the anisotropic margins specified in the ESTRO reference (20) for the adaptive and conventional workflows, respectively. Homogeneous plans were optimized with standard CTV-to-PTV margins whereas DPBN plans were created using minimax optimization (24) with "robustness distances" set to the same values as the CTV-to-PTV margins; the minimax optimizer generates a set of treatment scenarios with patient setup displacements along three axes, and aims to find a plan which is optimal for the worst case of these scenarios (i.e., a plan which is robust to geometric uncertainties). For each objective, the software allows it to be set as 'robust' or not, i.e. evaluated for all the scenarios. We selected only the TCP objective as robust.

### 2.3.3 Homogeneous dose prescriptions

We were interested to see whether DPBN, limited by normal tissue constraints, would be superior to homogeneous dose escalation. Dose escalation to the prostate is in general limited by gastrointestinal (GI) toxicities (e.g. diarrhea, rectal bleeding, proctitis), genitourinary (GU) toxicities (e.g. dysuria, hematuria, obstruction) and erectile dysfunction (25). A set of homogeneous plans were generated with prostate prescription doses ranging from 43.89 Gy up to 60.89 Gy in increments of 1 Gy, where 43.89 Gy (EQD2 = 91.6 $\text{Gy}_{1.93}$) corresponds to the dose level used in the previous works of Grönlund et al. (7, 11). In total, 7x18 homogeneous adaptive plans and 18 homogeneous conventional plans were optimized for each of the 5 patient anatomies. The 7x18+18 homogeneous plans were assigned to each of the 15 cases corresponding to the particular patient anatomy. The homogeneous plans per patient anatomy could be reused because no ADC information was used for planning of the homogeneous dose arms, Homo-conv and Homo-adap. For target dose uniformity, we used dose-volume *constraints* requiring that 98% of the target volume receives at least 95% of the prescribed dose, and that at most 2% of the volume receives doses larger than 105% of the prescribed dose.

### 2.3.4 Dose painting prescriptions

For the DPBN plans, the Grönlund et al. (11) TCP formalism (summarized briefly in Appendix A) was used to maximize the TCP subject to rectum constraints. The ADC maps were downsampled to the resolution of the dose grid, and subsequently transformed to Gleason score probabilities through a 'low precision' ADC-to-Gleason mapping constructed by Grönlund et al. (7, 11).

### 2.3.5 Dose to risk organs

Dose to the OARs other than the rectum could be held below the clinically set tolerance levels (femoral heads: $D_{2\%}$<30 Gy, bladder: $D_{\text{mean}}$<34 Gy) using a single 'dose falloff' objective (effectively aiming for a tight dose gradient around the target volume). For the rectum, we implemented the three volume-at-dose (VaD) metrics ($V_{33Gy} < 30\%$, $V_{38Gy} < 15\%$, $V_{41Gy} < 10\%$, where $V_D$ is the volume of the organ receiving doses larger than $D$) as *objectives* during homogeneous plan generation, and as *constraints* during DPBN plan generation. Rectum *objectives* were used in the homogeneous arms since rectum constraints could potentially conflict with the imposed target coverage constraints. We used these rectum DVH metrics since they are clinically implemented for plan evaluation at our clinic, Akademiska sjukhuset (Uppsala, Sweden). For all study arms, we *complemented* the clinically used evaluation criteria with $D_{2\%}$ <42.7 Gy (soft) objectives to limit high doses in the rectum, bladder and remaining normal tissues whenever possible (target goals were prioritized).

### 2.3.6 Evaluation of TCP and DVH endpoints based on accumulated dose from a full treatment course

For each case and study arm, we evaluated the TCP and rectum DVH endpoints based on the accumulated dose from 7 fractions. Biological dose (EQD2) was accumulated to calculate TCP according to Grönlund's formalism described in Appendix A. To be consistent with Grönlund's earlier work we used an α/β ratio of 1.93 Gy to calculate EQD2 (7). As input to the TCP calculations, we used the accumulated EQD2, the down sampled reference ADC map, and assigned to the full vesicle volume a Gleason score of 6, which is the lowest risk category in the TCP model (this assumption was made since we did not have any ADC information for the SV). The vesicle volume was included in the TCP formalism to be able to evaluate the effect of potential SV target loss (for all study arms, the SVs were prescribed a near-min dose of 43.89 Gy corresponding to an SV control probability larger than 99%). Physical dose was accumulated to generate cumulative DVHs. Out of the OARs, we decided to mainly focus on the rectum since it is the most dose limiting organ for PCa.

### 2.3.7 Deformable image registrations for mapping fraction doses to a common reference frame

For each case the dose was accumulated through mapping of all fraction doses to the patient's reference frame *via* the geometric transformation determined by a DIR. The result of the DIR consists of a rigid transformation matrix describing rotations and translations and a displacement vector field (DVF) describing the deformations. The 'hybrid DIR' option in the TPS was used to calculate DIRs between the reference- and each of the 7 fraction geometries. 'Hybrid DIR' is based on the ANACONDA algorithm (26) and employs three non-linear terms: an image similarity term, a grid regularization term, and a term considering the similarity between structures delineated in the two geometries. The prostate CTV, SV, rectum, and bladder were selected as 'controlling ROIs' (RayStation term for registration guiding structures) with weight 0.8.

The accumulated dose accuracy is sensitive to uncertainties in the DIR (27) and therefore it is important to assess the registration quality. To this end, we calculated the DICE score (DSC) (28) and Hausdorff distance (29) for the prostate CTV, SV, rectum, and bladder. Both measures quantify the 'similarity' (i.e., agreement) between structures defined in two different reference frames. A dosimetric evaluation was also performed to assess the quality of the registrations for the purpose of dose accumulation. This was done by comparing (per fraction) rectum VaD evaluated before and after dose mapping.

## 3 Results

In the present work we sought to investigate the potential benefit of plan adaptation for prescriptions based on DPBN. The study arms based on homogeneous prescriptions were used for reference to set the dose painting results in a clinical context. In Figures 3, (4) reference plans—showcasing the different study arms—for 1 of the 75 patient cases are presented together with the ADC map used to generate the particular DPBN plans. Compared to the homogeneous dose plans, the DPBN plans have distinct high dose regions that follow a low ADC structure. According to the model, low ADC structures are indicative of radiation resistant foci.

We begin the presentation of the results with a section comparing the overall difference in TCP- and rectum DVH endpoints calculated for the *conventional- and adaptive dose painting arms*, DPBN-conv and DPBN-adap. In the following section, these results are then grouped according to patient anatomy and compared against the results from homogeneous dose escalation in Homo-conv and Homo-adap. The results presented in the first two sections were calculated based on the accumulated dose from a full treatment course, and it is evident that rectum constraints were violated despite our intention to 'treat-to-tolerance'. Therefore, we break down the endpoint calculation per fraction in the third section to eliminate the role of dose mapping uncertainties that might confound potential differences between the conventional and adaptive workflows. In the fourth section, we then illustrate two mechanisms for the apparent rectum constraint violations. One of these mechanisms is inherent to the conventional arms, for which the use of non-representative rectum volumes can explain the observed constraint violations; the second mechanism deals with the complex task of accumulating dose to non-rigid organs that experience substantial volume changes over the treatment course. In the last section, we present the similarity measures calculated to analyze the quality of the deformable image registrations.

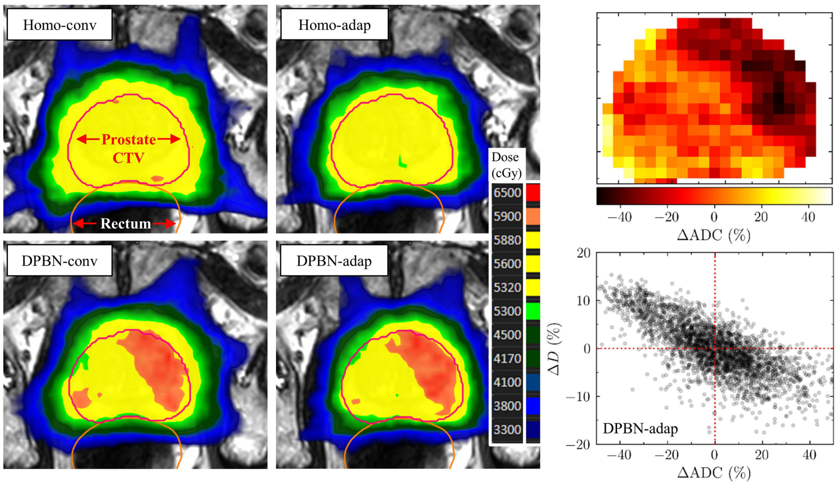


FIGURE 3

In the top 2x2 panels, transverse 2D-slices of 4 *reference* (i.e., not dose accumulated) plans are presented—showcasing the 4 respective study arms—for 1 of the 75 patient cases. The lower left panel shows the corresponding slice of the ADC map used to generate the DPBN plans. The lower right panel shows the prostate voxel dose (percentage deviation of 56 Gy) as a function of ADC (percentage deviation of ADC$_{\text{mean}}$=1046 $10^{-6}$mm$^2$s$^{-1}$) for the DPBN-adap reference plan.

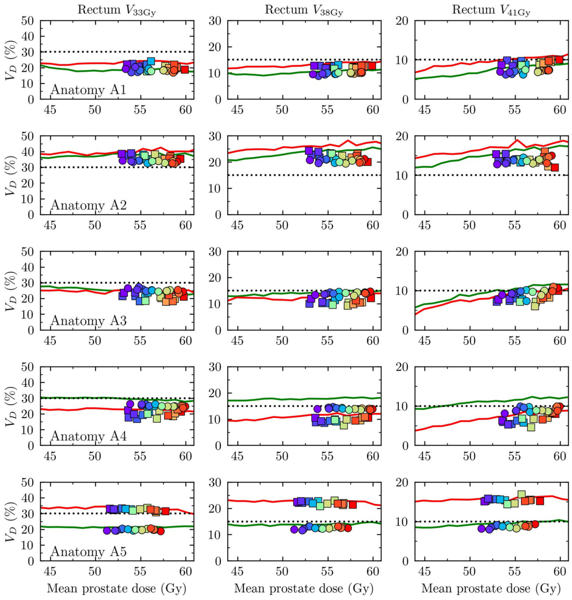


FIGURE 4

Rectum VaD as a function mean prostate dose. Each row corresponds to one of the five patient anatomies A1-A5 with A1 uppermost. The three columns correspond to $V_{33Gy}$, $V_{38Gy}$, $V_{41Gy}$, respectively. Rectum constraints are indicated by dotted lines ($V_{33Gy}$<30%, $V_{38Gy}$<15%, $V_{41Gy}$<10%). Homo-conv: red solid line, Homo-adap: green solid line, DPBN-conv: squares, DPBN-adap: circles. The squares/circles correspond to the 15 cases per anatomy and are color coded according to their mean Gleason score, as estimated through an ADC-to-Gleason-score probability mapping (the colormap scale is shown in Figure 5).

## 3.1 Conventional versus adaptive dose painting

Using the accumulated dose from 7 fractions, the conventionally and adaptively dose painted arms resulted in high tumor control probability for all cases (TCP: 96-99%). The difference per case between the two workflows was small; the mean difference was 0.5 percentage points, and the maximum difference was 2.6 percentage points. The adaptively dose painted arm resulted, on average, in lower rectal doses. However, analyzed per case, both negative and positive differences in rectum DVH metrics were observed (e.g., the difference in $V_{41\mathrm{Gy}}$ ranged between -7.9 and 4.5 percentage points; a negative difference implies that the adaptive workflow resulted in lower rectum dose). Table 2 summarizes the condensed DPBN results.

## 3.2 Homogeneous dose escalation versus dose painting to tolerance

Rectal doses varied slowly as a function of mean prostate dose in the homogeneously dose escalated arms. In other words, the mean prostate dose could be escalated without substantially increasing the rectum load (in terms of the clinically used evaluation criteria). Figure 4 shows the rectal doses as a function of mean prostate dose for all four study arms, presented separately for the five patient anatomies A1-A5. The DPBN arms resulted in equal or lower rectal doses for a given mean prostate dose compared to the homogeneous arms. For anatomies A3-A5, the DPBN-adap arm was successful in 'treating-to-tolerance', since at least one of the three rectum VaD metrics lies precisely on the tolerance limit and the other VaD metrics lie *on* or *below* the set tolerance limits. For anatomy A1, the rectum VaD metrics for the 15 DPBN-adap cases lie well below all three tolerance limits, and we thus failed to 'treat-to-tolerance'. On the other end, all rectum constraints were violated for the 15 DPBN-adap cases belonging to anatomy A2. Note, however, that these results were based on the accumulated dose from 7 fractions; we further explore these results in the following sections. A further interesting observation from Figure 4 is that the conventionally planned arms resulted in lower rectal doses compared to the adaptively planned arms for cases belonging to A3 and A4. To summarize, the outcome with regards to rectal doses from adaptive and conventional workflows depends to a large extent on the particular patient anatomy; moreover, dose painting is at least non-inferior to homogeneous dose and has the potential to decrease the rectum load. We did not prioritize near-maximum doses ($D_{2\%}$) for OARs in the optimization since we adopted the clinically used evaluation OAR criteria (30); however, condensed $D_{2\%}$ results for the four study arms are presented in Table 3.

The dose-response curves (i.e., the TCP as a function of mean prostate dose) for the homogeneously dose escalated arms are presented separately for the five patient anatomies A1-A5 in Figure 5 together with the resulting TCP from conventional and adaptive dose painting. Note that 15 dose-response curves were calculated for each anatomy using the same homogeneous dose but with different ADC maps; the resulting TCP is a function of dose, ADC, and implicitly a function of tumor volume (since the calculation is a product over the prostate voxels). The volume dependence explains why the dose-response curves are different for the five patient anatomies even though the same set of 15 ADC distributions were used. In the TCP model used, low ADC values indicate a high probability for high Gleason scores, and thus a worse prognosis. As is evident from Figure 5, the DPBN arms resulted in

TABLE 2 Comparison of the calculated endpoints for DPBN-conv and DPBN-adap.

| | DPBN-conv | DPBN-adap | Δ : adap-conv |
|---|---|---|---|
| TCP (%) | 98.6 [96.4, 99.3] | 99.1 [98.7, 99.4] | 0.5 [-0.3, 2.6] |
| Rectum $V_{33\mathrm{Gy}}$ (%) | 26 [17, 39] | 24 [17, 34] | -1.7 [-13.8, 9.4] |
| Rectum $V_{38\mathrm{Gy}}$ (%) | 15 [8, 24] | 14 [9, 21] | -1.2 [-11.1, 6.1] |
| Rectum $V_{41\mathrm{Gy}}$ (%) | 11 [5, 17] | 10 [6, 15] | -0.9 [-7.9, 4.5] |

The last column (Δ) shows the difference in endpoints evaluated per case. Data are presented as mean [min, max].

TABLE 3 Comparison of near-maximum doses ($D_{2\%}$) for the four study arms.

| | Homo-conv | DPBN-conv | DPBN-adap | Homo-adap |
|---|---|---|---|---|
| TCP (%) | 97.9 [95.3, 99.5] | 98.6 [96.4, 99.3] | 99.1 [98.7, 99.4] | 98.3 [95.3, 99.5] |
| Prostate $D_{\mathrm{mean}}$ (Gy) | 55 [54, 56] | 56 [52, 60] | 56 [51, 60] | 56 [55, 57] |
| Prostate $D_{2\%}$ (Gy) | 58 [57, 59] | 61 [56, 65] | 61 [56, 64] | 58 [56, 59] |
| Rectum $D_{2\%}$ (Gy) | 50 [47, 53] | 49 [45, 55] | 50 [46, 55] | 51 [49, 54] |
| Bladder $D_{2\%}$ (Gy) | 50 [47, 55] | 46 [41, 59] | 45 [41, 52] | 50 [47, 53] |

In this comparison, a single dose level per anatomy was selected for the homogeneous arms, corresponding to the mean DPBN dose. Data is presented as mean [min, max].

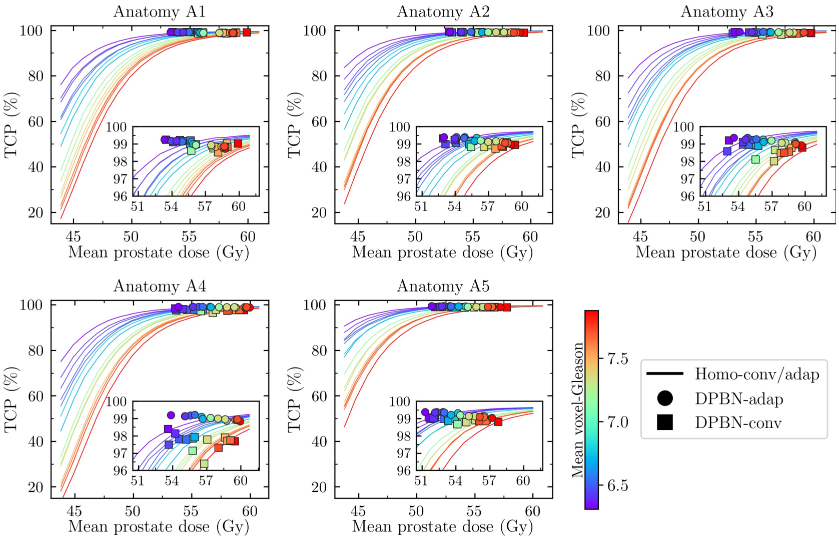


FIGURE 5
TCP as a function of mean prostate dose. Each panel corresponds to one of five patient anatomies (A1-A5). For each anatomy there are 15 cases corresponding to different prostate ADC maps. The cases are color coded according to their mean Gleason score, as estimated through an ADC-to-Gleason-score probability mapping. For a given mean prostate dose, DPBN is superior to homogeneous dose if the associated marker lies above the solid line (e.g., for cases belonging to A4, DPBN-conv markers are observed *below* the corresponding solid lines).

mean prostate doses located in the flat region of the dose-response curves. In this region there are diminishing marginal returns for additional increases in dose (i.e., the dose-response gradient is small); this explains why there is a spread in DPBN mean prostate doses resulting in similar TCP (approximately 99%) for all cases; the TPS optimizer pushed high Gleason score cases towards higher doses, since the marginal return is greater for these cases.

In Figure 5, the cases are color coded according to their mean Gleason score, as determined through an ADC-to-Gleason score mapping. The homogeneously dose escalated arms can be compared against the DPBN arms; DPBN is superior if the TCP for the corresponding case lies above the (homogeneous) dose-response curve at equal mean dose. For most cases, DPBN-adap and DPBN-conv was superior to homogeneous dose, except for some cases belonging to anatomy A4 for which DPBN-conv resulted in lower TCP. DPBN-adap resulted in better or similar TCP compared to DPBN-conv, with the largest difference observed for cases belonging to anatomy A4.

## 3.3 Breakdown of endpoint calculations per treatment fraction for the DPBN arms

The endpoints calculated based on accumulated dose from 7 fractions resulted in several rectum constraint violations even though we imposed hard constraints in the optimization. This was not expected for the DPBN-adap arm; the failure to meet these constraints was attributed to the dose mapping procedure. To eliminate any uncertainty associated with dose mapping, we calculated TCP and rectum DVH endpoints for each treatment fraction. Since the TCP model and DVH metrics are based on the total dose from 7 fractions, we scaled the fraction doses accordingly prior to the endpoint calculations. In Figure 6, the difference between DPBN-adap and DPBN-conv in TCP and rectum $V_{41Gy}$ is shown for all cases and treatment fractions; the results have been grouped according to anatomy A1-A5 (for each anatomy there are 15x7 = 105 data points). The analysis reveals a marked difference between DPBN-adap and DPBN-conv for some fractions belonging to anatomy A3, for which the difference in fraction specific TCP was larger than 95 percentage points. For most cases and treatment fractions, the difference in TCP was small. It appears that the difference in rectum load between the two workflows depends to a large extent on patient anatomy.

For cases belonging to A3, for which the adaptive workflow resulted in higher rectum load, the increase in rectum load resulted from adaptive avoidance of target misses to maintain the high TCP of ~99%. A target miss in the conventional workflow is illustrated in Figure 7. For the particular case illustrated, the nominal reference TCP was 98.6%, whereas the TCP for fraction 5 was 2.5% (the corresponding DPBN-adap fraction resulted in 99% TCP). The target misses suggest that the conventional margins used were insufficient to account for the interfraction prostate rotations- and deformations of anatomy A3.

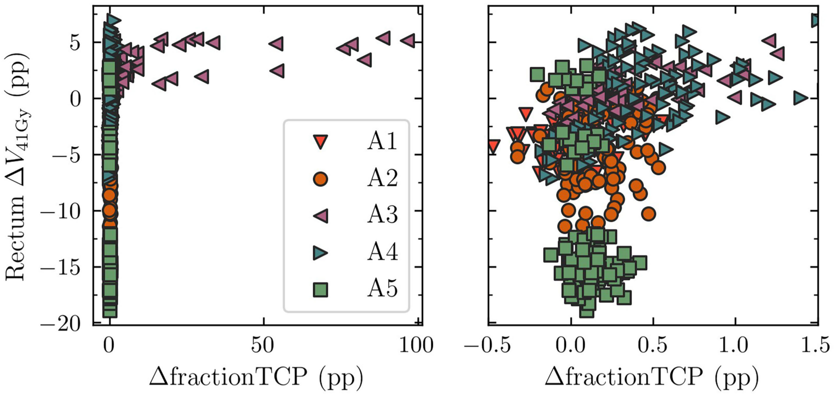


FIGURE 6
Comparison of rectum $V_{41Gy}$ and TCP between DPBN-adap and DPBN-conv. The results have been grouped according to patient anatomy A1-A5. The right and left panel contain the same data but have different horizontal scales. For anatomies A3 and A4 the difference in TCP between the two workflows was relatively large. DPBN-adap resulted in higher rectum dose for A3 and A4, but 'used' the extra rectum load to avoid target misses. $\Delta V_{41Gy}$=$(V_{41Gy})_{adap}$-$(V_{41Gy})_{conv}$, ΔfractionTCP=fractionTCP$_{adap}$-fractionTCP$_{conv}$, pp = percentage points.

## 3.4 Two mechanisms explaining rectum constraint violations in the DPBN arms

Since hard rectum constraints were violated for several cases despite the intention to treat-to-tolerance, we wanted to explain how this came about, and at the same time verify that no mistakes had entered the simulation pipeline. To this end, we looked at two separate operations in the pipeline: 1) *recalculation of DPBN-conv plans*, and 2) *dose mapping of DPBN-adap plans to the reference geometry*. As our analysis reveals, the optimized

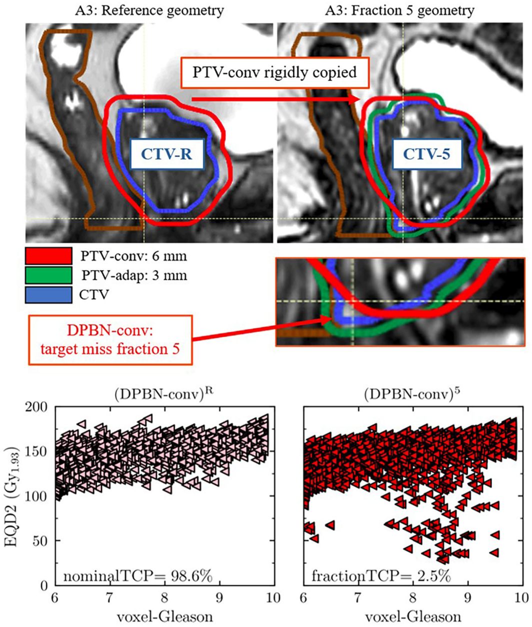


FIGURE 7
The upper panels show an obvious target miss in the DPBN-conv arm for patient anatomy A3 in fraction 5. The fraction specific CTV-5 is clearly outside of the conventional PTV-conv. The fraction specific TCP, calculated as if the fraction in question was delivered for an entire treatment, was 2.5% compared to 98.6% in the reference plan $(\text{DPBN-conv})^R$. The lower panels show the biological dose (EQD2) that was planned $(\text{DPBN-conv})^R$ (lower left panel) and delivered in fraction 5 $(\text{DPBN-conv})^5$ (lower right panel) to each voxel of the prostate, as a function of predicted Gleason score. The corresponding DPBN-adap plan resulted in 99% TCP.

treatment plans in fact met the imposed rectum constraints (illustrated in Figures 8 and 9, respectively), but the two operations independently altered the *planned* rectum DVH metrics. In Figure 11, the relationship between rectal volume changes, and the change in rectum DVH metrics is illustrated. The recalculation operation—which *simulates* the delivery of a conventional reference plan on the anatomy-of-the-day—makes apparent the complexity and limitations in planning and evaluating treatments using dose-volume metrics for organs that experience significant volume changes during the course of therapy; for consistency between planned and delivered dose, the reference geometry must be representative of the whole treatment course.

The dose mapping operation was implemented to be able to accumulate dose at the voxel level, and ultimately to evaluate the total dose from a full (simulated) treatment course. By comparing each fraction dose *before and after* dose mapping, we found that the rectum DVH metrics were not robust to such transformations (Figure 9). Accumulation of dose to organs that experience significant volume changes during the course of therapy is a difficult problem that needs attention.

### 3.4.1 Evaluation of deformable image registrations using Dice score and Hausdorff distance

Dose accumulation using DIR may lead to inaccurate dosimetric evaluation of treatments due to geometric errors in the DIR. Therefore, the DIRs were checked using two common geometrical properties, the DSC and Hausdorff distance. Figure 10 shows the level of agreement for the prostate CTV and rectum structures for each treatment fraction and patient anatomy. The relatively low DSC values and large Hausdorff distances of the rectum for patient A2 can help explain why the change in $V_{41\mathrm{Gy}}$ (as illustrated in Figure 11) after dose mapping was relatively large for cases belonging to patient A2 (for which the rectum size doubled in one fraction compared to reference). An explanation for the fact that contours do not agree perfectly may be the use of multiple, potentially conflicting, guiding structures, and the fact that we used a combination of guiding structures and image information. It is important to note that, even in the case of *perfect registration of structures* (high DSC and low Hausdorff distance), the change in rectum DVH metrics may still be substantial.

## 4 Discussion

A motivation behind our study were the results from the simulation study by Grönlund et al. (11) showing that there is a relationship between *geometric dose delivery uncertainties* and *the potential TCP gains from dose painting*. Adaptive RT is one possible strategy to reduce those uncertainties. Given the simulations with daily replanning in the present study, only small differences in terms of TCP were observed between adaptive and conventional dose painting. However, a notable difference was observed for some treatment fractions where the target was partially missed in the conventional workflow. Such target misses—especially if they are systematic—could very well compromise tumor control and ultimately the outcome for the patient motivating the use of positional feedback and adaptive measures.

TCP gains from dose escalation (in the studied range of 44 Gy to 60 Gy) were substantially larger compared to the TCP gains observed from dose painting alone; the flatness of the dose-response curve at high target doses yields diminishing returns for the additional cost of redistributing the dose using local radiation sensitivity information. The demonstrated gains are likely clinically insignificant but compared to homogeneous dose

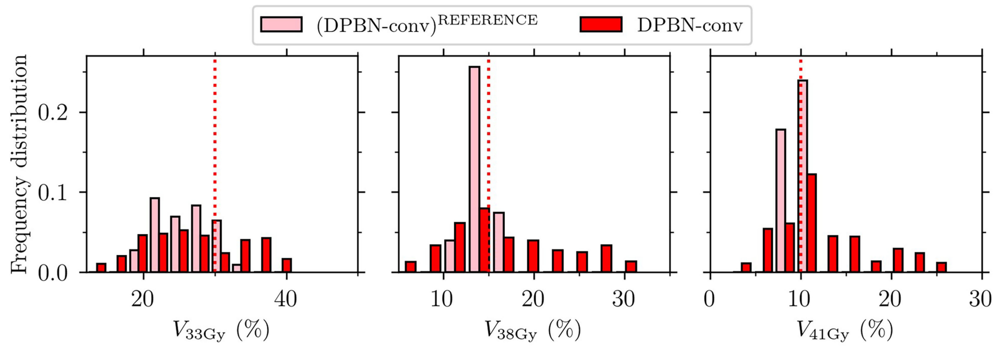


FIGURE 8
Rectum VaD for the conventional dose painted reference plans (DPBN-conv)$^{\mathrm{REFERENCE}}$ and the recalculated fraction doses DPBN-conv. Rectum constrains (indicated by the vertical dotted lines) were violated for the evaluated fraction doses, even though the constraints were met in the nominal reference plans.

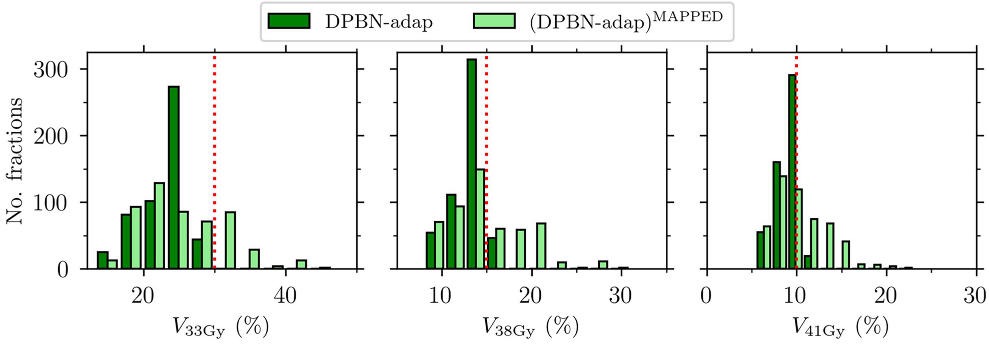


FIGURE 9
Rectum VaD for the adaptive dose painted fraction doses DPBN-adap and the corresponding mapped fraction doses (DPBN-adap)$^{\mathrm{MAPPED}}$. Rectum constrains (indicated by the vertical dotted lines) were violated after dose mapping.

escalation, dose painting prescriptions have the potential to decrease dose to the rectum while achieving similar or slightly larger probability for tumor control. Similar advances are likely also for prostate SBRT where narrow margins and steeper dose gradients are adopted to spare the rectum, while peaked dose distributions (higher dose maximum inside the PTV) are likely to increase TCP. However, the positions of dose maxima do not—in general—coincide with radiation resistant foci.

A different approach to lowering rectal doses (and reducing target motion) is to utilize a displacement device (e.g., hydrogel) to physically increase the space between the prostate and rectum (31, 32) but we have not considered such means in our planning study. The rectum DVH metrics used in this work were adapted from a clinical protocol with a homogeneous prescription dose of 42.7 Gy. The validity in using these metrics during dose escalation can be questioned. Moreover, the use of relative rectum VaD, as inherited from the current clinical practice, could potentially contribute to suboptimal results for the adaptive arms; an increased rectum volume a particular day (compared to reference) would allow for an increased absolute rectum volume receiving dose that day. On the contrary, conventional planning is oblivious to future gastrointestinal states. In other words, information about potential interfraction rectum deformations is unknown at the time of conventional planning, potentially resulting in lower-than-planned rectum doses and target misses. In our fraction-by-

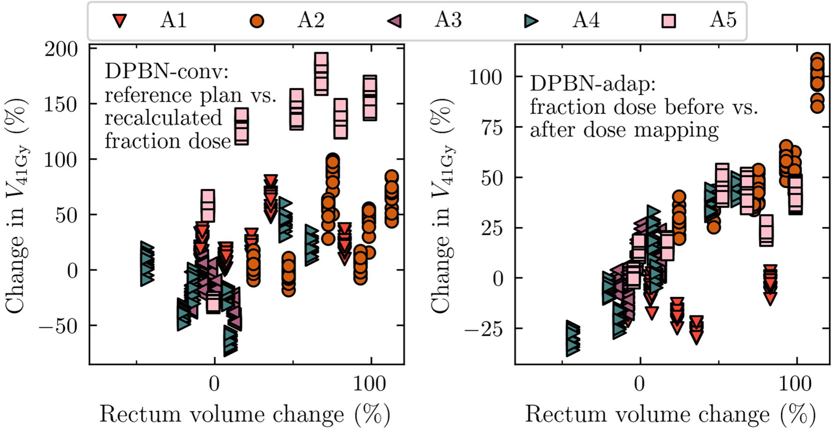


FIGURE 10
Relative change in rectum $V_{41\mathrm{Gy}}$ as a function of relative change in rectum volume (compared to the reference rectum volume delineated on the pre-treatment image). The 75 cases have been grouped according to anatomy A1-A5. The left panel shows the effect of recalculating the conventional dose painted reference plans on each fraction geometry, whereas the right panel shows the effect of mapping the adaptive dose painted fraction doses to the reference geometry.

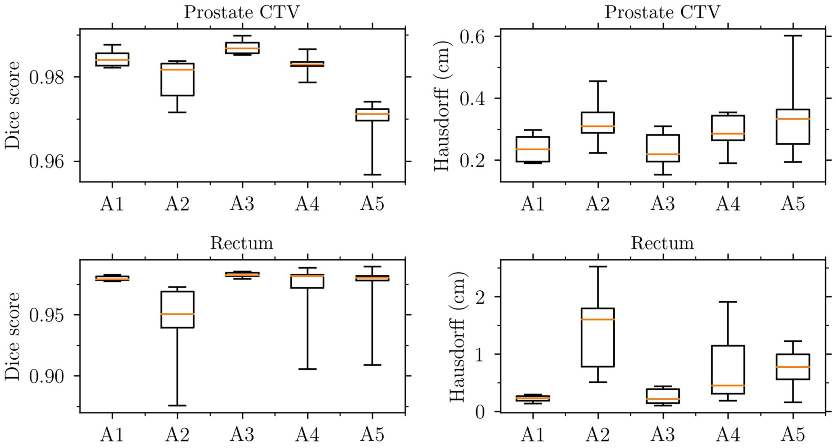


FIGURE 11
Evaluation of the deformable image registrations used for dose mapping. Dice score (left panels) and Hausdorff distance (right panels) for the prostate CTV (upper panels) and rectum (lower panels).

fraction analysis we indeed observed such results: for cases belonging to two of the patient anatomies, adaptive dose painting resulted in higher rectum doses compared to conventional dose painting; for these cases the extra rectum load was balanced by a relatively high TCP increase due to adaptive avoidance of target misses. Additionally, for some cases, substantial loss of SV target coverage was observed in the conventional workflow, whereas in the adaptive workflow loss of target coverage could be avoided. For patient anatomy A5, the SV target coverage ($V_{95\%}$) was as low as 53% in one fraction in the homogeneous conventional workflow (prostate prescription dose 44 Gy), whereas the adaptive workflow resulted in an SV target coverage of $V_{95\%}$=98%. The corresponding control probabilities of the vesicle volume were 85% and 100% for the conventional and adaptive workflow, respectively.

The combination of using ADC maps from one set of patients and applying them to a second set of patients enabled the current study (to our knowledge, this approach has not been used by others). However, the mapping of ADC values from one patient using DIR to a second patient results in a new ADC map due to the different shapes and sizes of the CTVs. The geometric accuracy of the DIRs was checked but the deformed ADC maps used for optimization of DPBN plans were hence not from actual patients. Nevertheless, the resulting ADC maps were visually checked, and the ADC histograms were compared before and after deformation (to ensure that the distributions of ADC values were consistent). We thus conclude that the spatial ADC distributions were realistic enough to investigate the potential of different dose painting strategies. The accuracy and precision with which functional imaging can be used to map out radiation resistant foci will influence the potential of DPBN strategies to increase TCP (11); the model used in this work has inherent limitations related to the uncertainty range of the ADC-to-Gleason mapping function. Further studies, exploring the potential of DPBN, should be conducted in parallel with technological advances in functional imaging and emerging knowledge on potential biomarkers for identifying radiation resistant foci.

The use of relative rectum volumes may confound the relationship between rectal wall doses and induced rectal toxicities. Future studies should be set up to improve the limitations related to the use of relative volumes; perhaps, in the era of daily replanning, one should focus on the matter that matters (e.g., the rectal wall) instead of scoring dose to feces. Future studies should also investigate the potential benefit of dose painting at different target dose levels along the dose-response curve since at low mean doses—where the gradient of the dose-response curve is higher—larger increases in TCP can in theory be achieved with DPBN (7).

## Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

## Ethics statement

The studies involving human participants were reviewed and approved by Etikprövningsmyndigheten, Uppsala. The patients/participants provided their written informed consent to participate in this study.

## Author contributions

All authors contributed equally to the study design. The corresponding author implemented the simulation pipeline, performed optimizations, and processed the resulting data. All authors contributed equally to the analysis as well as writing the manuscript. All authors contributed to the article and approved the submitted version.

## Acknowledgments

We acknowledge Agnes Angerud and Erik Traneus at RaySearch Laboratories for their help with the implementation of the MRL beam model and TCP objective in RayStation. Tufve Nyholm and Camilla Thellenberg are acknowledged for providing access to part of the image data for the PARAPLY trial.

## Conflict of interest

DT is employed part time by Elekta. However, this work was carried out solely as part of his employment with Uppsala University Hospital and affiliation with Uppsala University.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## References

1. Webb S, Evans PM, Swindell W, Deasy JO. A proof that uniform dose gives the greatest TCP for fixed integral dose in the planning target volume. *Phys Med Biol* (1994) 39(11):2091–8. doi: 10.1088/0031-9155/39/11/018

2. Ebert MA, Hoban PW. Some characteristics of tumour control probability for heterogeneous tumours. *Phys Med Biol* (1996) 41(10):2125–33. doi: 10.1088/0031-9155/41/10/019

3. van der Heide UA, Houweling AC, Groenendaal G, Beets-Tan RGH, Lambin P. Functional MRI for radiotherapy dose painting. *Quant Imaging Cancer.* (2012) 30(9):1216–23. doi: 10.1016/j.mri.2012.04.010

4. Bentzen SM, Gregoire V. Molecular imaging-based dose painting: a novel paradigm for radiation therapy prescription. *Semin Radiat Oncol* (2011) 21 (2):101–10. doi: 10.1016/j.semradonc.2010.10.001

5. Vogelius IR, Håkansson K, Due AK, Aznar MC, Berthelsen AK, Kristensen CA, et al. Failure-probability driven dose painting. *Med Phys* (2013) 40(8):081717. doi: 10.1118/1.4816308

6. Grönlund E, Johansson S, Montelius A, Ahnesjö A. Dose painting by numbers based on retrospectively determined recurrence probabilities. *Radiother Oncol J Eur Soc Ther Radiol Oncol* (2017) 122(2):236–41. doi: 10.1016/j.radonc.2016.09.007

7. Grönlund E, Johansson S, Nyholm T, Thellenberg C, Ahnesjö A. Dose painting of prostate cancer based on Gleason score correlations with apparent diffusion coefficients. *Acta Oncol Stockh Swed.* (2018) 57(5):574–81. doi: 10.1080/0284186X.2017.1415457

8. Turkbey B, Shah VP, Pang Y, Bernardo M, Xu S, Kruecker J, et al. Is apparent diffusion coefficient associated with clinical risk scores for prostate cancers that are visible on 3-T MR images? *Radiology* (2011) 258(S2):488–95. doi: 10.1148/radiol.10100667

9. Johansson S, Aström L, Sandin F, Isacsson U, Montelius A, Turesson I. Hypofractionated proton boost combined with external beam radiotherapy for treatment of localized prostate cancer. *Prostate Cancer.* (2012) 2012:654861. doi: 10.1155/2012/654861

10. Johansson S, Isacsson U, Sandin F, Turesson I. High efficacy of hypofractionated proton therapy with 4 fractions of 5 Gy as a boost to 50 Gy photon therapy for localized prostate cancer. *Radiother Oncol* (2019) 141:164–73. doi: 10.1016/j.radonc.2019.06.036

11. Grönlund E, Almhagen E, Johansson S, Traneus E, Nyholm T, Thellenberg C, et al. Robust treatment planning of dose painting for prostate cancer based on ADC-to-Gleason score mappings - what is the potential to increase the tumor control probability? *Acta Oncol Stockh Swed* (2020) 60:1–8. doi: 10.1080/0284186X.2020.1817547

12. Lagendijk JJ, Raaymakers BW, Van Vulpen M. The magnetic resonance imaging–linac system. *In: Semin Radiat Oncol Elsevier;* (2014) 24:207–9. doi: 10.1016/j.semradonc.2014.02.009

13. Winkel D, Bol GH, Kroon PS, van Asselen B, Hackett SS, Werensteijn-Honingh AM, et al. Adaptive radiotherapy: The elekta unity MR-linac concept. *Clin Transl Radiat Oncol* (2019) 18:54–9. doi: 10.1016/j.ctro.2019.04.001

14. Syed YA, Patel-Yadav AK, Rivers C, Singh AK. Stereotactic radiotherapy for prostate cancer: A review and future directions. *World J Clin Oncol* (2017) 8 (5):389–97. doi: 10.5306/wjco.v8.i5.389

15. Duprez F, De Neve W, De Gersem W, Coghe M, Madani I. Adaptive dose painting by numbers for head-and-Neck cancer. *Int J Radiat Oncol* (2011) 80 (4):1045–55. doi: 10.1016/j.ijrobp.2010.03.028

16. Malkov VN, Rogers DWO. Charged particle transport in magnetic fields in EGSnrc. *Med Phys* (2016) 43(7):4447–58. doi: 10.1118/1.4954318

17. Richmond N, Angerud A, Tamm F, Allen V. Comparison of the RayStation photon Monte Carlo dose calculation algorithm against measured data under homogeneous and heterogeneous irradiation geometries. *Phys Med* (2021) 82:87–99. doi: 10.1016/j.ejmp.2021.02.002

18. McPartlin AJ, Li XA, Kershaw LE, Heide U, Kerkmeijer L, Lawton C, et al. MRI-Guided prostate adaptive radiotherapy – a systematic review. *Radiother Oncol* (2016) 119(3):371–80. doi: 10.1016/j.radonc.2016.04.014

19. Castro P, Roch M, Zapatero A, Büchser D, Garayoa J, Ansón C, et al. Multicomponent assessment of the geometrical uncertainty and consequent margins in prostate cancer radiotherapy treatment using fiducial markers. *Int J Med Phys Clin Eng Radiat Oncol* (2018) 7(4):503–21. doi: 10.4236/ijmpcero.2018.74043

20. Ghadjar P, Fiorino C, Munck Af Rosenschöld P, Pinkawa M, Zilli T, van der Heide UA. ESTRO ACROP consensus guideline on the use of image guided radiation therapy for localized prostate cancer. *Radiother Oncol J Eur Soc Ther Radiol Oncol* (2019) 141:5–13. doi: 10.1016/j.radonc.2019.08.027

21. van Herk M. Errors and margins in radiotherapy. *Semin Radiat Oncol* (2004) 14(1):52–64. doi: 10.1053/j.semradonc.2003.10.003

22. van Herk M, Witte M, van der Geer J, Schneider C, Lebesque JV. Biologic and physical fractionation effects of random geometric errors. *Int J Radiat Oncol* (2003) 57(5):1460–71. doi: 10.1016/j.ijrobp.2003.08.026

23. Gordon JJ, Siebers JV. Convolution method and CTV-to-PTV margins for finite fractions and small systematic errors. *Phys Med Biol* (2007) 52(7):1967–90. doi: 10.1088/0031-9155/52/7/013

24. Fredriksson A, Engwall E, Andersson B. Robust radiation therapy optimization using simulated treatment courses for handling deformable organ motion. *Phys Med Biol* (2021) 66(4):045010. doi: 10.1088/1361-6560/abd591

25. Abdel-Wahab M, Begum N, Kovacs G, Lukka H, Miralbell R, Pellizzon A, et al. *Strategies for the management of localized prostate cancer: A guide for radiation oncologists [Internet].* Vienna: INTERNATIONAL ATOMIC ENERGY AGENCY (2014). Available at: https://www.iaea.org/publications/10618/strategies-

for-the-management-of-localized-prostate-cancer-a-guide-for-radiation-oncologists.

26. Weistrand O, Svensson S. The ANACONDA algorithm for deformable image registration in radiotherapy. *Med Phys* (2015) 42(1):40–53. doi: 10.1118/1.4894702

27. Tilly D, Tilly N, Ahnesjö A. Dose mapping sensitivity to deformable registration uncertainties in fractionated radiotherapy – applied to prostate proton treatments. *BMC Med Phys* (2013) 13(1):2. doi: 10.1186/1756-6649-13-2

28. Dice LR. Measures of the amount of ecologic association between species. *Ecology* (1945) 26(3):297–302. doi: 10.2307/1932409

29. Hausdorff F. Dimension und äußeres maß. *Math Ann* (1918) 79(1):157–79. doi: 10.1007/BF01457179

30. Widmark A, Gunnlaugsson A, Beckman L, Thellenberg-Karlsson C, Hoyer M, Lagerlund M, et al. Ultra-hypofractionated versus conventionally fractionated radiotherapy for prostate cancer: 5-year outcomes of the HYPO-RT-PC randomised, non-inferiority, phase 3 trial. *Lancet Lond Engl* (2019) 394(10196):385–95. doi: 10.1016/S1470-2045(20)30581-7

31. Afkhami Ardekani M, Ghaffari H, Navaser M, Zoljalali Moghaddam SH, Refahi S. Effectiveness of rectal displacement devices in managing prostate motion: a systematic review. *Strahlenther Onkol Organ Dtsch Rontgengesellschaft Al.* (2021) 197(2):97–115. doi: 10.1007/s00066-020-01633-9

32. Wang K, Mavroidis P, Royce TJ, Falchook AD, Collins SP, Sapareto S, et al. Prostate stereotactic body radiation therapy: An overview of toxicity and dose response. *Int J Radiat Oncol Biol Phys* (2021) 110(1):237–48. doi: 10.1016/j.ijrobp.2020.09.054

33. Grossfeld GD, Latini DM, Lubeck DP, Mehta SS, Carroll PR. Predicting recurrence after radical prostatectomy for patients with high risk prostate cancer. *J Urol.* (2003) 169(1):157–63. doi: 10.1097/01.ju.0000036470.57520.a0

# Appendix A

## Grönlund's TCP formalism

We applied the failure-driven TCP model for PCa derived by Grönlund et al. (7, 11). They defined TCP in terms of recurrence-free survival at five years using GS as a single variable for dose-response differentiation. Clinical endpoint and GS data from a patient cohort treated with photons (2 Gy x 25) and protons (5 Gy x 4) were used to derive TCP model parameter values. The total dose was converted to EQD2 assuming an alpha/beta ratio of 1.93 Gy. GS is related to the risk of recurrence (33) and is determined through histological assessment of prostate biopsies sampled prior to treatment as per routine clinical practice. GS is used as a prognostic tool together with the TNM system and prostate-specific antigen (PSA) concentration in blood to classify the cancer into low-, intermediate, or high-risk PCa. For a full derivation of the model, see Grönlund's previous work (7). In brief, the total control probability is given by a product of voxel control probabilities according to

$$\mathrm{TCP} = \prod\nolimits_j \mathrm{VCP}_j, j \in \mathrm{CTV} \qquad ,$$

i. e. it is assumed that potential bystander effects can be neglected. To establish a direct link between image intensities and TCP they used published correlations between ADC and GS to construct a mapping from ADC to Gleason probabilities according to

$$p\left(G_k|ADC\right) = \frac{p(ADC|G_k)}{\sum_m p(ADC|G_m)}$$

here $_p(ADC|G_k)$ are lognormal distributions of ADC values for each Gleason score category.$G_k$ The 25th and 75th percentiles of the lognormal distributions were set to correspond to the uncertainty ranges found in the work by Turkbey et al. (8). The voxel control probability $TCP_j$ of the j:th voxel is given by a geometric average

$$\mathrm{VCP}_j\left(d_j, \mathrm{ADC}\right) = \left(\prod_k R(d_j, G_k)^{p(\mathrm{G}_k|\mathrm{ADC}))^{\nu_k}}\right.$$

ver a set of dose-response functions $R(d_j, G_k)$defined as

$$R\left(d_j, G_k\right) = 1/(1 + \frac{D_{50}\left(G_k\right)}{d_j})^{4\gamma_{50}}$$

,here $d_j$ is the dose delivered to voxel $j$, $v_j$ is the fractional volume of the voxel within the prostate contour, $D_{50}(G_k)$ is the dose level corresponding to 50% control probability, and $\gamma_{50}$ is the normalized dose-response gradient (assumed to be independent of GS) evaluated at $D_{50}$. These parameter values were derived based on the assumption of a linear relationship between GS and voxel control probability at the homogeneous dose level prescribed to the patient cohort under study (EQD2 = 91.6 $\mathrm{Gy}_{1.93}$). Normal tissue voxels were treated as Gleason score 6.

In contrast to previous studies by Grönlund et al. (7, 11) we included the seminal vesicles in the TCP model to be able to include the effects of potential loss of target coverage. In previous studies, the seminal vesicles were simply assumed to be controlled by the dose level they received (i.e. $\mathrm{TCP}_{\mathrm{ves}}$=100%). In the present work, we treated the full SV volume as GS 6 by assigning to each SV voxel the highest possible ADC value in the ADC-to-Gleason mapping range.

TYPE Original Research
PUBLISHED 10 November 2022
DOI 10.3389/fonc.2022.1024160

# A novel approach for eliminating metal artifacts based on MVCBCT and CycleGAN

Zheng Cao[1,2], Xiang Gao[2], Yankui Chang[3], Gongfa Liu[1*] and Yuanji Pei[1*]

[1]National Synchrotron Radiation Laboratory, University of Science and Technology of China, Hefei, China, [2]Hematology and Oncology Department, Hefei First People's Hospital, Hefei, China, [3]School of Nuclear Science and Technology, University of Science and Technology of China, Hefei, China

Check for updates

OPEN ACCESS

EDITED BY
Wei Zhao,
Beihang University, China

REVIEWED BY
Chenbin Liu,
Chinese Academy of Medical Sciences and Peking Union Medical College, China
Yangkang Jiang,
Zhejiang Cancer Hospital, China

*CORRESPONDENCE
Gongfa Liu
gfliu@ustc.edu.cn
Yuanji Pei
yjpei@ustc.edu.cn

SPECIALTY SECTION
This article was submitted to Radiation Oncology, a section of the journal Frontiers in Oncology

RECEIVED 21 August 2022
ACCEPTED 27 October 2022
PUBLISHED 10 November 2022

CITATION
Cao Z, Gao X, Chang Y, Liu G and Pei Y (2022) A novel approach for eliminating metal artifacts based on MVCBCT and CycleGAN. *Front. Oncol.* 12:1024160. doi: 10.3389/fonc.2022.1024160

**Purpose:** To develop a metal artifact reduction (MAR) algorithm and eliminate the adverse effects of metal artifacts on imaging diagnosis and radiotherapy dose calculations.

**Methods:** Cycle-consistent adversarial network (CycleGAN) was used to generate synthetic CT (sCT) images from megavoltage cone beam CT (MVCBCT) images. In this study, there were 140 head cases with paired CT and MVCBCT images, from which 97 metal-free cases were used for training. Based on the trained model, metal-free sCT (sCT_MF) images and metal-containing sCT (sCT_M) images were generated from the MVCBCT images of 29 metal-free cases and 14 metal cases, respectively. Then, the sCT_MF and sCT_M images were quantitatively evaluated for imaging and dosimetry accuracy.

**Results:** The structural similarity (SSIM) index of the sCT_MF and metal-free CT (CT_MF) images were 0.9484, and the peak signal-to-noise ratio (PSNR) was 31.4 dB. Compared with the CT images, the sCT_MF images had similar relative electron density (RED) and dose distribution, and their gamma pass rate (1 mm/1%) reached 97.99% ± 1.14%. The sCT_M images had high tissue resolution with no metal artifacts, and the RED distribution accuracy in the range of 1.003 to 1.056 was improved significantly. The RED and dose corrections were most significant for the planning target volume (PTV), mandible and oral cavity. The maximum correction of Dmean and D50 for the oral cavity reached 90 cGy.

**Conclusions:** Accurate sCT_M images were generated from MVCBCT images based on CycleGAN, which eliminated the metal artifacts in clinical images completely and corrected the RED and dose distributions accurately for clinical application.

KEYWORDS

**tomography, MVCBCT, sCT CycleGAN, metal artifact reduction, radiotherapy dosimetry**

## Introduction

Metal artifacts are a common problem in kilovoltage CT images and radiation therapy. In the process of CT scanning, when X-rays pass through metal implants, such as metal dentures and metal hip joints in patients, erroneous X-ray projections will be produced due to the combined effects of beam hardening, scattering, photon starvation, noise enhancement, volume effects and other factors (1, 2), resulting in bright and dark stripes and radial areas in the reconstructed images; these are known as metal artifacts. Metal artifacts not only affect the diagnosis and the accurate delineations of the tumour target volume and normal tissues but also introduce dose calculation errors in radiation therapy by reducing the accuracy of relative electron densities (RED), which endanger the efficacy and safety of radiotherapy for patients (3–5).

Traditional metal artifact reduction (MAR) algorithms mainly include the interpolation method and iterative method (6–8), which often introduce new artifacts into images, resulting in image distortion (9–12). In recent years, deep learning technology has developed rapidly and has been widely applied in the field of image processing; it has provided new ideas for MAR in CT images. Yu et al. combined the traditional MAR method with a convolutional neural network (CNN) and achieved a higher accuracy than the traditional MAR method (13). Zhang et al. corrected metal artifacts in cervical CT images by using a CNN-based method (14). Zhu et al. trained U-Net based on a digital anthropomorphic head phantom and verified its MAR effect through PMMA phantoms containing aluminium rods and copper rods (15). Wang et al. developed an interpretable network model named InDuDoNet by combining sinogram and image data and embedding imaging geometric constraints in training (16). Yu et al. also designed a new deep learning framework by combining the advantages of the sinogram and image learning to obtain MAR images through multiple filtered back-projection reconstruction of the sinogram (17).

All the above studies are supervised methods that require paired CT images with the same anatomical structure, one with and the other without metal artifacts, for model training. However, it is clinically impractical to obtain such pairs of images. To obtain paired data, some studies used simulated phantoms for model training (15), and other studies artificially generated metal artifacts on metal-free CT images through theoretical calculations (13, 14, 16, 17). A simulated phantom is very different from the real human body, and the artificially generated metal artifacts cannot accurately simulate the real physical mechanisms of CT imaging. Therefore, the above two methods have poor generalization ability to real patient data (13–17). To solve the problem of the lack of paired training data, Liao et al. proposed an unsupervised network model named ADN, which used unpaired data for training (18), and its generalization ability was significantly improved compared to the supervised models that used synthetic data. Nevertheless, metal artifacts on CT images of real patients are still clearly residual and cannot be completely eliminated

This study aims to completely eliminate metal artifacts in CT images based on paired data from real patients. Compared with CT images, MVCBCT images have higher noise and lower soft tissue resolution, but the higher X-ray energy greatly reduces the photon starvation and radiation hardening effects, making the metal artifacts almost negligible, and this feature can be applied to MAR in CT images (19–21). In this work, we proposed a novel MAR approach using paired MVCBCT images and planning CT images. First, paired metal-free MVCBCT (MV_MF) images and metal-free planning CT (CT_MF) images were used for training the cycle-consistent adversarial network (CycleGAN) model. Then, synthetic metal-free CT (sCT_MF) images were generated from MV_MF images in the test dataset and compared with CT_MF images in terms of image quality, the RED distributions of organs at risk (OARs) and the dose calculation in radiation therapy. Finally, metal cases were used to evaluate the effect of MAR. The synthetic metal-containing CT (sCT_M) images were generated from the metal-containing MVCBCT (MV_M) images and compared with metal-containing CT (CT_M) images. The comparison of sCT_M and CT_M images was implemented with imaging and dosimetry to evaluate the radiation dosimetry improvement in the generated sCT_M images.

## Materials and methods

As illustrated in Figure 1, the process of this research was mainly divided into four stages. First, the CT images and MVCBCT images of the same patient were elastically registered in the registration stage. For metal-free images in the training set, CT numbers range from -1000 HU to 3000 HU for CT and from -1000 HU to 1400 HU for MV. Next, the CycleGAN model was trained using the metal-free images to generate sCT_MF images from MV_MF images. Then, in the third stage, the accuracy of the generated sCT_MF images was evaluated with imaging and dosimetry to judge whether the sCT images generated by the model were accurate enough to perform MAR. Finally, in the MAR stage, based on the well-trained CycleGAN model, metal-artifacts-free sCT_M images were generated from MV_M images; then, the metal pixels in the CT_M images were copied to the corresponding pixel positions in the sCT_M images. Specifically, the CT numbers in the MV_M images exceeding 1400 HU were modified to 1400 HU, and the sCT_M images without added metal pixels were generated through the CycleGAN model. In the works of Liao et al. and Wang et al., 2500 HU was used as the threshold of metal segmentation in CT images (16, 18). However, bright

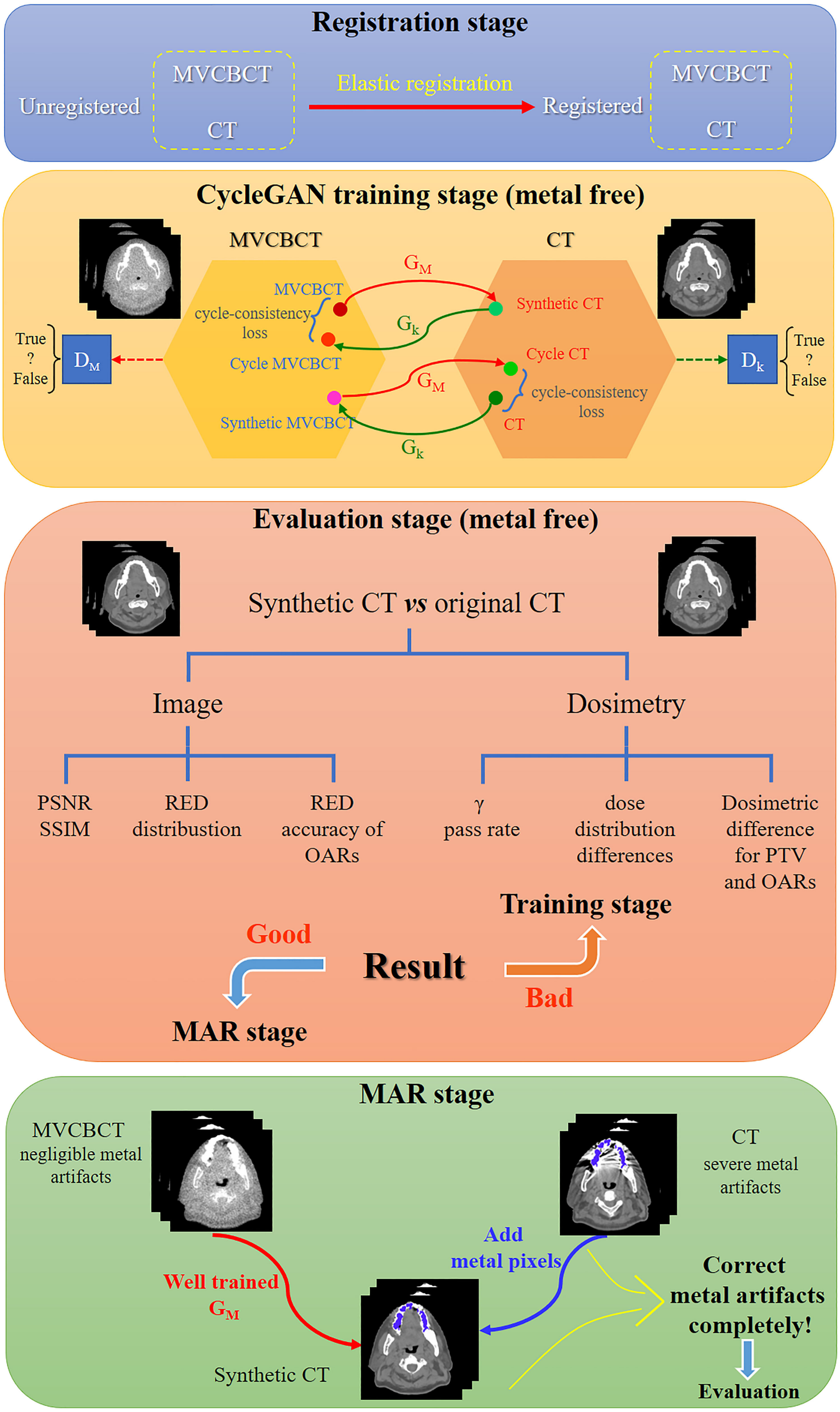


FIGURE 1
Schematic diagram of metal artifact correction based on MVCBCT and CycleGAN. The process of this research was divided into the registration stage, CycleGAN training stage, evaluation stage and MAR stage.

metal artifacts may still exist in the metal region segmented by this method. We observed that there is almost no metal artifact in the MVCBCT images and the CT number of metal is not less than 300 HU. Therefore, in order to reduce the metal artifacts contained in the segmented metal regions as much as possible, we identified the intersection regions with HU values greater than 2500 in the CT_M images and greater than 300 in the MV_M images as the metal regions. The final MAR images were obtained by copying the CT numbers of the metal pixels in the CT_M images into the previously generated sCT_M images.

## Data acquisition and preprocessing

Metal dentures have diverse materials and complex shapes. When their size is large or RED is high, severe metal artifacts appear in CT images, destroying the image quality and the accuracy of RED information. Therefore, the correction of metal artifacts caused by metal dentures has good clinical application value. In this study, CT and MVCBCT images of head cancer patients were obtained from the dataset.

Paired planning CT images and MVCBCT images of 126 patients without metal dentures and 14 patients with metal dentures were collected in this study, and the scans included the head. The CT images were derived from a Siemens SOMATOM Spirit helical CT scanner (tube voltage of 130 kV, slice thickness of 3 mm, 16-bit image output). The paired MVCBCT images were obtained in the first fraction (Siemens Artiste Medical Electron Linear Accelerator, 6 MV, 0.54 mm×0.54 mm×0.54 mm). The images from the temporomandibular joint to the mandible were selected for training and evaluation. The images from ninety-seven patients without metal in their scans were randomly selected for model training; these data included 1762 paired planning CT slices and MVCBCT slices. The remaining images of the 29 metal-free patients (457 slices) were used as the metal-free test set, and the 14 patients with metal dentures (86 slices) were used as the metal test set.

Data preprocessing was required before model input. First, the Elastix multiresolution B-spline registration method (22, 23) was used to elastically align the CT images and MVCBCT images of the same patient. Then, the images were resampled to 1 mm×1 mm and cropped to 256×256 pixels. Then, the hyperbolic tangent function (Tanh) was used to scale the CT values to (–1,1), and is defined as $\mathrm{Tanh(x)} = \frac{e^x - e^{-x}}{e^x + e^{-x}}$. Before being processed by Tanh, the HU values of the CT images and MVCBCT images were scaled linearly with three methods as follows:

1) $X(MVCBCT) = Tanh(HU(MVCBCT)/400)$ and $X(CT) = Tanh(HU(CT)/400)$.

2) $X(MVCBCT) = Tanh(HU(MVCBCT)/150)$ and $X(CT) = Tanh(HU(CT)/300)$.

3) $X(MVCBCT) = Tanh(HU(MVCBCT) - 800/400)$ and $X(CT) = Tanh(HU(CT) - 1600/960)$.

Processed by the above three methods, these data were used for model training separately to obtain three groups of results, named P1, P2 and P3.

## CycleGAN-based unsupervised model

Although the paired CT and MVCBCT images were selected as training data, there were still problems in supervised pixel-to-pixel learning. The setup error between the two scans, the differences in the mouth opening size and image distortions caused by elastic registration may introduce differences into the CT images and MVCBCT images. Therefore, this study used CycleGAN for unsupervised learning because pixel-level correspondence is not necessary.

Generative adversarial networks (GANs) are unsupervised deep learning models that mainly include a generator (G) and a discriminator (D). A trained GA-B could generate image A', which has the structure of image A and the style of image B. CycleGAN models (24) include two generators and two discriminators and add cycle-consistency loss for training. CycleGAN has been widely used for interconversion between different types of medical images (25–30). The structure of CycleGAN used in this study is consistent with that reported in the literature (24), and the model structure is shown in Figure 1. ResUNet (31) was used as the generator, and the Adam optimizer was selected to train the model with a batch size of 6 on one NVIDIA Quadro RTX 6000 GPU. The learning rate was constant at 0.0002 for the first 100 epochs of training and attenuated by 1% per epoch for the last 100 epochs. A previous study showed that paired data have better performance than unpaired data when using CycleGAN to generate sCT (32). Therefore, this study used deformation-registered paired data for training.

## Imaging evaluation

Compared to the planning CT images, the image quality of synthetic CT images was evaluated by the peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) index.

$$PSNR(I_1, I_2) = 10 \times \log_{10}\left(\frac{MAX^2}{RMSE(I_1, I_2)^2}\right) \qquad (1)$$

$$SSIM(I_1, I_2) = \frac{\left(2\mu_{I_1}\mu_{I_2} + c_1\right)\left(2\sigma_{I_1,I_2} + c_2\right)}{\left(\mu_{I_1}^2 + \mu_{I_2}^2 + c_1\right)\left(\sigma_{I_1}^2 + \sigma_{I_2}^2 + c_2\right)} \qquad (2)$$

To compute the dose using CT images in photon radiotherapy, the CT numbers need to be converted to RED values through the CT-ED conversion curve. Since the CT-ED curves are very different between CT images and MVCBCT images, it is necessary to compare their RED values rather than their CT numbers. The CIRS 062 electron density phantom was used to obtain CT numbers corresponding to the RED values in the range of 0 to 1.456. The correspondence between CT numbers and RED values of different metals was obtained through the head part of a CIRS ATOM 701-B dosimetry anthropomorphic phantom with aluminium alloy (RED: 2.43), titanium alloy (RED: 3.73) and stainless steel (RED: 6.83) plugs.

In addition, the RED distributions of OARs in CT images were analysed. The main OARs affected by metal artifacts, such as the mandible, oral cavity, parotid gland and spinal cord, were delineated, and their RED distributions were compared with those in the MVCBCT images and sCT images.

## Dosimetry evaluation

The target volume was redelineated according to the anatomical structure of each patient in the test set with reference to the actual target volume position of NPC patients. In the treatment planning system (TPS), the same prescription dose (PTV: 6000 cGy) was used to produce a dynamic intensity-modulated plan (Eclipse 15.6, AXB algorithm) on the CT images, and then the plan was copied to the corresponding sCT images. Finally, the global gamma pass rates and the three-dimensional dose distribution difference of the target area and the OARs were compared. The gamma pass rates between the radiotherapy plans of sCT images and CT images were calculated using PTW Verisoft software, version 6.0 (PTW, Frieburg, Germany), and the criteria included 2 mm/2% and 1 mm/1% (distance error/dose error), respectively. V95%, V100%, V110% (Vx% means the percentage of volume receiving at least x% of the prescription dose), D5, D95 (Dx means the doses to x % of the volume), Dmean (mean dose of the volume) for the PTV, D2 and Dmean for the mandible, D50 and Dmean for the oral cavity and parotid gland, and D0.1 cc (dose to 0.1 cc volume) for the spinal cord were investigated.

The significance test of the RED and dosimetry data was performed using IBM SPSS Statistics 26 software. Paired and unpaired t tests were used for normally distributed data, and the Mann-Whitney U test was used for nonnormally distributed unpaired data (33).

# Results

Figure 2 shows the effects of different preprocessing methods (P1, P2 and P3) on sCT image quality. For organs such as the mandible and teeth, more uniform CT numbers and higher similarity with the CT_MF images were achieved using sCT_MF_P3 compared with sCT_MF_P1 and sCT_MF_P2. The CT numbers of teeth for CT_MF, sCT_MF_P1, sCT_MF_P2 and sCT_MF_P3 were 1473 ± 554 HU, 1726 ± 863 HU, 2003 ± 995 HU and 1476 ± 481 HU, respectively. The CT numbers of the mandible were 838 ± 494 HU, 891 ± 631 HU, 917 ± 417 HU and 848 ± 425 HU, respectively. Table 1 shows a comparison of the accuracy of sCT_MF images with different preprocessing methods in the ranges of [-200, 400] HU, [400, 800] HU, [800, 3000] HU and [-1000, 3000] HU. sCT_MF_P2 performed best at [-200, 400] HU, sCT_MF_P1 performed best at [400, 800] HU, and sCT_MF_P3 performed best at [800, 3000] HU. Obviously, different image preprocessing methods have their own advantages in different CT number ranges. Therefore, the three trained models with different preprocessing methods were combined to produce the new sCT_MF (sCT_MF_P4), which used the part of sCT_MF_P2 below 400 HU, the part of sCT_MF_P1 at [400, 800] HU and the part of sCT_MF_P3 over

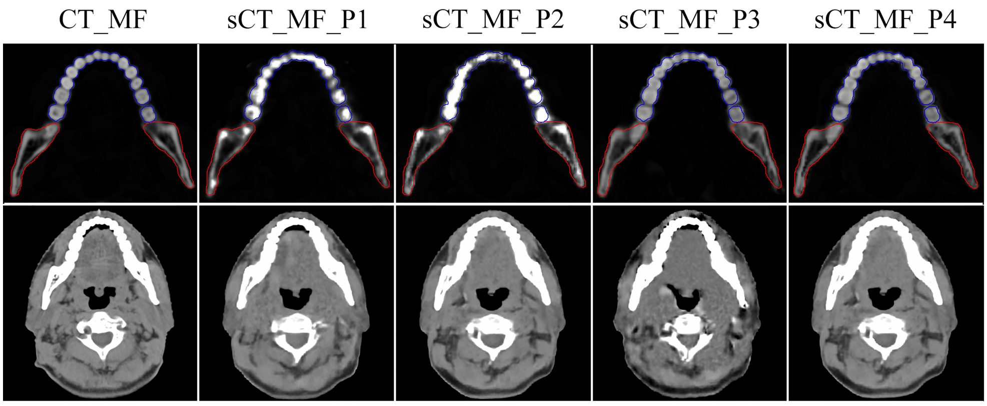


FIGURE 2
Visualized differences of CT_MF and sCT_MF images with different preprocessing methods. The display windows for the first and second rows were [0, 3000] HU and [-360, 440] HU, respectively. Blue lines represent the contour of the teeth, and red lines represent the contour of the mandible. The images in the first to fifth columns were CT_MF, sCT_MF_P1, sCT_MF_P2, sCT_MF_P3 and sCT_MF_P4, respectively. Different image preprocessing methods have their own advantages in different CT number ranges, and sCT_MF_P4 performs best.

TABLE 1 The evaluation of sCT_MF images with different preprocessing methods (PSNR (dB)/SSIM).

| CT number range(HU) | sCT_MF_P1 | sCT_MF_P2 | sCT_MF_P3 | sCT_MF_P4 |
|---|---|---|---|---|
| -1000~3000 | 30.0/0.9459 | 28.2/0.9435 | 30.7/0.9345 | **31.4/0.9484** |
| -200~400 | 23.9/0.8599 | **24.2/0.8689** | 22.6/0.8395 | / |
| 400~800 | **22.8/0.9558** | 22.2/0.9534 | 21.4/0.9470 | / |
| 800~3000 | 27.9/0.9652 | 24.9/0.9555 | **33.0/0.9701** | / |

sCT_MF_P1, metal-free sCT images obtained by the prepossessing method named P1; sCT_MF_P2, metal-free sCT images obtained by the prepossessing method named P2; sCT_MF_P3, metal-free sCT images obtained by the prepossessing method named P3; metal-free sCT images obtained by the combined prepossessing method named P4. The best PSNR and SSIM values in different HU ranges of sCT_MF images are marked in bold.

800 HU. The accuracy of the sCT_MF_P4 image was improved significantly (PSNR: 31.4 ± 1.3 dB; SSIM: 0.9484 ± 0.0090). It should be noted that the generated sCT_MF and sCT_M images in the following were processed by the combined P4 method.

The RED comparison of CT, MVCBCT and sCT images is shown in Figure 3. In Figure 3A, the difference in the RED values of CT_MF and sCT_MF images was significantly smaller than that of CT_MF and MV_MF images, especially in soft tissues. Figure 3F and part A in Figure 3E show that the RED curves of CT_MF and sCT_MF images were almost coincident, while the curves of CT_MF and MV _MF images were quite different. The RED distributions of OARs for CT_MF and sCT_MF images were almost the same (Figure 4), and the difference was not statistically significant (P > 0.05 in Table 2). Compared with the large difference in the RED values of CT_MF and MV_MF images, the RED values of the main OARs in sCT_MF images were sufficiently accurate to be used for radiotherapy dose calculations.

The dose distributions based on CT_MF and sCT_MF images were slightly different, as shown in Figure 5A. The gamma pass rates of the sCT_MF-based plans were 99.72% ± 0.29% (2 mm/2%) and 97.99% ± 1.14% (1 mm/1%) compared to the CT_MF-based plans. The blue part in Figure 6 shows the absolute dose errors of CT_MF and sCT_MF images, which were 8.9 ± 6.2 cGy, 11.9 ± 8.1 cGy, 9.3 ± 7.2 cGy, 0.04% ± 0.06%, 0.32% ± 0.28% and 0.76% ± 0.77% for Dmean, D5, D95, V95, V100%, and V110% of the PTV, respectively. For the mandible (D2 and Dmean) and oral cavity (D50 and Dmean), the maximum differences were all less than 40 cGy, and the average difference was approximately 10 cGy. For the parotid (D50 and Dmean) and spinal cord (D0.1 cc), the max differences were all less than 20 cGy, and the average difference was approximately 7 cGy. The above results demonstrate that the dose distribution of sCT_MF images was consistent with that of CT_MF images, which proves the accuracy of our proposed method for generating synthetic CT images from MVCBCT images.

LI (34) and NMAR (35) are widely used approaches to MAR. Supplementary figure 1 and Supplementary figure 2 show the qualitative comparisons of our MAR method with the LI and NMAR methods on the clinical data and phantom data, respectively. It is clear that our method completely eliminates metal artifacts in both clinical data and phantom data, whereas both the LI and NMAR methods not only fail to completely eliminate metal artifacts, but also create a large number of new artifacts in the images. For MAR of CT_M images, metal artifacts with varying severities were completely removed from sCT_M images (Figures 3B–3D). The sCT_M images had comparable quality to metal-artifact-free CT_MF images. Notably, according to the RED differences in Figures 3B–3D, the RED information corrupted by metal artifacts was corrected in the sCT_M images, and the RED difference in the area away from the metal artifacts was very small. It was evident from Figure 3F and parts B-D in Figure 3E that the difference in the RED values of the CT_M and sCT_M images was larger than that of the CT_MF and sCT_MF images. In Figure 4, the RED values of the CT_M and sCT_M images were 1.350 ± 0.254 and 1.355 ± 0.230 for the mandible (P = 0.813), 1.060 ± 0.081 and 1.032 ± 0.016 for the oral cavity (P< 0.001), 0.987 ± 0.036 and 0.994 ± 0.027 for the parotid (P = 0.174), and 1.029 ± 0.015 and 1.020 ± 0.010 for the spinal cord (P = 0.006), respectively.

The dose distributions based on CT_M and sCT_M images are shown in Figures 5B–5D, and the gamma pass rates of the sCT_M-based plans were 99.55% ± 0.35% (2 mm/2%) and 96.55% ± 1.54% (1 mm/1%) compared to the CT_M-based plans. The green part in Figure 6 shows the absolute dose errors from the CT_M and sCT_M images, which were 22.1 ± 17.9 cGy, 28.1 ± 20.8 cGy, 19.3 ± 19.0 cGy, 0.05% ± 0.07%, 0.37% ± 0.46% and 1.37% ± 1.46% for Dmean, D5, D95, V95%, V100%, and V110% of the PTV, respectively. For the PTV (Dmean, D5), mandible (Dmean), oral cavity (D50 and Dmean) and parotid (D50 and Dmean), the absolute dose errors of the sCT_M and CT_M images were statistically significant compared to the absolute errors of the sCT_MF and CT_MF images (Figure 6C and Table 3). The dose difference in the spinal cord far away from metal artifacts was not statistically significant (6.7 ± 3.5 vs. 11.7 ± 8.1, P > 0.05 in Table 3).

## Discussion

In this study, a novel approach, in which the advantages of the CycleGAN model and the characteristics of negligible metal artifacts in MVCBCT images were integrated, was proposed to address the MAR task. The results suggested that our proposed

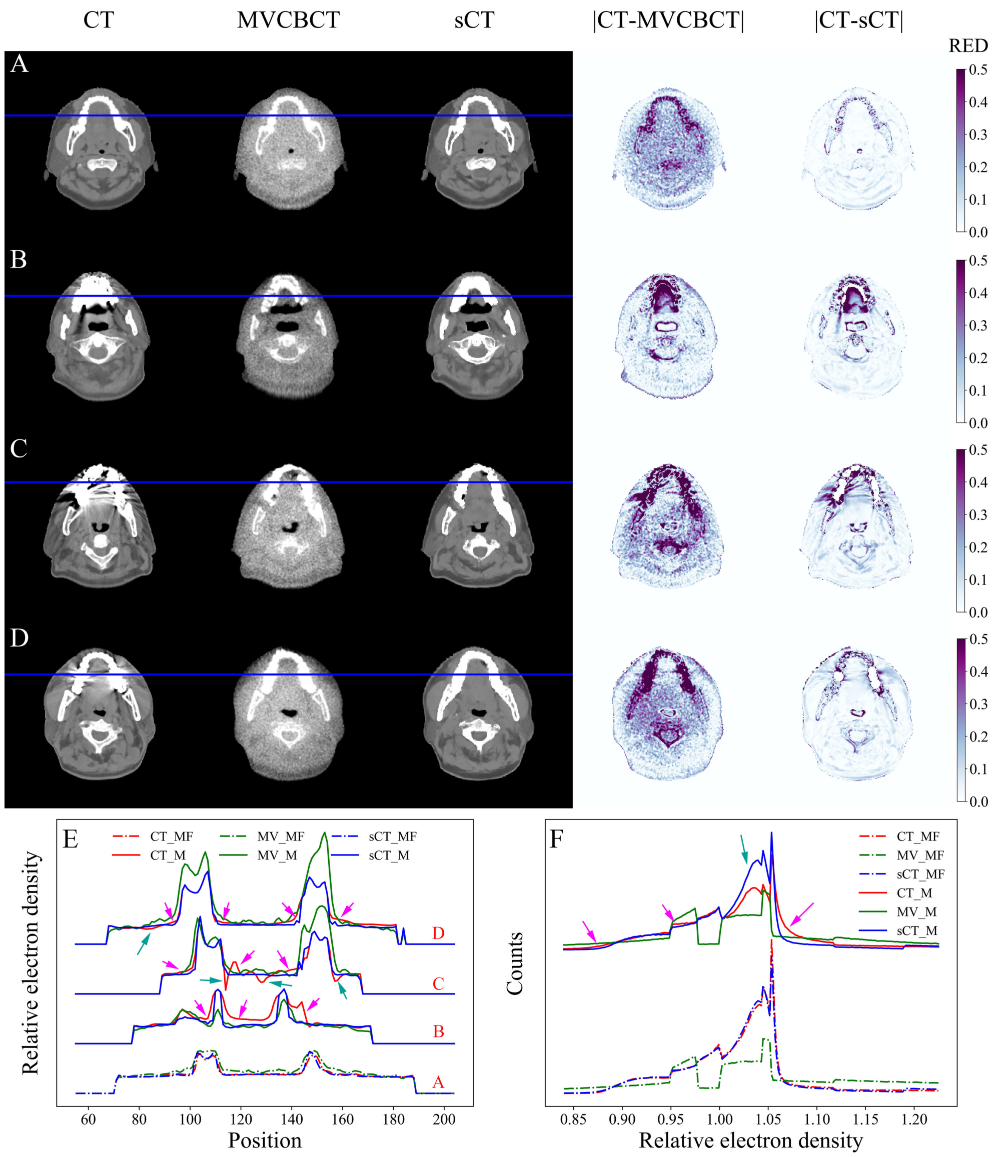


FIGURE 3
RED comparison of CT, MVCBCT and sCT images. **(A)** Metal-free images. **(B-D)** Metal-containing images. **(E)** RED distribution curves for the blue lines in **(A–D)**. **(F)** RED histograms of the images. The display window for CT, MVCBCT and sCT images was [-360, 440] HU. The RED distributions of CT_MF and sCT_MF images were almost coincident, and the metal artifacts were completely eliminated in sCT_M images after the MAR stage.

method could be used to completely remove metal artifacts in original CT images and correct the destroyed RED distributions, and hence a more accurate dose calculation for radiotherapy can be produced.

Different normalization methods in preprocessing could affect the accuracy of sCT images, as shown in Figure 2. The difference between the P1, P2 and P3 methods was mainly because the main range of the CT numbers involved in the training stage varied with the preprocessing methods. Therefore, the three trained models with different preprocessing methods were combined to produce the final sCT images.

TPS requires images to be calibrated for RED values before dose calculations are performed (36). Considering the large gap between the CT-ED curves of CT and MVCBCT images (19), it is not intuitive to directly compare the difference in CT numbers when evaluating image quality in the study by Zhao et al. (37). Therefore, our image evaluation approach mainly focused on the RED values.

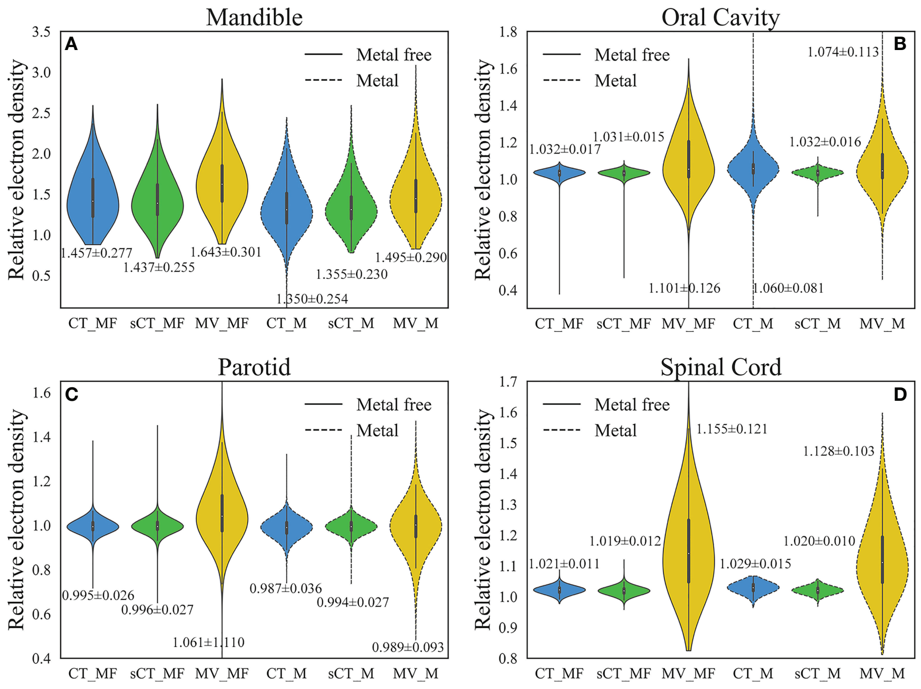


FIGURE 4

Comparison of RED distributions for OARs. **(A)** Mandible. **(B)** Oral cavity. **(C)** Parotid. **(D)** Spinal cord. The numbers marked in the figure are the average ± standard deviation. The RED values of the main OARs in sCT_MF images were accurate, and the inaccurate RED values caused by metal artifacts in CT_M images were corrected in sCT_M images after the MAR stage.

In the results, we analysed the image quality and dose calculation accuracy of the generated sCT images for the test sets with and without metal. Since we cannot obtain ground truth images for the clinical metal-containing images, in previous studies, quantitative evaluation could only be performed on synthetic data or simulated phantoms (13–18). In our study, we indirectly realized the quantitative evaluation of the MAR effect on clinical images through the quantitative evaluation of the sCT_MF images and the statistical analysis of sCT_MF and sCT_M images. The model was trained with paired CT_MF and MV_MF images, and CT_MF and sCT_MF images had high consistency in terms of image quality, RED values and dose distributions. The PSNR and SSIM values for the CT_MF and sCT_MF images comparison were 31.4 ± 1.3 and 0.9484 ± 0.0090, respectively, which are comparable to Liang et al.'s study (30.65 ± 1.36/0.85 ± 0.03), Vinas et al.'s study (29.7 ± 2.7/0.927 ± 0.028), Harms et al.'s study (PSNR: 32.3 ± 5.9) and Chen et al.'s study (30.75 ± 3.89/0.9642 ± 0.0186) for head patient images (25, 27, 38, 39). The gamma pass rates (1 mm/1%) of the sCT_MF-based plans (97.99% ± 1.14%) were better than those obtained in Liang et al.'s study (96.26% ± 3.59%) and Li et al.'s study (95.5% ± 1.6%) (25, 40). Therefore, we believe that the sCT_M images generated from MV_M images were sufficiently accurate to evaluate the effect of MAR.

In previous studies, the excellent MAR performance on simulated images could not be sustained on clinical images. Qualitative analyses showed that artifacts remained in images after MAR, and the image quality was also degraded (13, 14, 16–18). In contrast, metal artifacts in clinical images were eliminated completely in our study (Figure 3). Furthermore, quantitative

TABLE 2 P value comparison of RED distributions for OARs.

| OARs | CT_MF vs. sCT_MF[a] | CT_M vs .sCT_M[a] | CT_MF vs .CT_M[b] | CT_MF vs .sCT_M[b] |
|---|---|---|---|---|
| Mandible | 0.055 | 0.813 | < 0.001* | < 0.001* |
| Oral Cavity | 0.805 | < 0.001* | < 0.001* | 0.509 |
| Parotid | 0.499 | 0.174 | 0.355 | 0.744 |
| Spinal Cord | 0.056 | 0.006* | < 0.001* | 0.512 |

a: Paired-sample T test. b: Independent-sample T test.*: Statistically significant differences (P< 0.05). RED, Relative electron density; OARs, organs at risk; CT_MF, metal-free CT images; CT_M, metal-containing CT images; sCT_MF, metal-free sCT images; sCT_M, metal-containing sCT images.

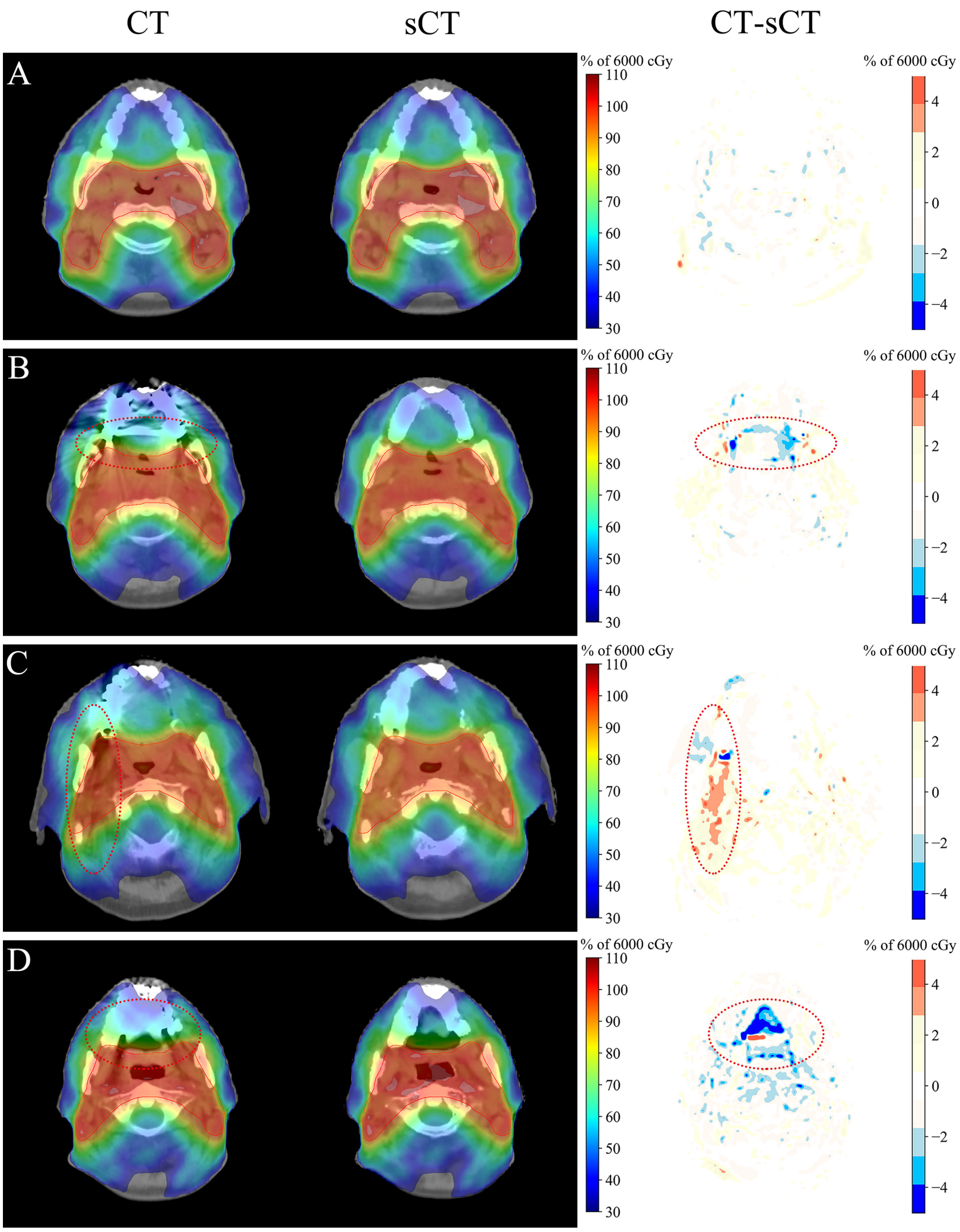


FIGURE 5
Dosimetric comparison of CT and sCT images. **(A)** Metal-free images. **(B–D)** Metal-containing images. The display window for the CT and sCT images was [-360, 440] HU. The dose distribution of sCT_MF images was consistent with that of CT_MF images, and there were obvious dose differences between CT_M and sCT_M images in the area with serious metal artifacts.

assessments of the MAR effect on clinical images were performed. The MV_M images were almost identical to the MV_MF images since the metal artifacts were barely visible (Figures 3B–3D). In the soft tissue region near the teeth, the RED distribution curves of CT_MF images were smooth, while those of CT_M images were not (Figure 3E). Some RED values were high (pink arrows) in CT_M images due to bright metal artifacts, while others (green arrows) were low due to dark metal artifacts. On the other hand, the curves in the corresponding areas in sCT_M images were as smooth as those in CT_MF images, which means that the RED values were accurately corrected in sCT_M images. In Figure 3F, the RED histograms of the CT_M and sCT_M images had obvious differences, especially in the RED value range of 1.003 to 1.056 (green arrow). This may be because metal artifacts mainly

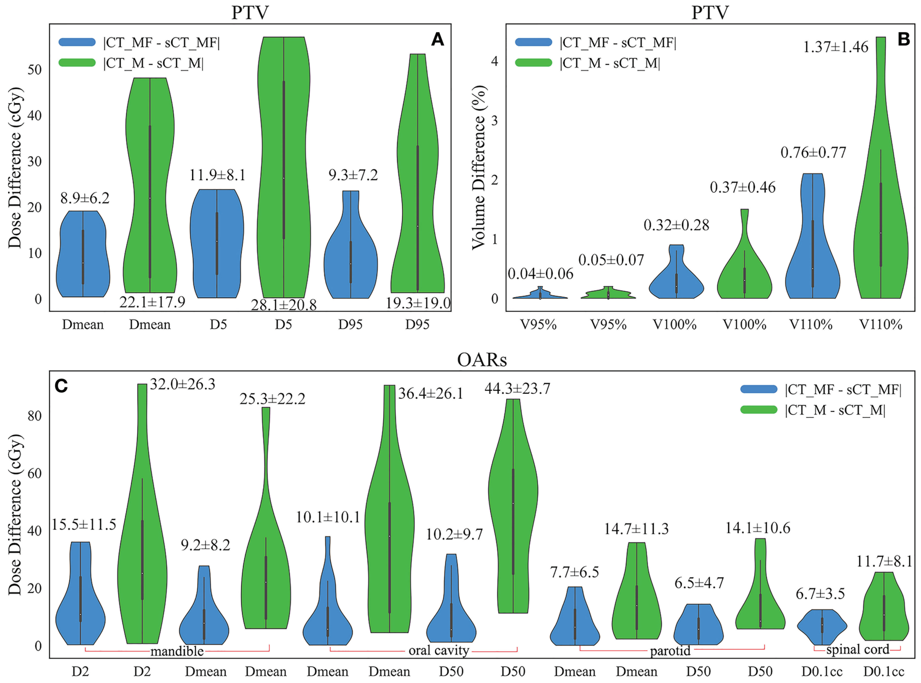


FIGURE 6
Comparison of the absolute dose errors for the PTV and OARs. **(A)** Dose difference (cGy) of the PTV. **(B)** Volume difference (%) of the PTV. **(C)** Dose difference (cGy) of the OARs. The numbers marked in the figure are the average ± standard deviation. Compared with metal-free cases, the average and standard deviation of the dose differences for the PTV and OARs doubled for cases with metal artifacts.

destroy RED values in the range of 1.003 to 1.056, while the damage is corrected in sCT_M images.

The RED distributions of OARs were different because of the different distances from the metal artifacts. Influenced by the metal artifacts, there were many pixels with low RED values in the mandible of CT_M images, and these pixels were corrected in the sCT_M images (Figure 4A). Due to the proximity to metal dentures, the RED values of the oral cavity in CT_M images were

TABLE 3 The difference significance test between the absolute dose errors of sCT_M and CT_M images and the absolute errors of sCT_MF and CT_MF images for the PTV and OARs.

| Structures | Dosimetry Parameter | Test Method | P Value |
|---|---|---|---|
| PTV | Dmean | T | 0.047* |
| | D5 | T | 0.038* |
| | D95 | T | 0.142 |
| | V95% | U | 0.487 |
| | V100% | U | 0.781 |
| | V110% | U | 0.517 |
| Mandible | D2 | U | 0.089 |
| | Dmean | U | 0.009* |
| Oral Cavity | Dmean | U | 0.002* |
| | D50 | U | < 0.001* |
| Parotid | Dmean | T | 0.047* |
| | D50 | U | 0.036* |
| Spinal Cord | D0.1 cc | T | 0.079 |

T: Independent-sample T test for normally distributed data. U: Independent-sample Mann-Whitney U Test for nonnormally distributed data. *: Statistically significant differences (P< 0.05). PTV, planning target volume; OARs, organs at risk; CT_MF, metal-free CT images; CT_M, metal-containing CT images; sCT_MF, metal-free sCT images; sCT_M, metal-containing sCT images; Vx%, the percentage of volume receiving at least x% of the prescription dose; Dx, the doses to x% of the volume; Dmean, mean dose of the volume; D0.1 cc, dose to 0.1 cc volume.

greatly affected by the metal artifacts (CT_MF vs. CT_M, P< 0.001 in Table 2), with a significantly higher mean and standard deviation (Figure 4B) and a large number of outliers that were too high or too low. The RED values of the oral cavity were almost consistent in the sCT_M and CT_MF images (P = 0.509 in Table 2), and there was a significant difference in the values of CT_M and sCT_M images (P< 0.001 in Table 2), which further proved the accuracy of RED correction for the oral cavity in sCT_M images. As shown in Table 2, the spinal cord and oral cavity had similar significance test results, which also proves that the RED values of the spinal cord were accurately corrected in sCT_M images.

For dose calculation, the gamma pass rates of sCT_M and CT_M images were lower than those of sCT_MF and CT_MF images, which was the results of MAR. As shown in Figures 5B–5D, there were obvious dose differences between CT_M and sCT_M images in the area with serious metal artifacts (elliptical dotted lines), and the maximum correction of the point dose could reach more than 5% of the total dose. Compared with metal-free cases, the average and standard deviation of the dose differences for the PTV and OARs doubled for cases with metal artifacts (Figure 6). The accuracy of the sCT_M-based dose calculation showed statistically significant improvements in the PTV and OARs (Table 3).

Finally, there are some works that need to be improved. The RED difference of the bone and tooth areas of the CT_MF and sCT_MF images was significantly greater than that of soft tissues. The results showed that the combination of multiple preprocessing methods could improve the accuracy of sCT images with high RED values, and this will be further researched in our next work.

## Conclusion

We proposed a novel MAR approach to complete the MAR task. In this approach, the advantages of the CycleGAN model and the characteristics of negligible metal artifacts in MVCBCT images are integrated. The model was trained on paired metal-free CT and MVCBCT images and generated metal-artifacts-free sCT images from metal-containing MVCBCT images to convert the task of MAR to the task of generating sCT images from MVCBCT images. The metal artifacts were completely removed in the sCT_M images, and the inaccurate RED values were corrected, which could significantly improve the accuracy of disease diagnosis and radiotherapy dose calculation.

## Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material. Further inquiries can be directed to the corresponding authors.

## Author contributions

ZC conceived the experiments. XG collected the clinical dataset. ZC, GL and YP designed the study and analyzed the result. ZC, XG, YC, GL and YP participated in writing manuscript. All authors contributed to the article and approved the submitted version.

## Funding

Hefei Municipal Health Commission (CN) Applied Medicine Research Project (key project: hwk2018zd012)

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fonc.2022.1024160/full#supplementary-material

## References

1. De Man B, Nuyts J, Dupont P, Marchal G, Suetens P. Metal streak artifacts in X-ray computed tomography: a simulation study. *IEEE Trans Nucl Sci* (1999) 46:691–6. doi: 10.1109/NSSMIC.1998.773898

2. Laukamp KR, Zopfs D, Lennartz S, Pennig L, Maintz D, Borggrefe J, et al. Metal artifacts in patients with large dental implants and bridges: combination of metal artifact reduction algorithms and virtual monoenergetic images provides an approach to handle even strongest artifacts. *Eur Radiol* (2019) 29:4228–38. doi: 10.1007/s00330-018-5928-7

3. Bazalova M, Beaulieu L, Palefsky S, Verhaegen F. Correction of CT artifacts and its influence on Monte Carlo dose calculations. *Med Phys* (2007) 34:2119–32. doi: 10.1118/1.2736777

4. Wei J, Sandison GA, Hsi WC, Ringor M, Lu X. Dosimetric impact of a CT metal artefact suppression algorithm for proton, electron and photon therapies. *Phys Med Biol* (2006) 51:5183–97. doi: 10.1088/0031-9155/51/20/007

5. Gao L, Li C, Lu Z, Xie K, Lin T, Sui J, et al. Comparison of different treatment planning approaches using VMAT for head and neck cancer patients with metallic dental fillings. *Radiat Med Prot* (2021) 2:128–33. doi: 10.1016/j.radmp.2021.05.002

6. Yu H, Zeng K, Bharkhada DK, Wang G, Madsen MT, Saba O, et al. A segmentation-based method for metal artifact reduction. *Acad Radiol* (2007) 14:495–504. doi: 10.1016/j.acra.2006.12.015

7. Fleischmann D, Boas FE. Computed tomography–old ideas and new technology. *Eur Radiol* (2011) 21:510–7. doi: 10.1007/s00330-011-2056-z

8. Gao L, Sui J, Lin T, Xie K, Ni X. Metal artifact reduction method based on noncoplanar scanning in CBCT imaging. *IEEE Access* (2020) 8:7236–43. doi: 10.1109/ACCESS.2019.2962386

9. Xia D, Roeske JC, Yu L, Pelizzari CA, Mundt AJ, Pan X. A hybrid approach to reducing computed tomography metal artifacts in intracavitary brachytherapy. *Brachytherapy* (2005) 4:18–23. doi: 10.1016/j.brachy.2004.11.001

10. Boas FE, Fleischmann D. Evaluation of two iterative techniques for reducing metal artifacts in computed tomography. *Radiology* (2011) 259:894–902. doi: 10.1148/radiol.11101782

11. Yazdi M, Lari MA, Bernier G, Beaulieu L. An opposite view data replacement approach for reducing artifacts due to metallic dental objects: Reducing artifacts due to metallic dental objects. *Med Phys* (2011) 38:2275–81. doi: 10.1118/1.3566016

12. Kalender WA, Watzke O. A pragmatic approach to metal artifact reduction in CT: merging of metal artifact reduced images. *Eur Radiol* (2004) 14:849–56. doi: 10.1007/s00330-004-2263-y

13. Zhang Y, Yu H. Convolutional neural network based metal artifact reduction in X-ray computed tomography. *IEEE Trans Med Imaging* (2018) 37:1370–81. doi: 10.1109/TMI.2018.2823083

14. Huang X, Wang J, Tang F, Zhong T, Zhang Y. Metal artifact reduction on cervical CT images by deep residual learning. *BioMed Eng Online* (2018) 17:175. doi: 10.1186/s12938-018-0609-y

15. Zhu L, Han Y, Li L, Xi X, Zhu M, Yan B. Metal artifact reduction for X-ray computed tomography using U-net in image domain. *IEEE Access* (2019) 7:98743–54. doi: 10.1109/ACCESS.2019.2930302

16. Wang H, Li Y, Zhang H, Chen J, Ma K, Meng D, et al. InDuDoNet: An interpretable dual domain network for CT metal artifact reduction. In: *Medical image computing and computer assisted intervention – MICCAI 2021*. Cham: Springer International Publishing (2021). p. pp 107–118. doi: 10.1007/978-3-030-87231-1_11

17. Yu L, Zhang Z, Li X, Ren H, Zhao W, Xing L. Metal artifact reduction in 2D CT images with self-supervised cross-domain learning. *Phys Med Biol* (2021) 66:175003. doi: 10.1088/1361-6560/ac195c

18. Liao H, Lin W-A, Zhou SK, Luo J. ADN: Artifact disentanglement network for unsupervised metal artifact reduction. *IEEE Trans Med Imaging* (2020) 39:634–43. doi: 10.1109/TMI.2019.2933425

19. Paudel MR, Mackenzie M, Fallone BG, Rathee S. Clinical evaluation of normalized metal artifact reduction in kVCT using MVCT prior images (MVCT-NMAR) for radiation therapy treatment planning. *Int J Radiat Oncol* (2014) 89:682–9. doi: 10.1016/j.ijrobp.2014.02.040

20. Gao L, Sun H, Ni X, Fang M, Cao Z, Lin T. Metal artifact reduction through MVCBCT and kVCT in radiotherapy. *Sci Rep* (2016) 6:37608. doi: 10.1038/srep37608

21. Paudel M, Kirvan P, Fallone B, Rathee S. SU-DD-A3-04: Evaluation of metal artifact reduction using MVCT and model based image reconstruction. *Med Phys* (2010) 37:3091–1. doi: 10.1118/1.3467997

22. Klein S, Staring M, Murphy K, Viergever MA, Pluim JPW. Elastix: A toolbox for intensity-based medical image registration. *IEEE Trans Med Imaging* (2010) 29:196–205. doi: 10.1109/TMI.2009.2035616

23. Shamonin DP, Bron EE, Lelieveldt BPF, 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

24. Zhu J-Y, Park T, Isola P, Efros AA. Unpaired image-to-Image translation using cycle-consistent adversarial networks. In: *IEEE International conference on computer vision (ICCV)*. Venice: IEEE (2017). p. pp 2242–2251. doi: 10.1109/ICCV.2017.244

25. Liang X, Chen L, Nguyen D, Zhou Z, Gu X, Yang M, et al. Generating synthesized computed tomography (CT) from cone-beam computed tomography (CBCT) using CycleGAN for adaptive radiation therapy. *Phys Med Biol* (2019) 64:125002. doi: 10.1088/1361-6560/ab22f9

26. Sun H, Fan R, Li C, Lu Z, Xie K, Ni X, et al. Imaging study of pseudo-CT synthesized from cone-beam CT based on 3D CycleGAN in radiotherapy. *Front Oncol* (2021) 11:603844. doi: 10.3389/fonc.2021.603844

27. Vinas L, Scholey J, Descovich M, Kearney V, Sudhyadhom A. Improved contrast and noise of megavoltage computed tomography (MVCT) through cycle-consistent generative machine learning. *Med Phys* (2021) 48:676–90. doi: 10.1002/mp.14616

28. Sun H, Xi Q, Fan R, Sun J, Xie K, Ni X, et al. Synthesis of pseudo-CT images from pelvic MRI images based on an MD-CycleGAN model for radiotherapy. *Phys Med Biol* (2022) 67:035006. doi: 10.1088/1361-6560/ac4123

29. Yang H, Sun J, Carass A, Zhao C, Lee J, Prince JL, et al. Unsupervised MR-to-CT synthesis using structure-constrained CycleGAN. *IEEE Trans Med Imaging* (2020) 39:4249–61. doi: 10.1109/TMI.2020.3015379

30. Yang B, Chang Y, Liang Y, Wang Z, Pei X, Xu X, et al. A comparison study between CNN-based deformed planning CT and CycleGAN-based synthetic CT methods for improving iCBCT image quality. *Front Oncol* (2022) 12:896795. doi: 10.3389/fonc.2022.896795

31. Xiao X, Lian S, Luo Z, Li S. (2018). Weighted res-UNet for high-quality retina vessel segmentation, in: 2018 9th International Conference on Information Technology in Medicine and Education (ITME), (Hangzhou, China: IEEE) 327–331. doi: 10.1109/ITME.2018.00080

32. Liu Y, Lei Y, Wang T, Fu Y, Tang X, Curran WJ, et al. CBCT-based synthetic CT generation using deep-attention cycleGAN for pancreatic adaptive radiotherapy. *Med Phys* (2020) 47:2472–83. doi: 10.1002/mp.14121

33. Bannas P, Li Y, Motosugi U, Li K, Lubner M, Chen GH, et al. Prior image constrained compressed sensing metal artifact reduction (PICCS-MAR): 2D and 3D image quality improvement with hip prostheses at CT colonography. *Eur Radiol* (2016) 26:2039–46. doi: 10.1007/s00330-015-4044-1

34. Kalender WA, Hebel R, Ebersberger J. Reduction of CT artifacts caused by metallic implants. *Radiology* (1987) 164:576–7. doi: 10.1148/radiology.164.2.3602406

35. Meyer E, Raupach R, Lell M, Schmidt B, Kachelriess M. Normalized metal artifact reduction (NMAR) in computed tomography. *Med Phys* (2010) 37:5482–93. doi: 10.1118/1.3484090

36. Morin O, Chen J, Aubin M, Gillis A, Aubry JF, Bose S, et al. Dose calculation using megavoltage cone-beam CT. *Int J Radiat Oncol* (2007) 67:1201–10. doi: 10.1016/j.ijrobp.2006.10.048

37. Zhao J, Chen Z, Wang J, Xia F, Peng J, Hu Y, et al. MV CBCT-based synthetic CT generation using a deep learning method for rectal cancer adaptive radiotherapy. *Front Oncol* (2021) 11:655325. doi: 10.3389/fonc.2021.655325

38. Harms J, Lei Y, Wang T, Zhang R, Zhou J, Tang X, et al. Paired cycle-GAN-based image correction for quantitative cone-beam computed tomography. *Med Phys* (2019) 46:3998–4009. doi: 10.1002/mp.13656

39. Chen L, Liang X, Shen C, Nguyen D, Jiang S, Wang J. Synthetic CT generation from CBCT images *via* unsupervised deep learning. *Phys Med Biol* (2021) 66:115019. doi: 10.1088/1361-6560/ac01b6

40. Li Y, Zhu J, Liu Z, Teng J, Xie Q, Zhang L, et al. A preliminary study of using a deep convolution neural network to generate synthesized CT images based on CBCT for adaptive radiotherapy of nasopharyngeal carcinoma. *Phys Med Biol* (2019) 64:145010. doi: 10.1088/1361-6560/ab2770

frontiers | Frontiers in Oncology

TYPE Original Research
PUBLISHED 11 November 2022
DOI 10.3389/fonc.2022.980312

# Automatic segmentation of nasopharyngeal carcinoma on CT images using efficient UNet-2.5D ensemble with semi-supervised pretext task pretraining

Jansen Keith L. Domoguen[1]*, Jen-Jen A. Manuel[2], Johanna Patricia A. Cañal[2] and Prospero C. Naval Jr[1]

[1]Computer Vision and Machine Intelligence Group, Department of Computer Science, University of the Philippines-Diliman, Quezon City, Philippines, [2]Division of Radiation Oncology, Department of Radiology, University of the Philippines-Philippine General Hospital, Manila, Philippines

Check for updates

OPEN ACCESS

EDITED BY
Xiaodong Wu,
The University of Iowa, United States

REVIEWED BY
Ziyue Xu,
Nvidia, United States
Lichun Zhang,
The University of Iowa, United States

*CORRESPONDENCE
Jansen Keith L. Domoguen
jldomoguen@up.edu.ph

SPECIALTY SECTION
This article was submitted to Radiation Oncology, a section of the journal Frontiers in Oncology

RECEIVED 28 June 2022
ACCEPTED 17 October 2022
PUBLISHED 11 November 2022

CITATION
Domoguen JKL, Manuel JA, Cañal JPA and Naval PC Jr (2022) Automatic segmentation of nasopharyngeal carcinoma on CT images using efficient UNet-2.5D ensemble with semi-supervised pretext task pretraining.
*Front. Oncol.* 12:980312.
doi: 10.3389/fonc.2022.980312

Nasopharyngeal carcinoma (NPC) is primarily treated with radiation therapy. Accurate delineation of target volumes and organs at risk is important. However, manual delineation is time-consuming, variable, and subjective depending on the experience of the radiation oncologist. This work explores the use of deep learning methods to automate the segmentation of NPC primary gross tumor volume (GTVp) in planning computer tomography (CT) images. A total of sixty-three (63) patients diagnosed with NPC were included in this study. Although a number of studies applied have shown the effectiveness of deep learning methods in medical imaging, their high performance has mainly been due to the wide availability of data. In contrast, the data for NPC is scarce and inaccessible. To tackle this problem, we propose two sequential approaches. First we propose a much simpler architecture which follows the UNet design but using 2D convolutional network for 3D segmentation. We find that this specific architecture is much more effective in the segmentation of GTV in NPC. We highlight its efficacy over other more popular and modern architecture by achieving significantly higher performance. Moreover to further improve performance, we trained the model using multi-scale dataset to create an ensemble of models. However, the performance of the model is ultimately dependent on the availability of labelled data. Hence building on top of this proposed architecture, we employ the use of semi-supervised learning by proposing the use of a combined pre-text tasks. Specifically we use the combination of 3D rotation and 3D relative-patch location pre-texts tasks to pretrain the feature extractor. We use an additional 50 CT images of healthy patients which have no annotation or labels. By semi-supervised pretraining the feature extractor can be frozen after pretraining which essentially makes it much more efficient in terms of the number of parameters since only the decoder is trained. Finally it is not only efficient in terms of parameters but also data, which is shown when the pretrained model with only portion of the labelled training data was able to achieve very close performance to the model trained with the full labelled data.

KEYWORDS

nasopharyngeal carcinoma, automatic volume segmentation, deep learning, radiotherapy, semi-supervised learning, pretext tasks

# 1 Introduction

Nasopharyngeal carcinoma is rare among Caucasians but one of the more common head and neck cancers found among Asians and North Africans (1). Standard treatment involves combination chemotherapy and radiotherapy. Surgery is generally done as salvage after treatment inadequacies or failures. Over the past few decades and with improved digitalization, radiation therapy has become more and more precise. This came about because of precision in both cross-sectional diagnostic imaging (CT and MRI) and radiation delivery. Precision is the key. In the process of radiotherapy, one of the most critical steps is contouring of the tumor. After all, if the target is incorrect or imprecise in any way, the subsequent treatment planning and treatment delivery will be incorrect and imprecise too.

With the advent of artificial intelligence, there is now software available for auto-contouring. All commercially available treatment planning systems contain software that can auto-contour normal structures or organs. At the present, much research is being done into auto-contouring the gross tumor volume (GTV), many of them coming out of China. Since nasopharyngeal carcinoma is considered endemic in China, it is logical that resources are being poured into creating artificial intelligence that can map nasopharyngeal tumors on CT scans and MRIs.

There are at least 6 studies that have dealt with auto-contouring of nasopharyngeal tumors using cross-sectional imaging, both CT scan and MRI (2–6). Work by (2) was one of the earliest works who applied deep learning methods on the segmentation of NPC. They proposed a modified UNet architecture where the downsampling and upsampling layers have similar number of parameters to ensure that the output resolution is exactly the same as the input. Moreover, their work also analyzed the performance of deep neural networks across different tumors stages as well as predicting gross nodal volumes. They observed significant performance degradation as the tumor stage increases and a much lower performance for gross nodal volumes. In contrast to our work, we don't distinguish tumor stage for our performance analysis. Work by (3) proposed a novel 3D convolutional network which uses cascaded multi-scale local enhancement for convolutional networks. Specifically they adopted the 3D Res-UNet as their backbone network and employed a multi-scale dilated convolutional block to enhance extracted receptive field and improve focus on the target tumor especially its boundary. This is then integrated to a central localization cascade model to concentrate on the gross tumor volume for fine segmentation. The work by (4) is most similar to ours as they also employed ensemble model based on multi-scale sampling, however they employed a projection block and attention block to improve the extracted representation. The projection block is similar to the popular "SqueezeExcite" (7) method used to improve the learned representation. However, in this case they squeeze the feature maps across the three dimensions which they later combined *via* summation operation across the spatial dimension and finally a projection to the depth dimension which recovers the original shape of the feature map. The attention module is a spatial attention block that focuses and refines extracted representation especially for very small tumors which is common in NPC. Despite the addition of more sophisticated blocks, we find their method under performs compared to purely using the UNet-2.5D which uses much fewer learning parameters. Although the work by (6) used magnetic resonance images in contrast to CT scans, they demonstrated that by combining the T1-weighted (T1W) and T2-weighted (T2W) MRI images of each patient provides significant performance boost. These two sequences were combined by their proposed dense connectivity embedding, which essentially fuses the feature maps of each modes across the layers in the encoder. Furthermore, a convolutional block is introduced to process the fused embedding which will then act as a skip connections to their corresponding decoder block in a UNet architecture. While MRI would instinctively be the better imaging modality to become the basis for auto-contouring, MRI is not always readily available in all countries, especially developing countries.

The use of deep learning in medical imaging has become a popular alternative for practitioners to automatically generate accurate target delineation. Furthermore, it does not only resolve the time-consuming and tedious task of manual contouring but can also alleviate the problem of inter-observer variability by generating more robust predictions since it learns from different sources. This problem occurs when radiation oncologists disagree on the delineated gross tumor volume brought by the inherent subjectivity of the annotation process itself. This depends on variety of factors notably years of experience of the practitioner. However, though deep learning models have the potential to generate significant benefits in the medical imaging field, it is also a poor field to apply these methods to. This is because data in this field are notoriously difficult and expensive to collect. And when this is coupled with the fact that deep learning models are only as good as the quantity and quality of the data you have, then the objective is to not only generate accurate models but also models that can perform well when there are few data. To this end, we employ self-supervised learning (SSL). SSL method has become the mainstream approach in mitigating problems regarding data scarcity when utilizing deep learning in the medical setting. It is able to leverage unannotated scans by using a predefined pre-text tasks (self-supervision task) which is used to train a feature-extractor or the encoder network. Ideally, this pre-text task should be able to help the encoder network or the feature extractor learn features and representations such as the generic structure, texture, and other salient features that can be re-used during the *downstream task* or the actual target task which is in our case the segmentation of GTV in NPC. Hence it will require much fewer annotated data during the downstream task making it data efficient. For our work, we used an equal number of unannotated and annotated NPC CT scans. In contrast with other works which used single pre-text task during SSL pretraining we used multiple pre-text tasks to pre-train our encoder network. Specifically we use a combination of relative-positional location (RPL) and rotation methods to pre-train our encoder network. This encoder network can then be frozen and attached to a decoder network used for the segmentation task. The goal is that by employing SSL pretraining, the feature extractor will be in a much better starting position to easily learn the diverse morphologies and sizes of the gross tumor volume even with much fewer data.

Self-supervised learning in medical images (8–14) is usually an extension of the self-supervised techniques used in 2D natural images. The seminal work by (15) proposed different pretext tasks for 3D medical image that were originally based on 2D images. Multiple pretext tasks specialized for 3D medical images are proposed such as: contrastive predictive coding, rotation prediction, jigsaw puzzles, relative patch location, and exemplar methods. The predictive coding pretext task first divides an input 3D cube into smaller cubes which are individually encoded by the network. Given a set of consecutive encoded cubes, the network must find and choose the next consecutive cube out of a set potential cubes based on their encoding. Hence, in order to accomplish this task, the network must be forced to learn the specific fine-grain morphology and structure of the volume in order to correctly predict the next adjacent cube. And because this uses contrastive learning (16), the encoding or representation of adjacent cubes are much closer than cubes that are farther away. This conforms with the actual input volume where adjacent volumes have very similar features. The important consideration here is that the network was able to learn and distinguish the feature, structure and morphology of the volume even without labels by doing this pretext task. This is essentially the same case with all pretext tasks, for rotation it randomly rotates the volume from a predefined class of orientation, then the network must predict the specific orientation but in order to correctly predict the orientation it must understand the structure of the volume. For relative-patch location, it randomly crops an input volume then divides the volume further into 27 non-overlapping cubes. It then uses the central cube to predict the location of a cube randomly queried which has a total of 26 possible locations or classes. In our work we employ the relative patch location and rotation pre-text tasks for our proposed SSL since they are much simpler and were able to produce significantly higher performance over the other pretext tasks. More recent work by (17) proposed a spatially guided self-supervised clustering network (SGSCN) for downstream medical image segmentation. They proposed using multiple loss functions to train a network in an end-to-end manner in order to group image pixels that are spatially connected and thus have similar representations. In addition, a context-based consistency loss is used to better learn the boundaries and shape of the target volume. Finally work by (18) proposed the use of auxiliary tasks for task-level consistency as an SSL approach. Specifically two auxiliary tasks are used where one task is responsible for foreground-background reconstruction aimed for in-formation segmentation while the other task employs a mean-teacher architecture to perform signed distance field (SDF) prediction to enforce shape constraints. All these SSL methods were proposed mainly to address the limited availability of labeled data while exploiting abundance of unlabeled data. Similar to ours we propose an SSL approach that uses a combination of pretext tasks to help a feature extractor learn representations from unlabeled dataset that are highly relevant to its downstream segmentation task.

Filipino oncologists have always been aware of the high number of cases of nasopharyngeal carcinoma in the Philippines based on their individual experiences in their own clinics and hospitals. The true number cannot be verified because of the absence of a government-run nationwide cancer registry. Because of the number of patients with nasopharyngeal carcinoma at our institution and the consequent volume of imaging data, we felt that it would be a good venue for the creation of auto-segmentation/auto-contouring software for radiation oncology use. Moreover, since modern deep learning methods are notoriously hungry for labelled data, we introduce a self-supervised method to compliment the development and training of our proposed deep learning method. This will mitigate overfitting introduced due to very few labelled data thereby improving performance as well as allows it to exploit unlabeled data which are often much more abundant than labelled data. This cuts costs in terms of the resources and time required to label more data to improve model performance.

Collaboration between researchers from the Department of Computer Science of the College of Engineering and the Division of Radiation Oncology of the Department of Radiology of the UP-Philippine General Hospital resulted in this study that set out to create a software that could accurately contour nasopharyngeal tumors on appropriately acquired CT scan images.

# 2 Methods

## 2.1 Network architecture

The primary network architecture used in this work is the UNet-2.5D (4) network based on the UNet3D (19). This is shown in Figure 1 where the main difference compared with UNet-3D lies in the 2D convolutional block that UNet-2.5D uses. Following (19) our architecture consists of nine convolutional blocks where each block consists of two convolutional layers interleaved with Batch Normalization (20) and RELU non-linearity (21).

The difference between UNet-3D and UNet-2.5D is the dimension of the convolutional layer. UNet-3D uses 3D convolutional layer across its block whereas UNet-2.5D utilizes 2D except for the center or bottleneck block which uses 3D convolution. In general, the performance of a model is better with higher number of parameters and convolutions. However, this is not the case as we will show in our results. The simpler and lighter UNet-2.5D network – in terms of number of parameters and operations – significantly outperforms the UNet-3D network. This is because heavier networks such as UNet-3D require more data as there are more parameters to train.

In a setting where limited data are available such as our case (i.e., NPC CT images), the parameters will easily overfit the training data making the model unable to generalize its prediction to test data.

## 2.2 Multi-scale training

In general, training time is proportional to the size of the data fed to the model before it converges. In our case, the GTV in NPC is smaller relative to the entire patient's body. Hence instead of feeding the entire volume as input to the model, we cropped the input volume along the x, y, and z directions as suggested by (4) using multiple scales encompassing the GTV. Five scales are extracted to generate five datasets. These are *extra-small*, *small*, *medium*, *large*, and *extra-large*. The smallest scale is randomly cropped across x, y, and z direction to extract a volume that contains the smallest spatial resolution and depth. By extracting the smallest volume, we ensure that we extract only the local information of the structure and feature of that given volume. The largest scale, on the other hand, captures almost the entire volume with the information extracted mostly globally.

The crop-size used to extract the data for each scale decreases in a fixed percentage as the scale decreases from extra-large to extra-small. Along the z direction, the length (number of slices) of the original input volume is cropped beginning at 90% with constant decrement of 10% as the scale decreases. For the x, y dimension, the patch for the large-scale starts at 100% of the resolution (i.e., 512 x 512 pixels) then cuts

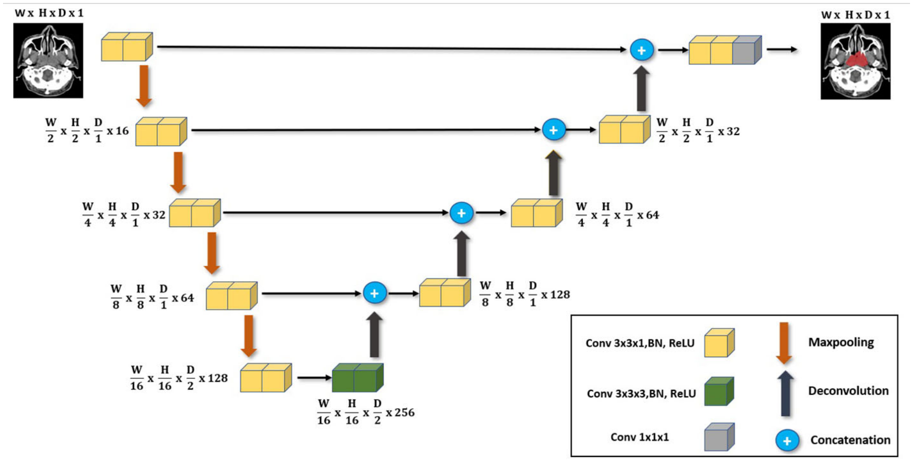


FIGURE 1
The main deep neural network architecture used in our work. It follows the same UNet architecture but with the main use of 3x3x1 convolutions instead of 3x3x3 convolutions employed for 3D volume segmentation. We highlight its effectiveness when used in data-scarce setting as it is less likely to overfit.

with decrement of 15% as the scale decreases. This can be seen in Figure 2. Each of the five datasets generated is then used to train a corresponding model.

During testing, the outputs of the five trained models given the same input are aggregated to produce a single result. Empirically, this produces a much more robust result compared to simply using a single model because each model is specialized to a certain scale. Since the features of NPC varies widely in terms of size and shape, models trained using the small-scale and large-scale datasets perform better in detecting tumors that are small and large, respectively.

The rationale behind using multi-scale training is specializing each model to a certain feature or context of the volume. As it will be shown, this approach achieves significantly superior results compared to training a single model.

## 2.3 Ensemble of models

Five models were trained using the five scaled dataset obtained *via* multi-scale cropping: extra-small, small, medium, large, extra-large. During the evaluation/testing stage, we used a fixed and uncropped raw CT input scans to evaluate the performance of each of the five models. Each of the models was to make a separate prediction in the form of probability maps. To create a model ensemble, the probability maps from all the five models for a specific input are averaged to produce a single probability map. This will be used to create the final segmentation mask.

Using this model ensemble approach produces a more robust prediction. Moreover, model ensembles also boost model performance compared to using a single model. This is due to having richer and more diverse predictions from each model that is specialized to a specific scale. The drawback of this approach, however is the higher computer and memory requirements. To mitigate this, the UNet-2.5D is used since it only uses a single 3D convolution with 2D convolution for the rest of the layers. This architecture is much more lightweight and less computationally intensive than 3D convolutions

## 2.4 Semi-supervised pretraining

### 2.4.1 Pretext tasks for self-supervised learning

The main pretext tasks used in this work are shown in Figure 3 which were introduced by (15), these are relative patch location and rotation pretext tasks. The rotation pre-text task shown in Figure 3B, is one of the simplest pretext tasks and therefore can easily be implemented in any setting. The goal of the rotation pretext task is to simply predict the angle of rotation for an input data that was rotated for a specific angle. We fix the possible angles of rotation to 0°, 90°, 180°and 270°. Since there are three axis of rotations, there will be a total of 10 possible angles (since 0°is redundant for the three axis) for an 3D input image. Hence this pretext task is essentially a multi-class (10 classes) classification task where each class consists of a particular rotation angle for a specific axis. The goal is by predicting the 3D rotation of each volume, the encoder network will be forced to learn the structure of the volume and hence relevant features that can be re-used when making downstream segmentation tasks. However due to its simplicity, the features learned at convergence of the rotation pretext task may not be enough in providing the necessary features for the downstream segmentation task.

To mitigate this insufficiency, we combine the relative patch location (RPL) pretext task shown in Figure 3A, which consists of predicting the location of a query patch relative to a fixed anchor patch. This self-supervision task enables the model to learn a much richer structural and finer grained information within the data. This is crucial for 3D segmentation task since it

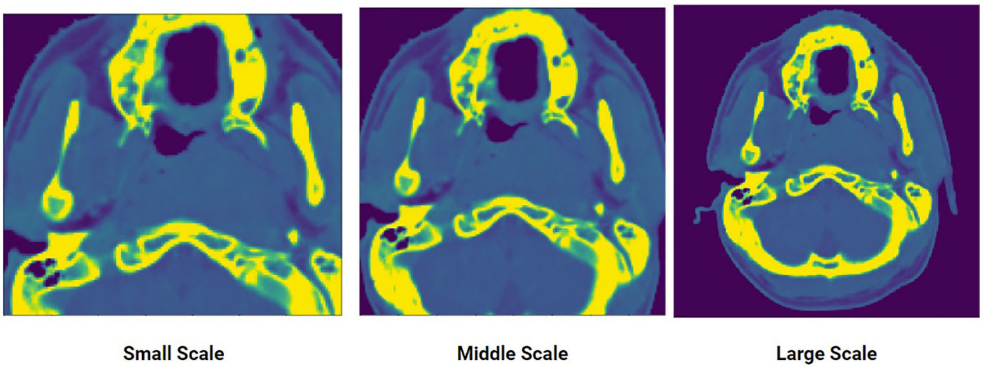


FIGURE 2
An input CT scan showing three different scales of the same scan: small, middle, large. Five scales were generated to create the multi-scale training data.

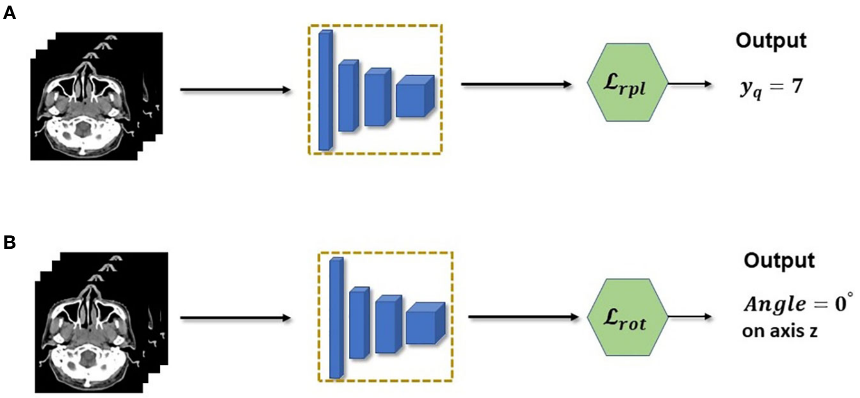


FIGURE 3
**(A)** Relative patch location is a pretext task used to pretrain the feature extractor. In practice, the task is a multi-class classification which predicts the location of a query patch. **(B)** Rotation pretext task is also casted as a multi-class classification but with rotations as the class.

needs to understand the structure and spatial features of the input to correctly predict each voxels. Discretely, the RPL pretext task is implemented by dividing a 3D input image into a 3×3×3 grid to create a total of 27 non-overlapping patches $\{x_{i\in\{1,\ldots,N\}}\}$. The central patch $x_c$will be used as the fixed anchor patch and a query patch $x_q$will be randomly sampled from the remaining set of patches $\{\{y_n\}$. The pretext task trains an encoder model to learn the location of the query patch with respect to the central patch by predicting a location $y\hat{}_q$. Since there are total of 26 patches (central patch is excluded), the encoder will be trained using a multi-class (26 classes) classification similar to the rotation pretext task. In this case it is predicting the class location instead of the angle of rotation. However, it is different in that it needs to fuse both the query and anchor patch together and make the location prediction based on this fused information. This is further shown in the following equation:

$$\mathcal{L}_{RPL} = -\sum_{k=1}^{K} \log p(y_q | \hat{y}_q, \{y_n\}) \quad (1)$$

where $y_q$corresponds to the groundtruth location of the patch.

### 2.4.2 Combining pretext tasks for richer representation

We combined the two pretext tasks shown above in order to force the encoder to learn a synergy in the feature representation that is extracted from each image. This is because the encoder needs to learn how to combine, segregate and choose the representations that are most relevant for the two tasks. And since each pretext task have varying objective, the representation should be compact and sufficient for the two tasks. Moreover, by using two pretext tasks simultaneously, the encoder will need to learn rich and diverse representations that will be much more useful for the downstream task. Since this is purely self-supervised, training our encoder network is much more data and parameter efficient. This is because it can leverage the use of unlabeled CT scans while using lighter network. The schematic of our SSL approach is shown in Figure 4 where each image is fed to two pre-processing blocks before being fed simultaneously to the encoder network. The overall loss function therefore is shown below:

$$L = \alpha L_{RPL} + (1 - \alpha)L_{Rot} \quad (2)$$

where $\alpha$ is the weighting factor to balance and control the contribution of each pretext task.

### 2.4.3 Efficient model ensemble

Although model ensemble have been very effective in improving the performance and robustness of the model by relying on independent weak learners in traditional machine learning, it is usually impractical to use it directly in deep learning. As was discussed above, to create the model ensemble, five models are trained on five different scales of the dataset which generates five trained models. This can be computationally prohibitive especially in very deep network which can be more expensive than the performance boost it provides. Our proposed SSL approach can help mitigate this since we can essentially *freeze* and *re-use* the encoder network that was pretrained during the SSL. We can have essentially a single unified encoder network while only training or finetuning the decoder of each model in the ensemble. A diagram of this approach is shown in Figure 5. This makes our method much more parameter efficient during both training and inference.

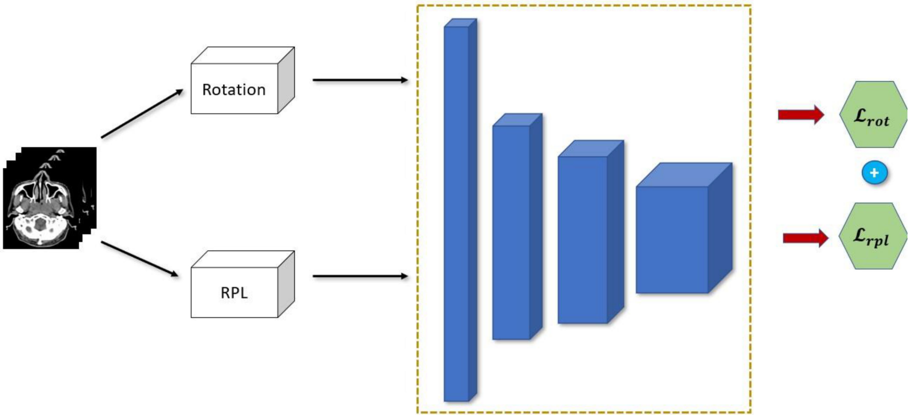


FIGURE 4
Proposed combination of semi-supervised learning pre-text tasks.

# 3 Materials and methods

## 3.1 Clinical material

At the outset, it was decided that we would attempt to create an auto-segmentation or auto-contouring program using nasopharyngeal tumors. With nasopharyngeal tumors, the tumors are confined to a single anatomic space and, there is no need to account for movement, swallowing and breathing. Additionally, nasopharyngeal cancers are relatively common in the Philippines, making this work impactful.

A review of census of nasopharyngeal cancer patients at the Division of Radiation Oncology was done, covering May 2017 — when operations of the linear accelerator started — until February 2020 — just before the start of the COVID lockdown. A total of 79 patient records were retrieved. Patients who were less than 18 years of age and had non-carcinoma tumors, i.e. lymphomas, sarcomas, were excluded from consideration. The images of the remaining 63 patients — 44 males and 19 females ranging in age from 18 to 73 and covering all tumor stages — were used in this paper. Individual patient consent was waived because of the use of just the images and the retrospective nature of this study. A total of 50 healthy patients were also collected to be used for the semi-supervised pretraining of the encoder network.

Shown in Table 1 are the baseline characteristics of the 63 NPC patients included in this study. Fifty-three (53) of these patients were randomly selected to be used in the training and validation of our models. The remaining ten (10) patients were utilized during testing. Majority of patients included in the study were male. More than half of patients had T4 disease based on the American Joint Committee on Cancer (AJCC) Cancer Staging Manual Eight edition for nasopharyngeal carcinoma.

Simulation computed tomography (CT) images with contrast were acquired using a SOMATOM Emotion 16 (Siemens Healthineers). All patients were positioned in supine and immobilized using a head, neck, and shoulder thermoplastic mask. Scanning range was from vertex to carina. Obtained CT images were reconstructed using a matrix of 512 × 512 with thickness of 3.0 mm. Delineation of the primary gross tumor volume on CT images was then performed by an experienced radiation oncologist. The contoured images — all in DICOM format — were anonymized before being subjected to computer "training." Ten (10) image sets were randomly chosen and used initially for testing. The remaining fifty-three (53) image sets were used in the training and validation of the software model. For training to commence on these images, these had to be in a suitable format to be processed by the proposed deep learning model. The array volumes (3D tensor) were extracted from the DICOM files, ensuring isotropic resolution. A uniform resolution of 1.0 ×1.0×3.0mm$^3$ was enforced. The Hounsfield Units (HU) of all images (originally ranging from -1024 to 3071) were truncated and normalized to [-150, 500]. All values above and below this range were set to zero.

## 3.2 Data preprocessing and augmentation

### 3.2.1 Data preprocessing

The actual raw data that is frequently used by radiation oncologists are in a DICOM (22) format and while it is extremely useful in their specialized software tools, it cannot be directly processed by our deep learning model. It needs to be cleaned and transformed to a suitable format for training and inference. The first step is the extraction of the array volume (3D

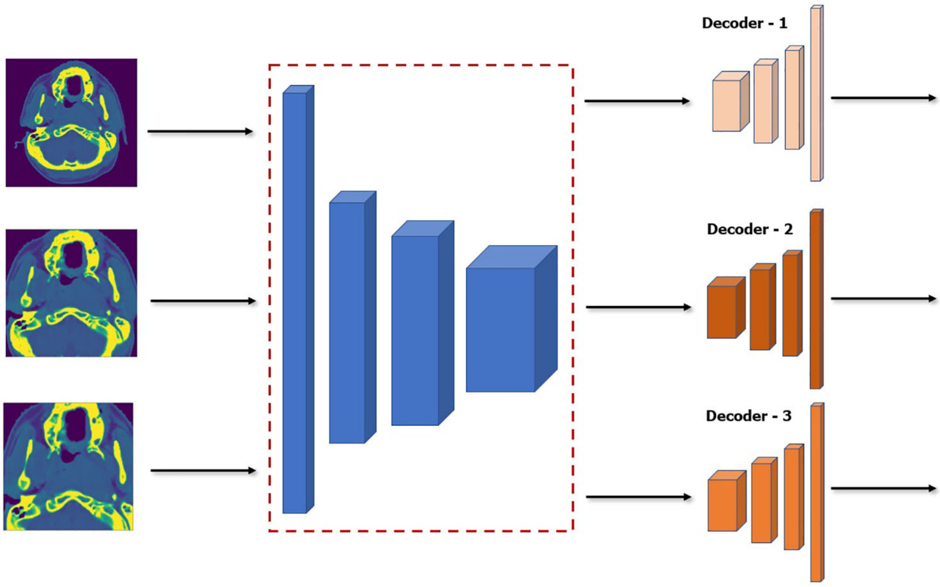


FIGURE 5
Architecture of model ensemble with a single encoder network while finetuning three decoder networks. Note that the encoder is frozen and thus will not be affected during finetuning.

tensor) from the DICOM files and ensuring an isotropic resolution. Although the thickness of each patient's CT scan are all 3.0 mm, the x,y pixel spacing range varies from patients to patient. We therefore enforce a uniform resolution of 1.0 ×1.0×3.0mm$^3$ for the x, y, z spacing by uniform interpolation. Afterwards, since the raw array values of the DICOM files are in Hounsfield unit which ranges between -1024 and 3071 HU for each voxel, we truncate and normalize their values. We find that for body and NPC, they have a distinct distribution of Hounsfield values. We therefore truncated the Hounsfield values to [-150,500] where outside these range are all automatically set to zero. This are then finally normalized to [0,1].

TABLE 1 Demographic characteristics of the NPC patients included in the study.

| Characteristics | Total number of patients<br>n = 63 |
|---|---|
| **Median age (range)** | **45 (18 – 73)** |
| Sex | |
| Male | 44 |
| Female | 19 |
| T classification | |
| T1 | 6 |
| T2 | 10 |
| T3 | 14 |
| T4 | 33 |

### 3.2.2 Data augmentation

Since we have very little training data, there is a high chance that the model may overfit hence we employ different data augmentation techniques that can increase data samples and thus improve model performance. We employ the most effective data augmentation techniques: rotation, flipping, cropping and transposing which are randomly applied across the x,y,z dimension for each batch size iteration during training. Flipping is also especially important as observed by (23), which highlighted that it improves the model's robustness on different tumor shapes and which is especially important for our use case because NPC has many different shapes.

## 3.3 Evaluation metrics

There are a total of eight evaluation metrics used in the experiment, the primary performance evaluation metric and most commonly used ones are the Dice-Similarity Coefficient (DSC) and Intersection-OverUnion (IOU) metrics. Both metrics measure the same overlap between the groundtruth and the predicted mask and have a range between 0 and 1 where 0 means totally no overlap and 1 means perfect overlap. Though it may be tempting to view both metrics functionally equivalent, their distinction arises when taking their average values across set of samples. Specifically, IOU score penalizes wrong predictions much more than DSC. Thus IOU score can be thought of as measuring the lower bound of the model performance while the

DSC measures the average model performance across the test data. We can expect therefore that IOU usually outputs a significantly lower score compared to DSC. This is highlighted later in the section.

The two other metrics are grouped under the distance metric which measures the distance between two sets that contain point coordinates from both the groundtruth and points predicted by the segmentation model. These metrics are the Average Symmetric Surface Distance (ASSD) and the Hausdorff Distance. The ASSD determines the average difference (24) between the surface of the predicted and groundtruth volumes. The surface points from both the prediction and groundtruth surfaces are sampled from a set of points that are not part of a predefined neighborhood. These points can be thought of as the outlier or the gap with respect to the groundtruth surface. The closest distance of each of these outlier points are then taken against the points in the other surface. The average distance of these points will be the ASSD which will be in a mm unit. ASSD score will be 0mm for perfect segmentation, with increasing score corresponding to worsening performance of the model. The Hausdorff Distance is similar to the ASSD except that it does not measure the average distance between outliers of two surfaces, but rather measures the maximum distance of randomly samples points from the two volumes to create two sets. The Hausdorff Distance is then the maximum distance from a point in one set to the closest point in the other set. Again the lower the distance, the closer the points between the groundtruth and predicted volumes are.

The final four metrics are: sensitivity, relative volume error, and positive predictive value (PPV). These are the most commonly used metrics for medical image segmentation in deep learning. The sensitivity, also referred to as true positive rate quantifies the model's ability to correctly detect the voxels that is indeed an NPC or tumor. It measures the proportion of voxels in the volume that are truly tumors and are correctly detected by the model. Finally the PPV is simply the ratio of voxels that were correctly identified as tumors to the voxels that were identified to be tumors. Or essentially it is the probability that the voxels that were predicted as tumors are indeed tumors.

### 3.4 Post-processing

Post-processing involves the aggregation of the individual model prediction in the ensemble and a heuristic-based post-processing to further refine the prediction. It has been observed that the aggregated output from the model ensemble still have some residual volumes that are sparsely distributed and are not attached from the largest volume prediction. Since it is assumed that we are only predicting the primary tumor volume, the final prediction should only have a single large solid volume. Hence we first perform a series of morphological operation (i.e., erosion and dilation) to remove the edges in the volume. Afterwards, for each 2D-slice of the volume, a contour search is applied to get only the largest 2D contiguous mask and removing the remaining 2D contours. This operation is applied across all the slice in a volume essentially taking only the largest connected region as the final volume prediction.

### 3.5 K-fold cross validation and implementation details

Since there are a total of 63 patients, we performed 7-fold cross-validation where 54 patients are used for training and validation and the remaining 9 patients will be used for evaluation for the final model. The final performance is averaged across the seven folds. During the training run, training validation data for each fold is split 80/20 respectively, where the validation is used to tune the hyperparameters. After finding the optimal hyperparameters, the training and validation data are combined to generate a final model which will be evaluated on the test dataset.

All the experiments were implemented using the Pytorch deep learning framework using NVIDIA RTX 2080Ti Graphical Processing Unit (GPU) 11GB. The ADAM (25) optimizer was employed to train our deep learning network using an initial learning rate of $1 \times 10^{-3}$ and with a decay factor of $1 \times 10^{-4}$ for every 150 epochs. The whole training-validation run takes a total of 900 epochs using 32 batchsize for each iteration. Moreover, random cropping is done where volume of patches is randomly extracted from each patient volume and fed to the network. This approach mitigates the memory constraint in the GPU and speeds up loss convergence. This is applied for the whole five-scaled dataset to generate five pretrained models for inference and testing.

## 4 Results and discussion

### 4.1 Method comparison

We evaluate first our proposed approach(UNet-2.5D) against different architectures commonly employed for medical image segmentation. The other architectures tested are UNet-3D, VNet and the UNet + Project Excite(PE) + Attention Module(AM) by (4) proposed specifically for the segmentation of GTV in NPC. We show that with the simpler UNet-2.5D architecture, it significantly outperforms the generic UNet architectures as well as the network proposed by (4).

Moreover, we compare our method on popular architectures that has gained state-of-the-art performance on multiple benchmark dataset. One is the Generic Autodidactic Models or Genesis model proposed by (26). The Genesis model aims to provide a generic source model that can be transferred on different application-specific target task. It achieved broad performance improvement over different medical segmentation benchmark dataset from chest to brain data. In our case we use their pretrained UGenesis model trained on chest CT as our base model then finetune it to our dataset. Another very popular method that achieved multiple SOTA results is the no-new-Unet or nnUNet by (27). They have shown that for a fully optimized network, "architectural tweaking" provides no improvement in the segmentation performance, and the influence of non-architectural aspects in segmentation methods is much more impactful. nnUNet offers an end-to-end automated pipeline that is adaptable to any medical dataset. It has an automated pipeline for preprocessing, data augmentation, and post-processing. It can also automatically infer important hyperparameters such as normalization, resampling and batchsize optimized for the given dataset. For our case, we employ the nnUNet for all the three available architecture types: 2D, Fully 3D and Low Resolution 3D. We use the same seven-fold cross validation for all the evaluation runs.

Except for nnUNet, all the different segmentation methods made use of the *medium-scale* preprocessed data as their training set. This is because data preprocessing from raw data is part of nnUNet's automated pipeline.

The quantitative results for DSC, IOU, PPV and RVE are shown in Table 2. These values are the average value (and standard deviation) from the seven fold cross-validation discussed above. Results show that UNet-2.5D network generally outperforms the other methods except in PPV. Since the bulk of the convolutional blocks used in our network is 2D convolutions, this may suggest that for the segmentation of gross tumor volume in NPC, the across-slice or depth-wise information does not really improve the performance. This also means that the 2D spatial information is more than enough to achieve high predictive performance. Moreover, it seems adding 3D information in predicting each voxel may actually hurt the segmentation performance as shown in VNet and UNet-3D architectures. This may be due to the structural characteristics of the NPC tumor itself, which has a random and irregular tumor structure. Aside from not adding any performance benefits, the added parameters using 3D convolution will only hurt performance because of overfitting. This makes the proposed approach not only much more powerful in segmenting NPC tumors but more efficient as it mostly uses 2D convolution with a single 3D convolution at the bottleneck region of the network.

Although the method proposed by (4) was able to achieve the highest PPV in Table 2, the addition of Projection-Excitation and Attention-Module blocks did not significantly achieve high performance on the other metrics.

Our method was also compared on other nnUNet and UGenesis family which were all outperformed by our method. The nnUNet "3D low resolution" variant was able to achieve the highest DSC score but generally under performed in relative to even the generic networks. UGenesis with the use of a pretrained model significantly underperformed across all the metrics. This is probably due to overfitting as the number of parameters and network architecture of UGenesis is much deeper.

Results for ASSD, Hausdorff distance and sensitivity for the different architectures are shown in Table 3. Compared to Table 2, our method was only able to decisively outperform other methods in the sensitivity metric. The highest performance for the ASSD and Hausdorff metrics were achieved generally by the nnUNet family although our method is still relatively competitive especially in ASSD metric where our method is statistically equal when taking into account their standard deviation.

## 4.2 Ensemble results

As discussed above in order to create a more robust, less data-scale dependent model as well as to boost performance, we generated five versions of the training dataset with different scales and generated five models to create an ensemble of model. We used our proposed architecture for the architecture of all the five models which we have established to be superior on majority of metrics in Tables 2, 3. These five models constituted the model ensemble. The performance of each model in the ensemble is shown in Table 4. As shown, models have different performance across different data scale. Notably the

TABLE 2 Comparative result of different deep neural network architectures for DSC, IOU, PPV and RVE.

| Method | DSC (%) ↑ | IOU (%) ↑ | PPV (%) ↑ | RVE (%) ↓ |
|---|---|---|---|---|
| UNet-3D | 66.01 ± 5.29 | 43.54 ± 3.49 | 86.03 ± 6.89 | 55.14 ± 4.42 |
| VNet | 64.25 ± 7.06 | 46.74 ± 5.13 | 70.23 ± 7.71 | 59.55 ± 0.86 |
| UNet-2.5D+PE +AM | 67.54 ± 2.16 | 51.15 ± 1.63 | 90.32 ± 2.87 | 38.21 ± 1.22 |
| UGenesis | 58.30 ± 7.31 | 41.68 ± 5.22 | 83.35 ± 10.44 | 45.24 ± 5.67 |
| nnUNet-2D | 63.14 ± 5.52 | 52.68 ± 4.61 | 63.69 ± 5.57 | 12.57 ± 1.10 |
| nnUNet-3D Full | 65.50 ± 8.43 | 54.65 ± 7.03 | 66.01 ± 8.50 | 13.05 ± 1.68 |
| nnUNet-3D Low Res. | 66.22 ± 7.94 | 55.25 ± 6.63 | 66.80 ± 8.01 | 13.19 ± 1.58 |
| UNet-2.5D (Ours) | 72.47 ± 4.10 | 60.46 ± 3.42 | 73.09 ± 4.14 | 14.43 ± 0.82 |

↑ means that higher means better while ↓ symbol means lower is better.

model trained with medium-scale outperformed the rest of the models including the aggregated ensemble performance for the DSC and RVE metrics, while the model trained on extra-small scale data achieved the best performance for the IOU metric. This means that some data scales offer the optimal information for different metrics, such as tumor's structure, topology and texture which are more likely to be emphasized in a specific data scale. The optimal inference therefore can be obtained by averaging and combining the predictions of the five models. In a way by coming the predictions, the voxel tumor that were missed by one model because it was trained on small scale dataset may be found by model trained on the large-scale dataset. This is very useful especially in the case of NPC segmentation where the GTV have diverse morphology and sizes. This mimics a kind of majority voting for a specific voxel across the models which makes it much more robust. This also offers a kind of confidence for the model prediction. Furthermore, this allows us to measure uncertainty of model prediction.

We also evaluated the performance of each model in the ensemble for ASSD, Hausdorff distance and sensitivity. The highest performance for ASSD and Hausdorff metrics where conclusively achieved by the ensemble-model. This makes sense since most of the uncertainty and difference in segmentation occurs around the boundary of the GTV. By using the prediction of the ensemble model, the boundary predictions have more confidence (when majority of models predict that a boundary voxel is a GTV) and false positive predictions are removed (when only a single model predicts that a voxel is a GTV). Although the ensemble model was not able to achieve the best performance for the sensitivity it is still relatively close and competitive.

## 4.3 Semi-supervised learning pretraining results

As mentioned in the discussion above, we used a semi-supervised learning method through the combined *Rotation +RPL* pretext tasks training to generate an encoder block that can extract sufficient representation even with few data. Moreover as the encoder block is assumed to be capable to extract sufficient representation for segmentation performance we can therefore freeze the encoder block during finetuning for the GTV segmentation. This effectively means that we will only finetune and train the decoder block which is very efficient especially when employing multi-scale training for model ensemble. This is quantitatively shown in the number of network parameters that needs to be trained when using a full model compared when the encoder is frozen, as shown in Table 5.

The number of parameters for the full UNet-2.5D network is more than *4x* the number of parameters compared to when the encoder is frozen which makes sense since the encoder or feature extractor is the backbone network. This efficiency is further increased when doing a full ensemble model as we do not need to create separate encoders across different models trained on different data scale since we can re-use the frozen encoder. Since the power of a network depends directly on the number of parameters that it can use to model the data, performance will naturally degrade if you use fewer parameters however since the encoder was pre-trained, the knowledge it gained during the pretext task is very useful and transferable during the segmentation of GTV and there might no significant performance degradation. In our case, we observed minimal performance degradation compared to the performance shown in Tables 4, 6, which we performed the same exact evaluation. These results are shown in Tables 7, 8. For Table 7, there is very small performance degradation in the model ensemble performance for DSC and RVE metrics. In fact for the IOU and PPV metric, the model ensemble performance with the SSL-trained encoder achieves higher performance albeit slight increase. Hence this method is not only much more parameter efficient but is actually on par with the performance of a full model.

The model ensemble performance of the SSL-trained encoder in the distance metrics shown in Table 8 shows a relatively steeper performance degradation. For the ASSD and

TABLE 3 Comparative results using distance metric as another measure between predicted and groundtruth contours.

| Method | ASSD (mm) ↓ | Hausdorff (mm) ↓ | Sensitivity (%) ↑ |
|---|---|---|---|
| UNet-3D | 7.55 ± 0.61 | 27.83 ± 2.23 | 59.56 ± 4.77 |
| VNet | 7.84 ± 0.86 | 33.28 ± 3.65 | 50.61 ± 5.45 |
| UNet-2.5D+PE +AM | 5.42 ± 0.17 | 25.52 ± 0.81 | 55.87 ± 1.78 |
| UGenesis | 6.15 ± 0.77 | 25.61 ± 3.21 | 49.86 ± 6.25 |
| nnUNet-2D | 3.31 ± 0.28 | 14.58 ± 1.27 | 63.91 ± 5.59 |
| nnUNet-3D Full | 3.43 ± 0.44 | 15.12 ± 1.95 | 66.29 ± 8.83 |
| nnUNet-3D Low Res. | 3.47 ± 0.42 | 15.29 ± 1.83 | 67.03 ± 8.04 |
| UNet-2.5D (Ours) | 3.79 ± 0.21 | 16.73 ± 0.94 | 73.35 ± 4.15% |

↑ means that higher means better while ↓ symbol means lower is better.

TABLE 4 Model ensemble performance for DSC, IOU, PPV and RVE, where each data scale corresponds to a separate and unique model trained on that specific data scale.

**Model ensemble performance**

| Data-scale | DSC (%) ↑ | IOU (%) ↑ | PPV (%) ↑ | RVE (%) ↓ |
|---|---|---|---|---|
| Extra-Small | 69.85 ± 4.06 | 61.23 ± 3.56 | 74.18 ± 4.31 | 16.56 ± 0.96 |
| Small | 72.13 ± 3.20 | 58.93 ± 2.62 | 76.32 ± 3.39 | 15.15 ± 0.67 |
| Medium | 72.47 ± 4.10 | 60.46 ± 3.42 | 73.01 ± 4.14 | 14.44 ± 0.82 |
| Large | 71.06 ± 3.31 | 58.12 ± 2.71 | 67.86 ± 3.16 | 26.35 ± 1.41 |
| Extra-Large | 66.41 ± 6.44 | 54.62 ± 5.30 | 68.92 ± 6.68 | 16.10 ± 1.56 |
| Ensemble Model | 72.02 ± 4.13% | 60.87 ± 3.40 | 74.61 ± 4.19 | 15.97 ± 0.83 |

↑ means that higher means better while ↓ symbol means lower is better.

Hausdorff metrics, the full model previously achieved 4.22mm and 15.73mm respectively while the SSL-trained encoder degraded to 6.49mm and 16.22mm. However in the sensitivity metric, SSL-trained encoder outperformed the sensitivity value achieved in the previous model-ensemble, 72.65% vs. 71.43%.

Aside from parameter efficiency, a more important benefits that the SSL-trained encoder can provide is data efficiency. More specifically labelled data efficiency which means that a model can achieve a specific level of performance with only a fraction or portion of the data. This has a greater practical advantage both to the doctors and annotators. They can save much more time and effort in generating, collecting and actually annotating data if the model requires much fewer data to achieve a specific baseline performance. We test to see the effectiveness of using SSL-trained encoder in achieving data efficiency. To see this we use different portions of the labelled data for training and evaluate their DSC. Moreover we compare the performance of the SSL-trained encoder to the full model to essentially see whether the use of SSL-trained model is indeed data efficient. The quantitative results for this experiment are shown in Table 9 which shows that an SSL-trained network with frozen encoder significantly outperforms the full model especially with very limited number of data as highlighted when the portion of labelled data is 10% and 30%, the SSL-trained frozen encoder was able to outperform the full model by 38.64% and 31.08% higher performance respectively. This shrinks as the number of data increases which is later outperformed by the full model when the full dataset is available. Again, the full model is using more than 4x the number of parameters compared to the SSL-trained network with frozen encoder. This effectively highlights the usefulness of using SSL-trained frozen encoder.

TABLE 5 UNet-2.5D parameter comparison.

| Network setting | Number of parameters |
|---|---|
| Full Model | 3,845,058 |
| Decoder Only (Frozen Encoder) | 895,122 |

# 5 Conclusion and recommendations

Nasopharyngeal carcinoma, a cancer common in Asia and Africa, is currently treated with a combination of chemotherapy and radiotherapy. To achieve precise radiation treatments, accurate target delineation is critical but not always easy, especially in the head and neck area where not only the gross tumor volume requires delineation and contouring, but also the lymph node drainage areas and the numerous organs at risk. Target delineation on CT scan images takes time, knowledge and experience. Automatic segmentation can make this task more objective and efficient.

The use of deep learning is a continuously progressing direction in advancing modern medical imaging. This work hopes to be an addition in advancing this goal. Although data scarcity has always been an issue especially in the medical field, we have been able to design and create a deep learning model that is able to perform automatic contouring of gross tumor volume of nasopharyngeal cancer (NPC).

Compared with other architectures, our proposed method is able to significantly outperform other architectures in segmenting NPC. Furthermore, our method is much more efficient as it uses only 2D convolution compared to 3D convolutions used by other architectures.

This highlights that in NPC, 2D convolution is enough and may suggest that across slice information does not only improve performance but degrades it. This may be a result of NPC's structure and topology, in that it forms no regular pattern in its structure but are random and irregular. Hence, the across-slice information adds little information to the model during training.

Moreover, we also leverage a multi-scale training data using five different scales. This allowed us to generate an ensemble of models that is more robust than the individual model. More importantly we have employed the use of semi-supervised learning through the combined rotation and relative-patch-location pre-text tasks to pretrain and freeze an encoder network. This made it 4 times more efficient in terms of the number of parameters required as well as very data efficient. We have shown that even with a portion of labelled data we are able

TABLE 6 Distance metric results of the model ensemble for each data scale highlighting the effectiveness of the model ensemble.

Model ensemble performance with distance metric

| UNet-2.5D-Scale | ASSD (mm) ↓ | Hausdorff (mm) ↓ | Sensitivity (%) ↑ |
|---|---|---|---|
| Extra-Small Scale | 3.87 ± 0.22 | 21.98 ± 1.28 | 67.69 ± 3.93 |
| Small | 4.28 ± 0.19 | 27.70 ± 1.23 | 69.80 ± 3.01 |
| Medium | 3.79 ± 0.21 | 16.73 ± 0.95 | 73.35 ± 4.15 |
| Large | 4.64 ± 0.21 | 22.04 ± 1.03 | 77.74 ± 3.62 |
| Extra-Large | 7.07 ± 0.68 | 21.21 ± 2.05 | 65.72 ± 6.37 |
| Ensemble-Model | 4.22 ± 0.29 | 15.73 ± 0.89 | 71.43 ± 4.20 |

↑ means that higher means better while ↓ symbol means lower is better.

TABLE 7 Model ensemble performance for DSC, IOU, PPV and RVE using a single unified SSL-pretrained encoder hence, effectively training only the decoder block.

Ensemble performance with a frozen SSL-Pretrained encoder

| Data-Scale | DSC (%) ↑ | IOU (%) ↑ | PPV (%) ↑ | RVE (%) ↓ |
|---|---|---|---|---|
| Extra-Small | 71.05 ± 2.32 | 62.28 ± 2.04 | 75.46 ± 2.46 | 16.84 ± 0.55 |
| Small | 71.68 ± 2.55 | 58.56 ± 2.09 | 75.84 ± 2.70 | 15.05 ± 0.54 |
| Medium | 71.76 ± 2.48 | 59.87 ± 2.07 | 72.38 ± 2.51 | 14.29 ± 0.49 |
| Large | 70.97 ± 3.17 | 58.06 ± 2.59 | 67.80 ± 3.03 | 30.32 ± 1.35 |
| Extra-Large | 65.18 ± 4.87 | 53.61 ± 4.01 | 67.64 ± 5.05 | 15.80 ± 1.18 |
| Ensemble Average | 71.16 ± 2.93 | 61.62 ± 2.44 | 75.59 ± 2.95 | 15.37 ± 0.52 |

↑ means that higher means betterwhile ↓ symbol means lower is better.

TABLE 8 Distance metric performance results using a single unified encoder block for multiple decoder for a specific data scale.

Distance metric performance with a Frozen SSL-Pretrained encoder

| UNet-2.5D-Scale | ASSD (mm) ↓ | Hausdorff (mm) ↓ | Sensitivity (%) ↑ |
|---|---|---|---|
| Extra-Small Scale | 3.93 ± 0.13 | 22.36 ± 0.73 | 68.86 ± 2.25 |
| Small | 5.23 ± 0.15 | 28.70 ± 0.98 | 69.36 ± 2.47 |
| Medium | 3.56 ± 0.13 | 15.57 ± 0.57 | 72.64 ± 2.51 |
| Large | 4.14 ± 0.20 | 20.42 ± 0.85 | 78.38 ± 3.47 |
| Extra-Large | 8.47 ± 0.78 | 20.82 ± 1.55 | 64.51 ± 4.82 |
| Ensemble-Average | 6.49 ± 0.37 | 16.22 ± 0.64 | 72.65 ± 3.08 |

↑ means that higher means better while ↓ symbol means lower is better.

TABLE 9 The DSC segmentation performance using the medium-scale dataset for SSL-trained encoder vs. full model.

| % of labelled data | frozen encoder (%) | Full model (%) | Percentage difference |
|---|---|---|---|
| 20% | 64.18 ± 2.35 | 46.29 ± 1.70 | +38.64% |
| 30% | 64.95 ± 2.07 | 49.55 ± 1.58 | +31.08% |
| 50% | 68.89 ± 2.93 | 61.75 ± 2.62 | +11.56% |
| 100% | 71.76 ± 2.48 | 72.47 ± 4.10 | -0.98% |

to reach close performance by a model trained from scratch but using all the training. This has a much greater practical usage in terms of the time and resources needed to collect and annotate the data. Moreover it allows one to exploit and take advantage of the abundant data of healthy patients. We believe for future works, that achieving higher performance with fewer data will gradually become the central focus of researchers as use of medical data tightens.

## Data availability statement

The datasets presented in this article are not readily available because of data security. Requests to access the datasets should be directed to jldomoguen@up.edu.ph.

## Ethics statement

The informed consent was waived given the retrospective nature of this study.

## Author contributions

All authors discussed and conceptualized the study. JD wrote the program and drafted the first manuscript together with JM. All authors analyzed and interpreted the data. PN and JC overviewed the study and guided the preparation of the manuscript. All authors contributed to manuscript revision. All authors read and approved the final manuscript.

## Funding

The authors would like to acknowledge Engineering Research and Development for Technology (ERDT) consortium and Philippine Department of Science and Technology (DOST) for the funding of this work.

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## References

1. Salehiniya H, Mohammadian M, Mohammadian-Hafshejani A, Mahdavifar N. Nasopharyngeal cancer in the world: Epidemiology, incidence, mortality and risk factors. *World Cancer Res J* (2018) 5(1): e1046. doi: 10.32113/wcrj_20183_1046

2. Li S, Xiao J, He L, Peng X, Yuan X. The tumor target segmentation of nasopharyngeal cancer in ct images based on deep learning methods. *Technol Cancer Res Treat* (2019) 18:1533033819884561. doi: 10.1177/1533033819884561

3. Yang G, Dai Z, Zhang Y, Zhu L, Tan J, Chen Z, et al. Multiscale local enhancement deep convolutional networks for the automated 3d segmentation of gross tumor volumes in nasopharyngeal carcinoma: A multi-institutional dataset study. *Front Oncol* (2022) 12:827991–1. doi: 10.3389/fonc.2022.827991

4. Mei H, Lei W, Gu R, Ye S, Sun Z, Zhang S, et al. Automatic segmentation of gross target volume of nasopharynx cancer using ensemble of multiscale deep neural networks with spatial attention. *Neurocomputing* (2021) 438:211–22. doi: 10.1016/j.neucom.2020.06.146

5. Men K, Chen X, Zhang Y, Zhang T, Dai J, Yi J, et al. Deep deconvolutional neural network for target segmentation of nasopharyngeal cancer in planning computed tomography images. *Front Oncol 7* (2017) 7. doi: 10.3389/fonc.2017.00315

6. Ye Y, Cai Z, Huang B, He Y, Zeng P, Zou G, et al. Fully-automated segmentation of nasopharyngeal carcinoma on dual-sequence mri using convolutional neural networks. *Front Oncol* (2020) 10:166. doi: 10.3389/fonc.2020.00166

7. Hu J, Shen L, Sun G. (2018). Squeeze-and-excitation networks, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, Salt Lake City, UT, USA. pp. 7132–41.

8. Bai W, Oktay O, Sinclair M, Suzuki H, Rajchl M, Tarroni G, et al. Semi-supervised Learning for Network-Based Cardiac MR Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2017: 20th International Conference*, September 11-13, 2017, Proceedings, Part II. Springer-Verlag, Berlin, Heidelberg, 253–260. doi: 10.1007/978-3-319-66185-8_29

9. Tseng KK, Zhang R, CM C, Hassan MM. Dnetunet: a semi-supervised cnn of medical image segmentation for super-computing ai service. *J Supercomputing* (2021) 77:3594–615. doi: 10.1007/s11227-020-03407-7

10. Mahapatra D. Semi-supervised learning and graph cuts for consensus based medical image segmentation. *Pattern recognition* (2017) 63:700–9. doi: 10.1016/j.patcog.2016.09.030

11. Bortsova G, Dubost F, Hogeweg L, Katramados I, Bruijne Md. Semi-supervised medical image segmentation via learning consistency under transformations. In: *International conference on medical image computing and computer-assisted intervention*. Springer, Shenzhen, China (2019). p. 810–8.

12. Chen S, Bortsova G, Garc´ıa-Uceda Juarez A, Tulder Gv, Bruijne Md. Multi-task attention-based´ semi-supervised learning for medical image segmentation. In: *International conference on medical image computing and computer-assisted intervention*. Springer, Shenzhen, China (2019). p. 457–65.

13. Wang D, Zhang Y, Zhang K, Wang L. Focalmix: Semi-supervised learning for 3d medical image detection. In: Conference on *Computer Vision and Pattern Recognition* (2020) Institute of Electrical and Electronics Engineers (IEEE), Seattle, Online, USA. p. 3951–60.

14. Cheplygina V, de Bruijne M, Pluim JP. Not-so-supervised: A survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. *Med image Anal* (2019) 54:280–96. doi: 10.1016/j.media.2019.03.009

15. Taleb A, Loetzsch W, Danz N, Severin J, Gaertner T, Bergner B, et al. 3d self-supervised methods for medical imaging. *Adv Neural Inf Process Syst* (2020) 33:18158–72. doi: 10.48550/arXiv.2006.03829

16. Chen T, Kornblith S, Norouzi M, Hinton G. A simple framework for contrastive learning of visual representations. In: *International conference on machine learning*. PMLR, Vienna, Austria (2020). p. 1597–607.

17. Ahn E, Feng D, Kim J. A Spatial Guided Self-supervised Clustering Network for Medical Image Segmentation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021: 24th International Conference*, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I. Springer-Verlag, Berlin, Heidelberg, 379–388. doi: 10.1007/978-3-030-87193-2_36

18. Li H, Xue FF, Chaitanya K, Luo S, Ezhov I, Wiestler B, et al. Imbalance-Aware Self-supervised Learning for 3D Radiomic Representations. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021: 24th International Conference*, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part II. Springer-Verlag, Berlin, Heidelberg, 36–46. doi: 10.1007/978-3-030-87196-3_4

19. Çiçek O, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O. 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016: 19th International Conference*, Athens, Greece, October 17-21, 2016, Proceedings,

Part II. Springer-Verlag, Berlin, Heidelberg, 424–432. doi: 10.1007/978-3-319-46723-8_49

20. Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: *International conference on machine learning*. PMLR, Lille, France (2015). p. 448–56.

21. Xu B, Wang N, Chen T, Li M. Empirical evaluation of rectified activations in convolutional network. *Proceedings of the 32nd International Conference on Machine Learning* (2015) JMLR: W & CP 37. doi: 10.48550/arXiv.1505.00853.

22. Bidgood Jr.WD, Horii SC, Prior FW, Van Syckle DE. Understanding and using dicom, the data interchange standard for biomedical imaging. *J Am Med Inf Assoc* (1997) 4:199–212. doi: 10.1136/jamia.1997.0040199

23. Ma S, Tang J, Guo F. Multi-task deep supervision on attention r2u-net for brain tumor segmentation. *Front Oncol* (2021), 11, 3651. doi: 10.3389/fonc.2021.704850

24. Fechter T, Adebahr S, Baltas D, Ben Ayed I, Desrosiers C, Dolz J. Esophagus segmentation in ct *via* 3d fully convolutional neural network and random walk. *Med Phys* (2017) 44:6341–52. doi: 10.1002/mp.12593

25. Kingma DP, Ba J. Adam: A method for stochastic optimization. *arXiv preprint arXiv* (2014) 1412:6980. doi: 10.48550/arXiv.1412.6980

26. Zhou Z, Sodha V, Rahman Siddiquee MM, Feng R, Tajbakhsh N, Gotway MB, et al. Models genesis: Generic autodidactic models for 3d medical image analysis. In: *International conference on medical image computing and computer-assisted intervention*. 13-17 October 2019 Springer, Shenzhen, China (2019). p. 384–93.

27. Isensee F, Jaeger PF, Kohl SA, Petersen J, Maier-Hein KH. Nnu-net: A self-configuring method for deep learning-based biomedical image segmentation. *Nat Methods* (2021) 18:203–11. doi: 10.1038/s41592-020-01008-z

TYPE Original Research
PUBLISHED 24 November 2022
DOI 10.3389/fonc.2022.1028382

# An attention base U-net for parotid tumor autosegmentation

Xianwu Xia[1,2,3,4], Jiazhou Wang[3,4], Sheng Liang[2], Fangfang Ye[2], Min-Ming Tian[5], Weigang Hu[3,4*] and Leiming Xu[1*]

[1]The Second Affiliated Hospital, School of Medicine, Zhejiang University, Hangzhou, Zhejiang, China, [2]Department of Oncology Intervention, The Affiliated Municipal Hospital of Taizhou University, Taizhou, China, [3]Department of Radiation Oncology, Fudan University Shanghai Cancer Center, Shanghai, China, [4]Department of Oncology, Shanghai Medical College, Fudan University, Shanghai, China, [5]Department of Oncology Intervention, Jiangxi University of Traditional Chinese Medicine, Nanchang, Jiangxi, China

A parotid neoplasm is an uncommon condition that only accounts for less than 3% of all head and neck cancers, and they make up less than 0.3% of all new cancers diagnosed annually. Due to their nonspecific imaging features and heterogeneous nature, accurate preoperative diagnosis remains a challenge. Automatic parotid tumor segmentation may help physicians evaluate these tumors. Two hundred eighty-five patients diagnosed with benign or malignant parotid tumors were enrolled in this study. Parotid and tumor tissues were segmented by 3 radiologists on T1-weighted (T1w), T2-weighted (T2w) and T1-weighted contrast-enhanced (T1wC) MR images. These images were randomly divided into two datasets, including a training dataset (90%) and an validation dataset (10%). A 10-fold cross-validation was performed to assess the performance. An attention base U-net for parotid tumor autosegmentation was created on the MRI T1w, T2 and T1wC images. The results were evaluated in a separate dataset, and the mean Dice similarity coefficient (DICE) for both parotids was 0.88. The mean DICE for left and right tumors was 0.85 and 0.86, respectively. These results indicate that the performance of this model corresponds with the radiologist's manual segmentation. In conclusion, an attention base U-net for parotid tumor autosegmentation may assist physicians to evaluate parotid gland tumors.

KEYWORDS

parotid, auto-segmentation, artificial intelligence (AI), neoplasms, diagnosis

OPEN ACCESS

EDITED BY
Michele Avanzo,
Aviano Oncology Reference Center (IRCCS), Italy

REVIEWED BY
Kuo Men,
Chinese Academy of Medical Sciences and Peking Union Medical College, China
David G. Ellis,
University of Nebraska Medical Center, United States

*CORRESPONDENCE
Weigang Hu
jackhuwg@gmail.com
Leiming Xu
mdxu@zju.edu.cn

SPECIALTY SECTION
This article was submitted to Radiation Oncology, a section of the journal Frontiers in Oncology

RECEIVED 25 August 2022
ACCEPTED 26 October 2022
PUBLISHED 24 November 2022

CITATION
Xia X, Wang J, Liang S, Ye F, Tian M-M, Hu W and Xu L (2022) An attention base U-net for parotid tumor autosegmentation.
*Front. Oncol.* 12:1028382.
doi: 10.3389/fonc.2022.1028382

## Introduction

Parotid tumors are uncommon neoplasms, accounting for less than 3% of all head and neck cancers (1). Unfortunately, a lack of early detection may lead to tumor progression, and nearly 20% of untreated polymorphic adenomas will become malignant tumors (2). In addition, 80% of salivary gland tumors occur in the parotid gland, of which 21% to 64% are malignant (3). Due to the absence of specific imaging findings (parotid tumor may have different appearance in MR images), their heterogeneous clinical nature, accurate diagnosis before surgery remains a challenge (4).

Similar to lung nodule detection, automatic parotid tumor segmentation may facilitate physicians evaluating these parotid tumors. It can be used to inspect the MRI image and highlight the tumor region. At the same time, with the progress of quantitative image analysis technology, we can construct a quantitative imaging model of parotid gland tumors through accurate and consistent automatic segmentation of tumors, which can be used to predict the pathological type and prognosis of the patients (5).

In this study, we developed and assessed an autosegmentation model for parotid tumors that can be used to improve the imaging evaluation of these conditions. This proposed model was also compared to other model architectures. Since we combined three MRI sequences, the value of each MRI sequence was investigated.

# Methods

The study workflow is presented in Figure 1. Patient parotid MR images were exported from PACS. Parotid and tumor tissues were segmented by 3 radiologists based on T1-weighted (T1w), T2-weighted (T2w) and T1-weighted contrast-enhanced (T1wC) MR images. A 10-fold cross-validation was performed to assess the segmentation performance. These images were randomly divided into two datasets, including a training dataset (90%) and an validation dataset (10%). The autosegmentation model was trained on the training dataset, and its performance was then tested on the validation dataset. This retrospective study was approved by the Institutional Review Board of Fudan University Shanghai Cancer Center and Taizhou Municipal Hospital, and all methods were performed in accordance with the guidelines and regulations of this ethics board. The Hospital Ethics Committee agreed to the informed consent waiver.

## Patients and MRI image acquisition

Two hundred eighty-five patients diagnosed with benign or malignant parotid tumors from two institutions were enrolled in this study. Among these patients, 185 were male and 100 were female; the mean age of the patients was 52.4 years (range, 21–93 years). These patients were treated from 2014 to 2018. All patients received surgical resection and had a pathology report. The patient characteristics are shown in Table 1. All patients received a parotid site MRI scan before treatment. Three MR scanners were used to acquire these images, and details of the image parameters are shown in Table 2. The scan parameters were based on our parotid image protocol and were adjusted during scanning based on image quality by the MRI operator.

## Tumor and parotid manual delineation

Parotid tumors were distinguished on axial thin-Section T1w, T2w and T1wC MR images and segmented by three experienced radiologists (>5 years of experience) in MIM (version 6.8.10, Cleveland, US). These three series were registered and fused before segmentation. The radiologists were required to distinguish the pathology type of the parotid tumor before delineation. Each radiologist segmented approximately 90 patients. To make the delineation between different radiologists consistent, all delineations were reviewed by one senior radiologist (more than 10 years' experience). To improve the performance of the tumor delineation, the parotids were also segmented.

## The attention U-net

A 2D U-Net with an attention module was used in this task. This network was inspired by the application of an attention mechanism to medical image deep learning-based segmentation (6–8). The basic structure of the model is shown in Figure 2. The input (512 x 512 x 3) was obtained from MR images. The channels were combined from the T1w, T2w and T1wC sequences. The output (512 x 512 x 4) contained 4 channels for 4 ROIs, including the left parotid, right parotid, left tumor and right tumor. The U-net was constituted by encoder and decoder parts. The encoder part was constituted by 12 convolution blocks and 4 max pooling blocks. The convolution block had a 3x3 convolution layer, batch normalization layer and rectified linear unit (ReLU) layer. The maximum pooling layer was used to downsample the features. Similarly, the decoder part was constituted by many convolution blocks and upsampling blocks. The convolution block was the same as the encoder part, using a 3x3 convolution layer, batch normalization layer and rectified linear unit (ReLU) layer. The skip connection was used to connect the encoder and decoder parts with the same feature map size. An attention gate was placed in these skip connections to improve the segmentation results. Because the slices thickness (4~7.2 mm) was larger than the pixel size (0.4~1mm), MR images were not be resampled to isotropy resolution. And

The tumor and parotid tissues were relatively small compared to the entire image size. The attention mechanism was used to create a model focused on local regions that extracted more relevant features from the feature maps. A mask with pixel values between 0 and 1 was generated by a sigmoid activation function. By multiplying the mask by feature maps, the region of interest remained unchanged, and the rest of the feature map was set to zero.

## Model training

Before input into the model, the gray value of the MR images was centralized to 0.5 and scaled to [0, 1]. No spatial resampling was performed in the preprocessing stage. We used the original pixels,

**Abbreviations:** DICE, Dice similarity coefficient.

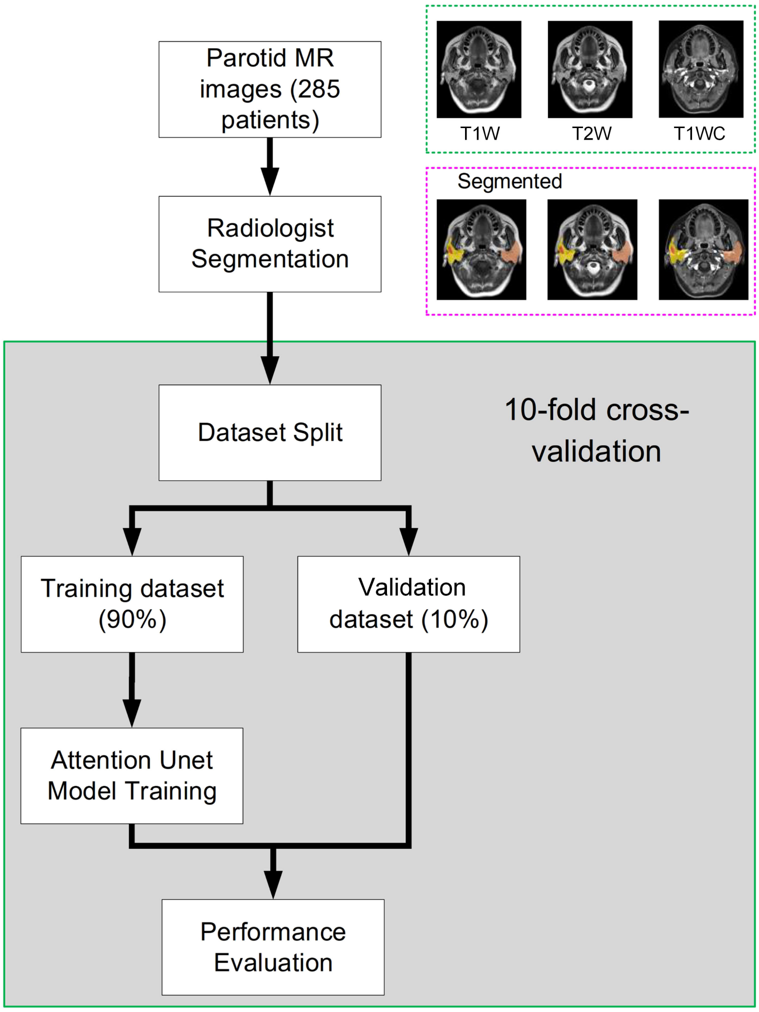


FIGURE 1
The whole study workflow. The parotid MR images were randomly divided into two datasets, including training and evaluation. Then, the performance was assessed on the validation dataset. Tenfold cross-validation was used to obtain a reliable result.

TABLE 1 Patient characteristics.

| | | Characteristics |
|---|---|---|
| Age | | 52.4 (21~93) years |
| Sex | Male | 185 (65%) |
| | Female | 100 (35%) |
| Pathology Type | Warthin tumor | 62 (21.5%) |
| | Pleomorphic adenoma | 90 (31.4%) |
| | Adenocarcinoma | 80 (28.0%) |
| | Basal cell adenoma | 6 (2.0%) |
| | Lymphoma | 30 (10.1%) |
| | Others | 20 (7.0%) |
| Site | Left | 127 (44.6%) |
| | Right | 140 (49.1%) |
| | Both | 18 (6.3%) |

which means that different patients may have different pixel spacings. The loss used in this phase was 1- DICE index. The whole model was trained for 200 epochs with a learning rate of 1e-4, and the optimizer was RMSprop. The training procedure took approximately 20 h to complete on one 2080 ti GPU (Nvidia, Santa Clara, CA). The Python deep learning library pytoch (version 1.5) was applied to establish this autosegmentation system.

Next, a data augmentation method was performed. Two argumentation processes were implemented: gray level disturbance and shape disturbance. For gray disturbance, the gray value of the MR image was multiplied by a random number [0.9~1.1], and a random number [-0.1~0.1] was added. This random number was added to the normalized image. For shape disturbance, MR images and binary contour images were deformed using affine transformation. The augmentation method was the same as that in our previous study (9).

TABLE 2 MR scan parameters.

| | | Signa HDxt (GE) | Verio (SIEMENS) | Skyra (SIEMENS) |
|---|---|---|---|---|
| Patients | | 218 (76.5%) | 34 (11.9%) | 33 (11.6%) |
| T1-weighted | TR (Repetition Time) | 280~540 ms | 450~620 ms | 250~1560 ms |
| | TE (Echo Time) | 8.5~10.4 ms | 12~16 ms | 2.5~12 ms |
| T2-weighted | TR (Repetition Time) | 2740~3600 ms | 2500~5240 ms | 2500~5790 ms |
| | TE (Echo Time) | 84~88 ms | 78~91 ms | 78~83 ms |
| T1-weighted contrast enhanced | TR (Repetition Time) | 175~280 ms | 4.1~6.0 ms | 3.7~6.0 ms |
| | TE (Echo Time) | 1.8~3.4 ms | 1.5~2.5 ms | 1.4~2.4 ms |
| | Contrast Agent | Gadopentetic acid | Gadopentetic acid | Gadopentetic acid |
| Slice Thickness | | 5~7 mm | 4.5~7.2 mm | 4.0~6.0 mm |
| Pixel size | | 0.4~0.6 mm | 0.65~0.97 mm | 0.4~0.85 mm |

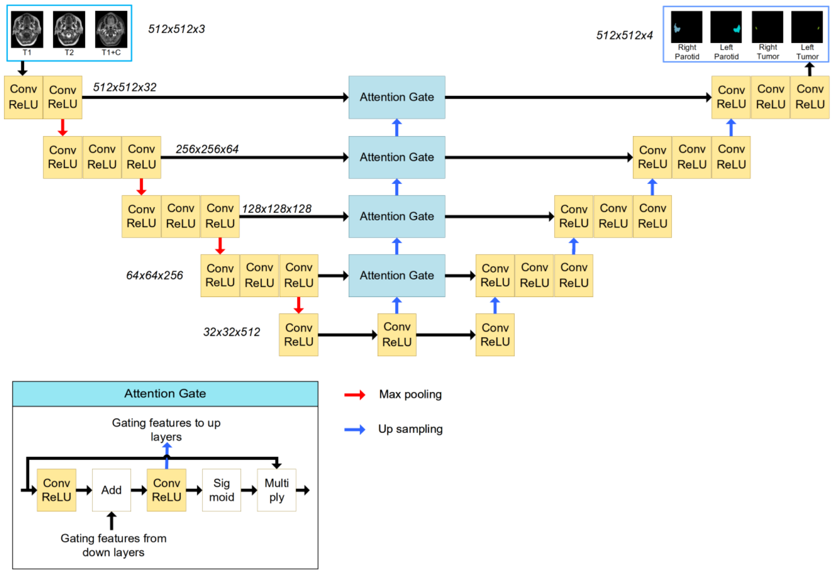


FIGURE 2
The structure of the attention-based U-Net. The input of the network is three MR images, and the output of the network is the four segmentations. The attention gate structure is shown in the left corner.

Meanwhile, to increase the training samples, we mirrored images (and adjusted for the corresponding left and right labels) with a probability of 0.5.

To investigate the impact of each MRI sequence, 6 models with different image sequence combinations were trained and evaluated, including T1w only, T2w only, T1wC only, T1w+T2w, T1w+T1wC and T2w+T1wC.

## Comparison to other models

Three other models, including DeepLab Version 3 (10), attention U-Net (11) and PSPNet (12), were trained on the same dataset. Some modifications were performed, such as changing the output channels and changing the softmax function to a sigmoid function. The same training hyperparameters were used, and all models converged after 200 epoch iterations.

## Performance evaluation

Four indices were calculated for performance evaluation, including the Dice similarity coefficient (DICE), the Jaccard similarity coefficient (JACCARD), the 95% percentile of Hausdorff distance (HD95) and the average Hausdorff distance (AHD). The DICE and JACCARD are computed by the following:

$$DICE = 2|A \cap B|/(|A| + |B|) \tag{1}$$

$$JACCARD = (|A \cap B|)/(|A \cup B|) \tag{2}$$

where A represents the volume of the manual segmentation, B represents the volume of the autosegmentation, | · | denotes the volume of truth or predicted ROIs, |A∩B| indicates the volume shared by A and B and |A∪B| represents the total volume of A and B. Larger DICE and JACCARD values indicate more accurate results.

# Results

## Segmentation results

A 10-fold cross-validation was used in this study. A total of 256 (90%) patients were used for model training, and 29 (10%) patients were used for model evaluation and performance assessment. Training was converged after 200 epoch iterations. The results of the validation dataset are shown in Figure 3. It can be observed that the performance of the validation dataset has a relatively large variation.

For the results of the cross-validation, the mean DICE for both parotids was 0.88, and the mean DICE for left and right tumors was 0.85 and 0.86, respectively. The mean JACCARD for left and right parotids was 0.79. The mean JACCARD for left and right tumors was 0.78 and 0.80, respectively. The 95% ranges for left and right parotid DICE were 0.77-0.94 and 0.75-0.95, respectively. The 95% ranges for left and right tumor DICE were 0.37-1.00 and 0.30-1.00, respectively. Detailed values of these results are provided in Supplementary Table S1. Figure 4 demonstrates a result on a left parotid tumor patient.

## Comparison to other models

The performance of three other models, including DeepLab Version 3 (10), attention U-Net (11) and PSPNet (12), is presented in Table 3. Since all of the models were trained on the same training dataset, this comparison provides insight into the performance of the proposed model.

## The impact of MRI sequences

The performance of models with different MRI sequences is presented in Table 4. For parotid gland segmentation, one MRI

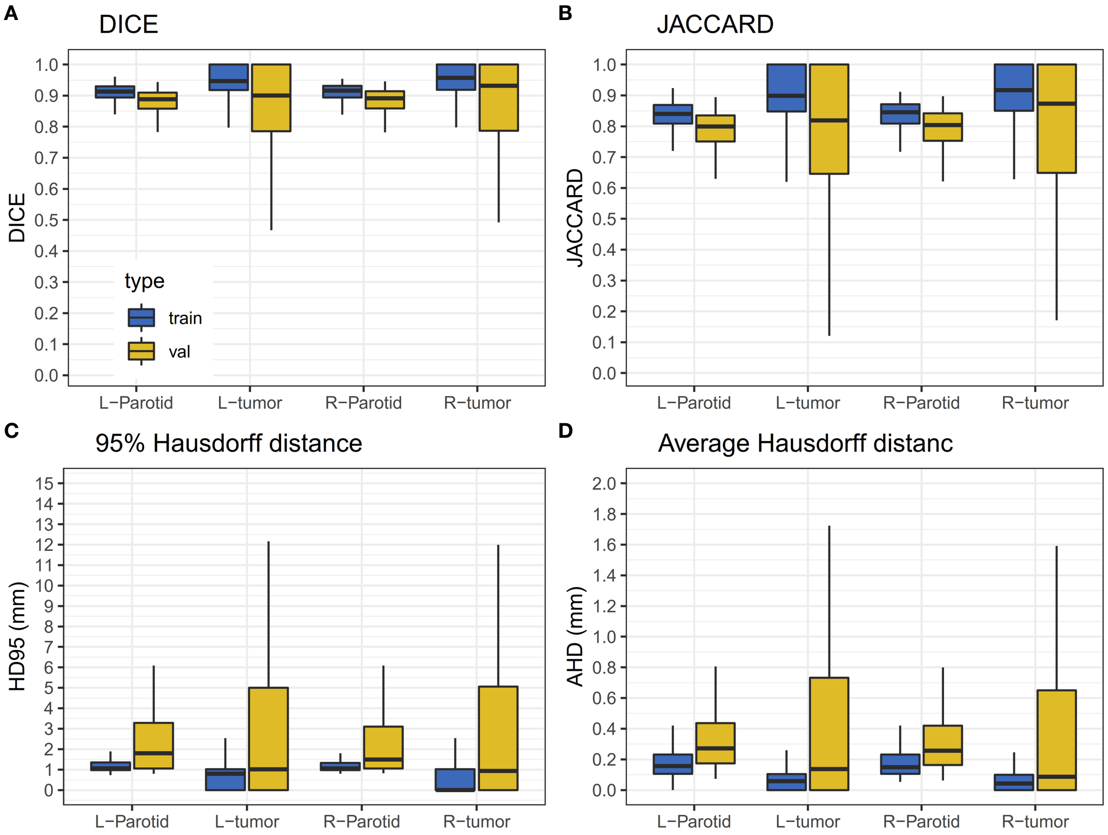


FIGURE 3
Results of the validation dataset. The horizontal lines indicate the median values. **(A)** The DICE value for the training and validation dataset. **(B)** JACCARD value for the training and validation datasets. **(C)** The HD95 value for the training and validation datasets. **(D)** The AHD value for the training and validation datasets.

sequence can achieve segmentation performance similar to that of a combination of three MRI sequences. However, for tumor segmentation, combining three image sequences can provide the best performance. Among the three MRI sequences, T1w performed better than the other two.

## Discussion

In this study, we implemented an attention base U-net for parotid tumor autosegmentation on MRI T1w, T2w and T1wC images. For a rare tumor, the entire dataset was relatively large, including 285 patients, and multiple MRI scanners were used for image acquisition. All whole images were acquired over the course of 4 years with many adjustments to the scan parameters. We believe these images are representative of most parotid tumor MRI scenarios.

An attention mechanism was applied to optimize the extracted spatial information of the feature maps in our study (13). Here, we used a mask with pixel values between 0 and 1 that was generated by transformation, and then feature maps were multiplied by the mask. The region of interest remained unchanged, and the rest of the feature map was set to zero because the regions of the parotid and tumor tissues were relatively small compared to the other organs. This will facilitate model training to focus on critical regions and provide improved results. Compared to the original attention U-Net, our proposed model extracts the gate feature from the bottom of the network. This architecture may help the network focus consistently on only a small region. For hyper-parameters tuning, the major parameters were learning rate. We have use 3 different learn rate (1e-2, 1e-3 and 1e-4), the results showed that 1e-4 can provide the stable results (Figure S1).

There are some differences in the difficulty of organs and tumors delineating. Organ delineation is a relatively simple task. Compare to other's study, our research on the performance of parotid gland segmentation is similar (DICE = 0.88) (14, 15). Few studies have reported using MR imaging for parotid gland autosegmentation. Kieselmann et al. performed atlas-based autosegmentation for parotids (14). The DICE values for Kieselmann's study were 0.83 and 0.84 for the left and right parotid, respectively. Nuo et al. used deep learning technology on a low-field MR segment of the parotid gland and found that the best performance was 0.85 (15). Compared

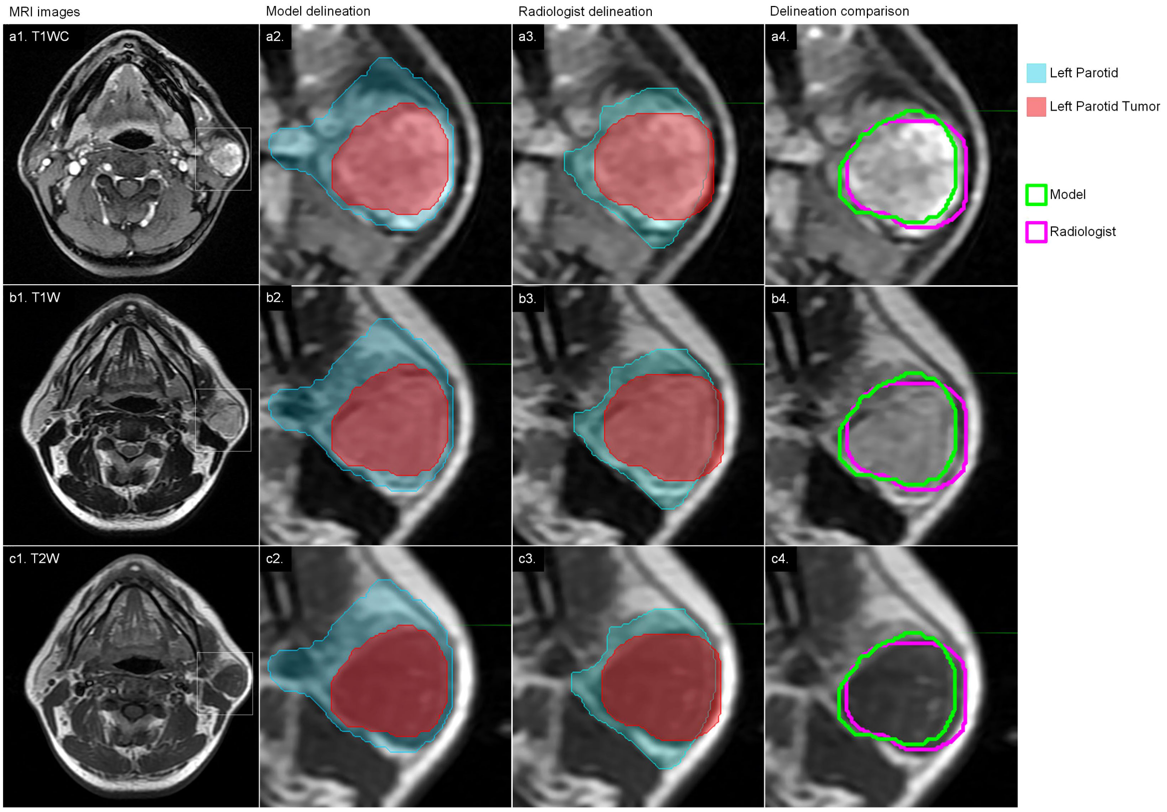


FIGURE 4
An example of the results. a1, b1 and c1 represent one slice of the MR images; a2, b2 and c2 represent the results of autosegmentation; a3, b3 and c3 represent the results of manual segmentation; a4, b4 and c4 show the comparison of the tumor segmentation.

TABLE 3 The comparison with other models.

| Model | Right parotid | | Left parotid | | Right tumor | | Left tumor | |
|---|---|---|---|---|---|---|---|---|
| | DICE | 95% CI | DICE | 95% CI | DICE | 95% CI | DICE | 95% CI |
| DeepLab V3 | 0.87 | [0.65-0.98] | 0.85 | [0.63-0.95] | 0.77 | [0.51-0.95] | 0.83 | [0.45-0.93] |
| Attention U-Net | 0.88 | [0.73-0.96] | 0.86 | [0.71-0.94] | 0.84 | [0.35-1.00] | 0.81 | [0.45-1.00] |
| PSPNet | 0.87 | [0.72-0.90] | 0.85 | [0.78-0.89] | 0.78 | [0.25-1.00] | 0.85 | [0.38-1.00] |
| Proposed Model | 0.88 | [0.75-0.95] | 0.88 | [0.77-0.94] | 0.85 | [0.30-1.00] | 0.86 | [0.37-1.00] |

CI, confidence interval.

TABLE 4 The comparison between different MRI sequences.

| MRI Sequences | Right parotid | | Left parotid | | Right tumor | | Left tumor | |
|---|---|---|---|---|---|---|---|---|
| | DICE | 95% CI | DICE | 95% CI | DICE | 95% CI | DICE | 95% CI |
| T1w | 0.88 | [0.74-0.95] | 0.86 | [0.74-0.92] | 0.82 | [0.30-1.00] | 0.81 | [0.37-1.00] |
| T1wC | 0.84 | [0.73-0.92] | 0.82 | [0.69-0.90] | 0.71 | [0.14-1.00] | 0.73 | [0.12-1.00] |
| T2w | 0.88 | [0.74-0.95] | 0.88 | [0.72-0.93] | 0.81 | [0.30-1.00] | 0.79 | [0.31-1.00] |
| T1w+T1wC | 0.88 | [0.75-0.94] | 0.85 | [0.75-0.92] | 0.78 | [0.30-1.00] | 0.84 | [0.53-1.00] |
| T1w+T2w | 0.88 | [0.74-0.95] | 0.87 | [0.75-0.93] | 0.84 | [0.30-1.00] | 0.83 | [0.43-0.94] |
| T1wC+T2w | 0.88 | [0.75-0.93] | 0.85 | [0.76-0.94] | 0.75 | [0.20-1.00] | 0.78 | [0.30-0.95] |
| T1w+T1wC+T2w Proposed | 0.88 | [0.75-0.95] | 0.88 | [0.77-0.94] | 0.85 | [0.30-1.00] | 0.86 | [0.37-1.00] |

CI, confidence interval.

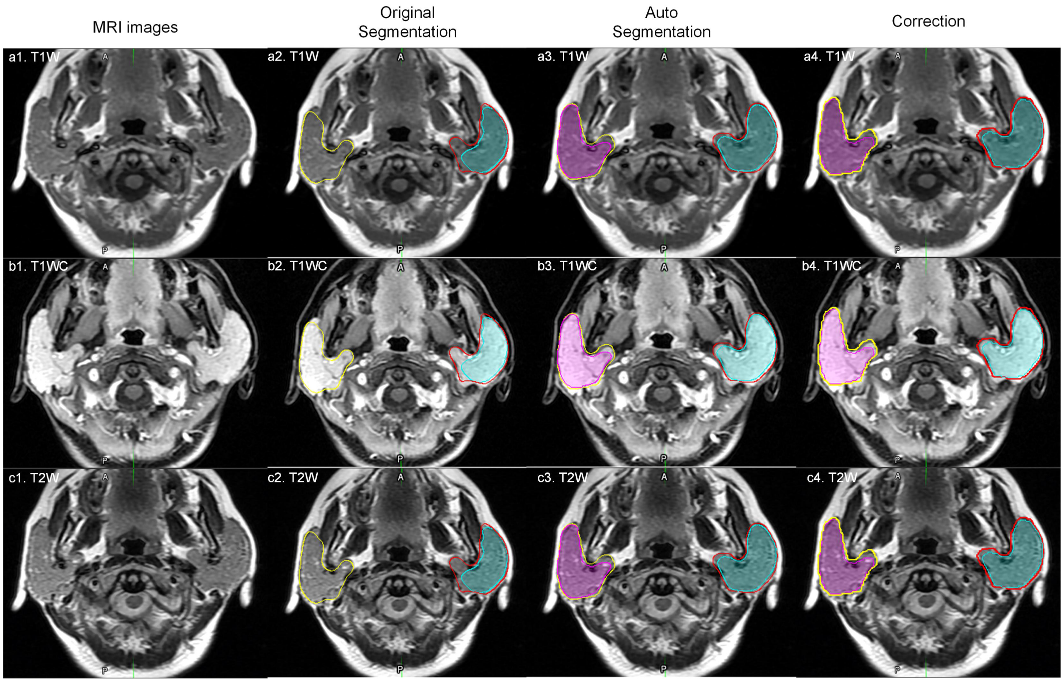


FIGURE 5
An outlier example. The yellow and red lines represent the right and left parotid. The pink and cyan colored filling represents the right and left tumor. a1, b1 and c1 represent one slice of MR images; a2, b2 and c2 represent original manual segmentation. The right tumor was not delineated correctly; a3, b3 and c3 represent the results of autosegmentation; a4, b4 and c4 represent the corrected segmentation by manual delineation by physicians.

with these studies, our data were delineated by radiologists with the same protocol on both the training and validation data. The data consistency was relatively good.

Parotid tumor delineation is a relatively difficult task. The main problem is the lack of training samples and the lack of consistent delineation standards (16). Parotid tumor delineation is challenging in medical image segmentation due to the infrequency of this disease, which physician may not have enough experience to precisely delineate the tumor. Even after carefully reviewed the manual segmentation, there still exist some uncertainty in the manual delineation. Figure 5 shows a patient with a DICE of 0.127 for a right tumor. After carefully checking the data and reviewing this patient's history, we found that the delineation in training dataset only segmented part of the tumor, while this patient exhibited a bilateral diffuse MALT (mucosa-associated lymphoid tissue, mucosa-associated lymphoid tissue) lesion. Given this, our model correctly marked the entire tumor, and in this case, the tumor comprised nearly the entire parotid.

There is an overfitting between training and validation. We believe this degree of overfitting is acceptable. While the deviation of performance between different patients still large. For example, the 95% CI of DICE was [0.30-1.00] for of right tumor. This phenomenon indicates that training sample may too small to cover different types of parotid tumors. And the training dataset also may have some uncertainty in delineation.

For the clinical application, because the parotid cancer is a rare cancer, physicians may not have enough experience to assess tumor-infiltrating area. Tumor autosegmentaion may help physicians to do this. Further researches may require to demonstrate the benefit of this model.

There are some limitations to this study. First, we did not validate our model on an external dataset, which might be valuable for providing reliability information. However, because there were 3 MR scanners were used to acquire these images, and the parameters of image protocol were changed during 4 years, using cross validation can precisely estimate the model performance. Second, we combined three images, T1w, T2w and T1wC. For routine diagnostic purposes, some of these images may not be acquired, and a model accounting for missing data may need to be developed in the future.

## Conclusion

An attention base U-net for parotid tumor autosegmentation may assist physicians to evaluate parotid gland tumors.

## Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material. Further inquiries can be directed to the corresponding authors.

## Ethics statement

The studies involving human participants were reviewed and approved by Institutional Review Board of Fudan University Shanghai Cancer Center and Taizhou Municipal Hospital. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

## Author contributions

JW, XX, LX, and WH designed study. Data collection: SL, FY, and M-MT. Image segmentation: SL and FY. Data analysis and interpretation XX and JW. Manuscript written – all authors contributed. All authors read and approved final manuscript.

## Funding

The work was supported by the Zhejiang Provincial Health Science and Technology Project. Grant number: 2021KY396.

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fonc.2022.1028382/full#supplementary-material

## References

1. Lewis AG, Tong T, Maghami E. Diagnosis and management of malignant salivary gland tumors of the parotid gland. *Otolaryngol Clin North Am* (2016) 49:343–80. doi: 10.1016/j.otc.2015.11.001

2. Zhou N, Chu C, Dou X, Li M, Liu S, Zhu L, et al. Early evaluation of irradiated parotid glands with intravoxel incoherent motion MR imaging: correlation with dynamic contrast-enhanced MR imaging. *BMC Cancer* (2016) 16:865. doi: 10.1186/s12885-016-2900-2

3. Tao X, Yang G, Wang P, Wu Y, Zhu W, Shi H, et al. The value of combining conventional, diffusion-weighted and dynamic contrast-enhanced MR imaging for the diagnosis of parotid gland tumours. *Dentomaxillofac Radiol* (2017) 46 (6):20160434. doi: 10.1259/dmfr.20160434

4. Sentani K, Ogawa I, Ozasa K, Sadakane A, Utada M, Tsuya T, et al. Characteristics of 5015 salivary gland neoplasms registered in the Hiroshima tumor tissue registry over a period of 39 years. *J Clin Med* (2019) 8(5):566. doi: 10.3390/jcm8050566

5. Zheng YM, Li J, Liu S, Cui JF, Zhan JF, Pang J, et al. MRI-Based radiomics nomogram for differentiation of benign and malignant lesions of the parotid gland. *Eur Radiol* (2021) 31(6):4042–52. doi: 10.1007/s00330-020-07483-4

6. Dolz J, Gopinath K, Yuan J, Lombaert H, Desrosiers C, Ben Ayed I. HyperDense-net: A hyper-densely connected CNN for multi-modal image segmentation. *IEEE Trans Med Imaging* (2019) 38(5):1116–26. doi: 10.1109/TMI.2018.2878669

7. Mishra D, Chaudhury S, Sarkar M, Soin AS. Ultrasound image segmentation: A deeply supervised network with attention to boundaries. *IEEE Trans BioMed Eng* (2019) 66:1637–48. doi: 10.1109/TBME.2018.2877577

8. Ronneberger O, Fischer P, Brox T. *Medical image computing and computer-assisted intervention – MICCAI 2015*. Navab N, Hornegger J, Wells WM, Frangi AF, editors. Springer International Publishing p. 234–41.

9. Wang J, Lu J, Qin G, Shen L, Sun Y, Ying H, et al. Technical note: A deep learning-based autosegmentation of rectal tumors in MR images. *Med Phys* (2018) 45(6):2560–4. doi: 10.1002/mp.12918

10. Chen L-C, Papandreou G, Schroff F, Adam H. Rethinking atrous convolution for semantic image segmentation. In: *arXiv:1706.05587* (2017). Available at: https://ui.adsabs.harvard.edu/abs/2017arXiv170605587C.

11. Oktay O, Schlemper J, Le Folgoc L, Lee M, Heinrich M, Misawa K, et al. Attention U-net: Learning where to look for the pancreas. In: *arXiv:1804.03999* (2018). Available at: https://ui.adsabs.harvard.edu/abs/2018arXiv180403999O.

12. Zhao H, Shi J, Qi X, Wang X, Jia J. Proceedings of the IEEE conference on computer vision and pattern recognition In *2017 IEEE Conference onComputer Vision and Pattern Recognition (CVPR)*. pp. 2881–90.

13. Chen L-C, Yang Y, Wang J, Xu W, Yuille AL. Proceedings of the IEEE conference on computer vision and pattern recognition In *2017 IEEE Conference onComputer Vision and Pattern Recognition (CVPR)*. pp. 3640–9.

14. Kieselmann JP, Kamerling CP, Burgos N, Menten MJ, Fuller CD, Nill S, et al. Geometric and dosimetric evaluations of atlas-based segmentation methods of MR images in the head and neck region. *Phys Med Biol* (2018) 63(14):145007. doi: 10.1088/1361-6560/aacb65

15. Tong N, Gou S, Yang S, Cao M, Sheng K. Shape constrained fully convolutional DenseNet with adversarial training for multiorgan segmentation on head and neck CT and low-field MR images. *Med Phys* (2019) 46:2669–82. doi: 10.1002/mp.13553

16. Balagopal A, Morgan H, Dohopolski M, Timmerman R, Shan J, Heitjan DF, et al. PSA-net: Deep learning-based physician style-aware segmentation network for postoperative prostate cancer clinical target volumes. *Artif Intell Med* (2021) 121:102195. doi: 10.1016/j.artmed.2021.102195

TYPE Original Research
PUBLISHED 10 February 2023
DOI 10.3389/fonc.2023.1130119

# Leveraging intelligent optimization for automated, cardiac-sparing accelerated partial breast treatment planning

Joel A. Pogue*, Carlos E. Cardenas, Yanan Cao, Richard A. Popple, Michael Soike, Drexell Hunter Boggs, Dennis N. Stanley and Joseph Harms

Department of Radiation Oncology, University of Alabama at Birmingham, Birmingham, AL, United States

**Background:** Accelerated partial breast irradiation (APBI) yields similar rates of recurrence and cosmetic outcomes as compared to whole breast radiation therapy (RT) when patients and treatment techniques are appropriately selected. APBI combined with stereotactic body radiation therapy (SBRT) is a promising technique for precisely delivering high levels of radiation while avoiding uninvolved breast tissue. Here we investigate the feasibility of automatically generating high quality APBI plans in the Ethos adaptive workspace with a specific emphasis on sparing the heart.

**Methods:** Nine patients (10 target volumes) were utilized to iteratively tune an Ethos APBI planning template for automatic plan generation. Twenty patients previously treated on a TrueBeam Edge accelerator were then automatically replanned using this template without manual intervention or reoptimization. The unbiased validation cohort Ethos plans were benchmarked *via* adherence to planning objectives, a comparison of DVH and quality indices against the clinical Edge plans, and qualitative reviews by two board-certified radiation oncologists.

**Results:** 85% (17/20) of automated validation cohort plans met all planning objectives; three plans did not achieve the contralateral lung V1.5Gy objective, but all other objectives were achieved. Compared to the Eclipse generated plans, the proposed Ethos template generated plans with greater evaluation planning target volume (PTV_Eval) V100% coverage ($p$ = 0.01), significantly decreased heart V1.5Gy ($p$< 0.001), and increased contralateral breast V5Gy, skin D0.01cc, and RTOG conformity index ($p$ = 0.03, $p$ = 0.03, and $p$ = 0.01, respectively). However, only the reduction in heart dose was significant after correcting for multiple testing. Physicist-selected plans were deemed clinically acceptable without modification for 75% and 90% of plans by physicians A and B, respectively. Physicians A and B scored at least one automatically generated plan as clinically acceptable for 100% and 95% of planning intents, respectively.

Check for updates

OPEN ACCESS

EDITED BY
Jiahan Zhang,
Emory University, United States

REVIEWED BY
Yibo Xie,
Winship Cancer Institute, Emory University, United States
Fu Jin,
Chongqing University, China

*CORRESPONDENCE
Joel A. Pogue
japogue@uabmc.edu

SPECIALTY SECTION
This article was submitted to Radiation Oncology, a section of the journal Frontiers in Oncology

RECEIVED 22 December 2022
ACCEPTED 02 February 2023
PUBLISHED 10 February 2023

CITATION
Pogue JA, Cardenas CE, Cao Y, Popple RA, Soike M, Boggs DH, Stanley DN and Harms J (2023) Leveraging intelligent optimization for automated, cardiac-sparing accelerated partial breast treatment planning.
*Front. Oncol.* 13:1130119.
doi: 10.3389/fonc.2023.1130119

**Conclusions:** Standard left- and right-sided planning templates automatically generated APBI plans of comparable quality to manually generated plans treated on a stereotactic linear accelerator, with a significant reduction in heart dose compared to Eclipse generated plans. The methods presented in this work elucidate an approach for generating automated, cardiac-sparing APBI treatment plans for daily adaptive RT with high efficiency.

KEYWORDS

**APBI, Ethos, automated planning, cardiac-sparing breast radiotherapy, intelligent optimization, adaptive radiotherapy (ART)**

## 1 Introduction

The incidence rate of early stage breast cancer is steadily increasing due to improved detection and screening strategies (1). Equivalent overall survival rates of lumpectomy followed by external beam radiation therapy (RT) compared to mastectomy have been shown (2), and post-lumpectomy pathologic analysis by Vicini et al. demonstrated that residual disease occurred within 1cm of the lumpectomy cavity for more than 90% of patients (3). Until recently, external beam accelerated partial breast irradiation (APBI) has been less preferred to brachytherapy APBI due to the large planning target volume (PTV) margins necessary to account for set-up uncertainty, resulting in increased healthy tissue exposure and inferior cosmetic outcomes relative to whole breast RT (4). However, technical improvements in patient immobilization, imaging, and dosimetry have more recently piqued interest in stereotactic body radiation therapy (SBRT), which allows for reduced margins and steeper dose fall-off outside of the target.

To that end, Vermeulen et al. observed no toxicities ≥ grade 3 for 46 stage 1 patients receiving supine SBRT treatment with a 2mm PTV expansion (5, 6). Additionally, Timmerman et al. published methods and cosmetic outcomes for a 75 patient, five arm dose-escalation SBRT trial in which high rates of good or excellent cosmesis were achieved (7, 8). Livi et al. demonstrated that compared to conventionally fractionated (50Gy in 25 fractions) breast treatment, intensity modulated radiation therapy (IMRT) based PBI resulted in significantly fewer short and long term toxicities and improved cosmetic satisfaction compared to whole breast RT using 1cm PTV margins (9). Based on these findings, our institution initiated the UAB RAD 1802 trial (Pilot Trial of LINAC Based Stereotactic Body Radiotherapy for Early Stage Breast Cancer Patients Eligible for Post-Operative Accelerated Partial Breast Irradiation (APBI); clinicaltrials.gov identifier NCT03643861). The purpose was to combine the SBRT techniques and accelerated fractionation schemes, which were previously exclusively utilized on the Cyberknife platform, with the IMRT capabilities of a traditional linear accelerator. Methods and preliminary findings for the first 23 patients (16 prone, 7 supine) have since been published (10).

While novel platforms such as RapidPlan and HyperArc (Varian Medical Systems) have provided a means of automating planning processes (11, 12), many institutions still heavily rely on iterative, manual planning (13, 14). Developing alternatives to manual planning would be ideal as the time required to train personnel and manually generate high-quality treatment plans remains costly (15). Furthermore, planning skill varies greatly by planner and site (16, 17), manual plan constraints and optimization structures are often inconsistent, and time limitations greatly impact the quality of manual plans. Thus, the aim of automation is to increase plan consistency, reduce planning time, and maintain or improve plan quality. Popular forms of automation include, but are not limited to, knowledge-based planning (KBP) (18–22), multi-criteria optimization (MCO) (23, 24), and template-based planning (25–28).

There have been few studies showing effective automated planning implementation for the Ethos system. The Ethos (Varian Medical Systems, Palo Alto, CA) is an independent treatment planning system (TPS) that utilizes a unique Intelligent Optimization Engine (IOE) that is designed to mimic the way a skilled planner generates treatment plans. The IOE attempts to minimize the impact of planner ability on final treatment plan quality, reducing interpatient plan quality variation (29). While the IOE was designed to reduce the need for the iterative planner interactions, we have found that plan quality can be highly heterogeneous without adequate tuning of a treatment planning template. To the authors' knowledge, there is only a single study outlining planning for stereotactic radiation therapy using the IOE (30). This study by Byrne et al. focused on treatment in the brain and lungs, where the planning focus is on plan conformity and dose falloff. However, for stereotactic APBI planning, sparing of proximal organs-at-risk can be just as important as conformity and dose falloff. Because of this, the template proposed by Byrne et al. is not easily translatable for APBI planning. Thus, the primary endpoint of the proposed work is to develop a treatment planning template which creates clinically acceptable treatment plans for stereotactic APBI in a fully automated fashion, which can easily be disseminated in XML format to any institution wishing to treat APBI using the Ethos. Clinical acceptability will be judged by adherence to published clinical trial guidelines created with traditional planning techniques and *via* evaluation by radiation oncologists with experience treating linear-accelerator based APBI. As a secondary endpoint, the automatically-generated plans will be compared to previously treated plans using standard dose-volume histogram metrics used to evaluate overall plan quality

# 2 Methods and materials

## 2.1 Cohort description

29 patients (30 plans due to one patient with bilateral disease) previously receiving supine APBI treatment for early stage breast cancer (stages 0-2) at our institution between 2019 and 2022 were utilized in this Institutional Review Board (IRB-1207033005) approved study. Nine patients (10 plans) were used to iteratively tune an optimization template and an independent 20 patient cohort was utilized to validate the template *via* automatic replanning without intervention or reoptimization. Seven patients met RAD 1802 inclusion criteria and were simulated and contoured according to trial protocol. Inclusion criteria consisted of age ≥ 50, estrogen receptor (ER) positive, and negative margins of at least 2mm for invasive histology or 3mm for ductal carcinoma *in situ*, carcinoma *in situ*, or T1 disease. Patients receiving neoadjuvant chemotherapy or having multifocal cancer, pure invasive lobular histology, surgical margins< 2mm, a lumpectomy cavity within 5mm of the body contour, or unclear cavity delineation on the planning scan were excluded. Additionally, patients with evaluation PTV (PTV_Eval) volumes exceeding 124cc were excluded based on fat necrosis observed by Timmerman et al. above this threshold (8). 23 patients were not included in the RAD 1802 study, but were simulated, contoured, and planned with the same methods and intent. For all patients, an isotropic 1cm gross tumor volume (GTV) expansion was utilized for clinical target volume (CTV) generation and an isotropic 3mm CTV expansion was utilized for PTV generation. PTV_Eval volumes were created by carving out the PTV at anatomical boundaries (i.e., lung, rib, chest wall, and 5mm from the skin). PTV_Eval volume ranged from 28.6cc to 217.9cc, with an average of 85.2cc. Patients were prescribed 30Gy in five fractions, with an average 98.3% of the PTV_Eval receiving 30Gy in the original clinical plans. Patient characteristics are summarized in Table 1.

## 2.2 Treatment planning

Nine patients previously treated on the Ethos were selected for our tuning cohort (one bilateral patient, four left breast plans, six right breast plans). The tuning cohort was used to establish an Ethos planning template that generated plans meeting RAD 1802 treatment planning goals (Table 2) through iterative planning and fine-tuning of the optimization objectives. A particular emphasis was placed on lowering heart dose to the extent possible while maintaining otherwise similar plan quality to clinical plans. Twenty patients originally receiving supine RT on a Varian TrueBeam Edge were assigned to the validation cohort (seven left breast, thirteen right breast), and were automatically planned using the template resulting from the tuning cohort. Clinically approved Eclipse contours were exported from Eclipse to Ethos and were used for plan generation without modification (i.e., the manually-generated Eclipse lung contour was used in optimization instead of the Ethos auto-contoured lung volume). Ethos validation cohort plans were not reoptimized or renormalized prior to evaluation and were thus evaluated “as-is”.

Clinical Edge plans were originally calculated with Acuros XB (AXB version 15.5.11, Varian Medical Systems) with heterogeneity correction on and dose-to-water selected. Because Ethos automatically calculates with AXB, dose-to-medium (version 16.1.0), all 20 Edge plans were recalculated using dose-to-medium prior to plan comparison. Recalculations preserved beam geometries and field weightings, but plans were re-normalized to the clinically accepted PTV_Eval prescription isodose coverage. A 2.5mm grid was used for dose calculation in both TPS. The Varian TrueBeam Edge is a stereotactic linear accelerator equipped with a 10MV flattening filter free (FFF) beam, high definition MLCs (HDMLC: 0.25cm in the center, 0.50cm in the periphery), and a maximum dose rate of 2400 MU/min. The Ethos is a CBCT-guided adaptive capable rotational linear accelerator equipped with a 6MV FFF beam, dual stacked and staggered MLC banks as its primary form of collimation, and a maximum dose rate of 800 MU/min (29).

The Ethos pre-defined planning geometries selected for this work include equidistant 9- and 12-field IMRT plans, an ipsilateral 7-field IMRT plan, a 2 full-arc VMAT plan, and a 2 half-arc (180-degree arc span) VMAT plan. While Eclipse optimization is dictated by an internal cost function that varies with assigned priority number, Ethos plans are optimized according to the ascending order of planning objectives submitted in the dose preview workspace. The optimum plan geometry generated from each intent was selected by the reviewing physicist based on adherence to RAD 1802 objectives. Selected Ethos plans were exported to Eclipse, where they were benchmarked dosimetrically against clinically delivered Edge plans.

TABLE 1 Patient cohort description.

| Descriptor | median (range) |
|---|---|
| Age (years) | 67 (50 - 85) |
| Laterality | 11 left, 19 right |
| GTV volume (cc) | 10.6 (3.0 - 43.9) |
| CTV volume (cc) | 57.1 (15.0 – 165.6) |
| PTV_Eval volume (cc) | 82.9 (28.6 – 217.9) |

TABLE 2 APBI planning goals utilized in this study.

| Plan metric | Constraint |
|---|---|
| PTV V100% (%) | ≥ 95.0 |
| Ipsilateral breast V30Gy (%) | < 20.0 |
| Ipsilateral breast V15Gy (%) | < 40.0 |
| Contralateral breast V5Gy (%) | < 20.0 |
| Heart V1.5Gy (%) | < 5.0 (right)<br>< 40.0 (left) |
| Ipsilateral lung V9Gy (%) | < 10.0 |
| Contralateral lung V1.5Gy (%) | < 10.0 |
| Skin D0.01cc (Gy) | < 39.5 |
| Rib D0.01cc (Gy) | < 43.0 |
| RTOG CI | < 1.30 |

Eclipse and Ethos objective metrics and dose volume histograms (DVH) were extracted *via* the Eclipse Scripting Application Programming Interface (version 16.1). In addition to presenting the RTOG CI (30), high dose spillage (8) and Paddick gradient index (GI) (31) values were calculated to enable a more holistic plan quality evaluation. The CI, high-dose spillage, and GI are defined in equations (1) – (3).

$$\text{RTOG CI} = \frac{\text{PIV}}{\text{TV}} \tag{1}$$

$$\text{High dose spillage}\ (\%) = 100 * \frac{\text{PIV}_{105\,\%} - \text{TV}_{105\,\%}}{\text{TV}} \tag{2}$$

$$\text{Paddick GI} = \frac{\text{PI}V_{50\,\%}}{\text{PIV}} \tag{3}$$

Here PIV and TV are the prescription isodose volume and treated volume (i.e., PTV_Eval volume), respectively. Subscripts specify the isodose volumes evaluated if different than 100%. The Wilcoxon paired, non-parametric test was utilized to test for significant difference between Eclipse and Ethos plan metrics. When conducting multiple tests on the same dependent variable, the likelihood of observing a significant result by pure chance increases. Thus, a Bonferroni correction was applied to adjust for multiple testing, and $p< 0.004$ is considered significant ($\alpha = 0.05/12$). Statistical analyses were performed in the Python ScyPy library without removal of outliers.

## 2.3 Physician review

Two board certified radiation oncologists specializing in accelerated partial breast treatment qualitatively evaluated all twenty automatically generated Ethos validation cohort plans according to a previously-utilized in-house grading scheme, which is outlined in Table 3 (32). To avoid scoring bias, the physicians were not shown the Ethos optimization template before evaluation; in addition, the physicists did not provide feedback or respond to physicians during evaluation, nor were the physicians aware of the cardiac-sparing emphasis of this study. Rather, the physicians graded each plan based on their past clinical experience and their unique interpretation of the scoring criteria. Physicians were not provided case-specific information and performed evaluations solely with anonymous patient identifiers. In cases where plans selected by the physicist would require modification prior to treatment (i.e., a clinically unacceptable physician score of 1-3), the physicians were asked to evaluate their preferred alternative geometry plan for clinical acceptability. The proportion of planning intents that automatically generated at least one plan that the physician deemed clinically acceptable without re-optimization was then evaluated.

# 3 Results

## 3.1 Planning template and intent

From each APBI planning intent submitted in Ethos, five plans with varying geometries were automatically generated. A total of 110 intent and intent revisions were created in this work, equating to the generation and evaluation of 550 unique APBI plans. Ninety intents were required to iteratively plan the nine patient tuning cohort, and the twenty validation cohort patients were each only planned with one unbiased intent. When plans were optimized solely using the RAD 1802 dosimetric objectives in Table 2, many plans failed to meet planning goals. Thus, the planning template in Table 4 was iteratively procured to maximize the likelihood of achieving all planning objectives. The left sided template is shown as an example, but the right sided template is included in Supplementary Material. Both templates in XML format are available upon request for easy reproduction of this work by other researchers.

The template prioritizes GTV coverage the highest, followed by PTV coverage and heart avoidance. The contralateral lung V1.5Gy was given lower priority in the right-sided than in the left-sided template because heart metrics were more challenging to meet for left sided treatments. This lead the optimizer to spill low dose into the contralateral lung in the absence of a higher priority objective. The left and right templates were identical besides the contralateral lung V1.5Gy constraint. The PTV was cropped out of the ipsilateral breast to avoid conflicting objectives prior to optimization (i.e., asking the optimizer to irradiate the PTV but spare the breast + PTV). The entire ipsilateral breast (including PTV) was designated as a report only structure and was thus not optimized. The template contains three rings constituting seven objectives focused solely on conformity, fall-off, and limiting high dose spillage. The inner, middle, and outer rings are derived from (0 - 0.5)cm, (0.5 - 1.0)cm, and (1.0 – 3.0)cm PTV_Eval expansions inside of the Body, respectively.

TABLE 3 Physician qualitative review grading scheme.

| Score | Description |
|---|---|
| 5 | **Use as-is.** Clinically acceptable plan that could be used for treatment without change. |
| 4 | **Minor edits that are unnecessary.** Reviewer prefers stylistic changes but considers current plan acceptable for treatment. |
| 3 | **Minor edits that are necessary.** Reviewer would require changes prior to treatment and the changes, in the judgment of the reviewer, can be implemented by minimal editing of the objectives. |
| 2 | **Major edits.** Reviewer would require changes prior to treatment and the changes in the judgment of the reviewer would require significant modification of the objectives. |
| 1 | **Unusable.** The plan quality is so poor that it is deemed unsafe to deliver, i.e. would likely result in harm to the patient. |

TABLE 4 Ethos left-sided APBI planning template. The skin was generated using a 3mm inward expansion of the body surface.

| Priority | Structure | Planning Goal | Acceptable Variation |
|---|---|---|---|
| 1 | GTV | V30Gy ≥ 100% | V30Gy ≥ 99% |
| | GTV | D100% ≥ 30.05Gy | D100% ≥ 30Gy |
| | CTV | V30Gy ≥ 99% | V30Gy ≥ 98% |
| | Heart | V1.5Gy ≤ 3% | V1.5Gy ≤ 5% |
| | PTV_Eval | V28.5Gy ≥ 99% | V28.5Gy ≥ 98% |
| | Heart | Dmean ≤ 1.5Gy | Dmean ≤ 2.0Gy |
| | Heart | V7Gy ≤ 0.5% | V7Gy ≤ 10% |
| | Heart | D0.03cc ≤ 12Gy | D0.03cc ≤ 15Gy |
| | PTV_Eval | V30Gy ≥ 97.5% | V30Gy ≥ 95% |
| | PTV_Eval | D0.03cc ≤ 37Gy | D0.03cc ≤ 39Gy |
| | Rib | V30Gy ≤ 0.80cc | V30Gy ≤ 1.00cc |
| | _Lung_R | V1.5Gy ≤ 5% | V1.5Gy ≤ 10% |
| 2 | _Lung_L | V9Gy ≤ 5% | V9Gy ≤ 10% |
| | _Lung_L | V5Gy ≤ 15% | V5Gy ≤ 20% |
| | _RingInner | V30Gy ≤ 6% | V30Gy ≤ 8% |
| | _RingInner | D0.03cc ≤ 30Gy | D0.03cc ≤ 30Gy |
| | _RingInner | Dmean ≤ 20Gy | Dmean ≤ 22Gy |
| | _Lung_L | V15Gy ≤ 1% | _ |
| | _RingMiddle | D0.03cc ≤ 20Gy | D0.03cc ≤ 21Gy |
| | _RingMiddle | Dmean ≤ 11.5Gy | Dmean ≤ 20.0Gy |
| | _Breast_L - PTV_Eval | V15Gy ≤ 15% | V15Gy ≤ 40% |
| | _RingOuter | Dmean ≤ 4.5Gy | Dmean ≤ 14Gy |
| | _RingOuter | D0.03cc ≤ 14Gy | D0.03cc ≤ 15Gy |
| | _Breast_L - PTV_Eval | V20Gy ≤ 5% | V20Gy ≤ 30% |
| | _Breast_L - PTV_Eval | V30Gy ≤ 2% | V30Gy ≤ 20% |
| | _Breast_R | V5Gy ≤ 15% | V5Gy ≤ 20% |
| | _Breast_R | V15Gy ≤ 0.02cc | V15Gy ≤ 0.03cc |
| | Skin | D0.01cc ≤ 37.5Gy | D0.01cc ≤ 39.5Gy |
| | Skin | V36.5Gy ≤ 8cc | V36.5Gy ≤ 10cc |

## 3.2 Plan selection

The twenty patient validation cohort was originally treated on the Edge using 6-field (n = 1), 8-field (n = 1), and 9-field IMRT (n = 5), as well as 2 partial VMAT arcs (n = 13). Validation cohort plan geometries chosen by the physicist to benchmark against the clinical Edge plans are as follows: seven equidistant 9-field plans (35%), four equidistant 12-field plans (20%), five ipsilateral 7-field plans (25%), three VMAT plans with 2 partial arcs (15%), and one VMAT plan with 2 full-arcs (5%). Because sparing the heart is a primary emphasis of this study, Figure 1 shows Ethos and Eclipse axial, sagittal, and coronal dose distributions (1Gy – 38Gy) for the manually generated plan with the highest heart V1.5Gy metric. As is visually evident, the Ethos IOE automatically produces an equidistant 9-field IMRT plan with significant cardiac sparing relative to the manual lateral 6-field IMRT plan.

## 3.3 Dosimetry evaluation

The proposed template automatically generated plans meeting all RAD 1802 objectives for 85% (17/20) of plans without reoptimization. Three initially-selected plans failed to meet the contralateral lung V1.5Gy constraint. No other constraints were violated in any validation cohort plan. 90% (18/20) of the manually generated clinical Edge plans met all objectives; one plan had less than 95% of the PTV receiving prescription dose and one plan exceeded the contralateral lung V1.5Gy constraint. Boxplots showing validation

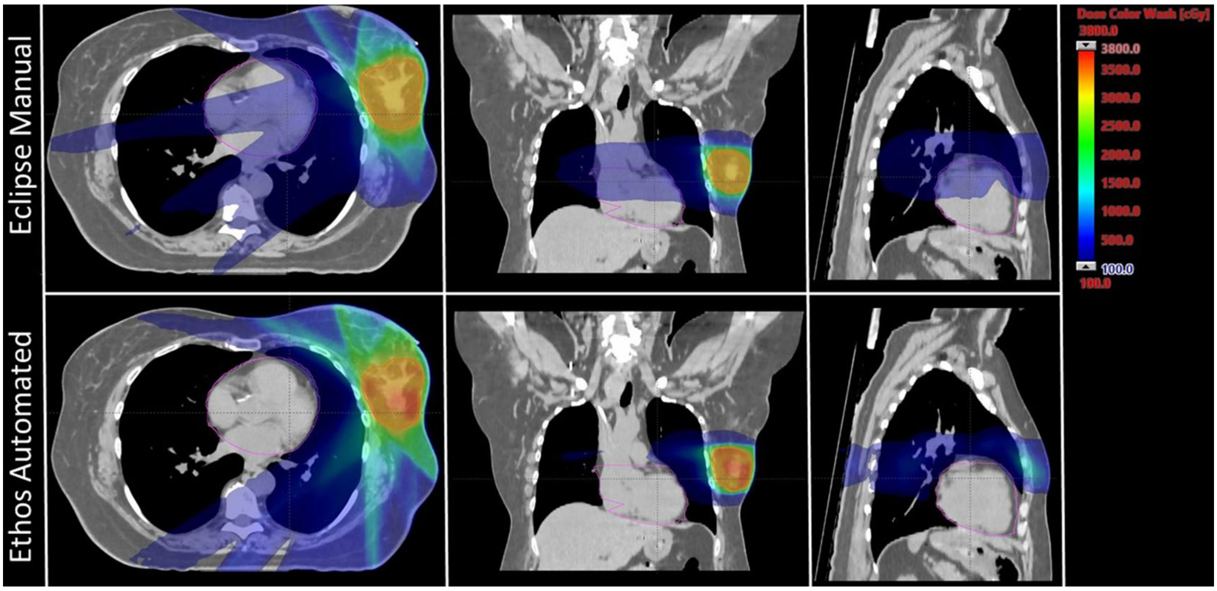


FIGURE 1
Axial, coronal, and sagittal dose distributions of the manual plan with the highest heart V1.5Gy metric for both Eclipse and Ethos. The Eclipse and Ethos plans utilize 6 lateral fields and 9 equidistant fields, respectively. The planning target volume and heart were contoured in red and pink, respectively.

cohort metric summaries for Ethos and Eclipse are displayed in Figure 2. Ethos plans had greater PTV_Eval V100% coverage ($p$ = 0.01), decreased heart V1.5Gy ($p$< 0.001), but increased contralateral breast V5Gy and skin D0.01cc. ($p$ = 0.03 and $p$ = 0.03 respectively). Although several metrics have medians and interquartile ranges (IQR) that differ, only the heart V1.5Gy distributions are significantly different when a Bonferroni correction is applied to adjust for multiple hypothesis testing. The Eclipse left sided heart V1.5Gy IQR and maximum value and are 21.7% and 29.3%, respectively, whereas they are 0.4% and 0.5% for Ethos. The minimum Eclipse right sided heart V1.5Gy metric is 1.4% while the maximum Ethos V1.5Gy metric is 0.6%.

All Ethos and Eclipse plans easily met the 1.30 CI planning objective; one Ethos outlier was much greater than all other plans and one Eclipse plan had a CI of 0.95 due to 92.7% PTV_Eval coverage. The median Eclipse and Ethos CI were 1.05 and 1.06, respectively. 100% of the Eclipse and Ethos plans met the 15% high-dose spillage constraint planning suggested in the Timmerman study (8). There is little discernable difference in high-dose spillage and GI distributions between both TPS when outliers are excluded. Ethos plans generally had more compact high-dose spillage values, but a greater GI IQR. The median Ethos GI was lower, but mean values were similar. While Eclipse CI values were lower than Ethos ($p$ = 0.01), there were no significant quality metric differences between both TPS.

Validation cohort mean DVHs with standard deviation bounds are presented for both TPS in Figure 3. The inferior/superior triangle tips illustrate planning objectives and the insets elucidate DVH difference between both TPS (i.e., Ethos volume minus Eclipse volume as a function of dose). Ethos had superior PTV coverage between approximately 29.5Gy and 31.50Gy, but a lower portion of the target received above 105% of prescription dose, which is generally preferred for SBRT. Ethos significantly spares the heart above 0.25Gy, and on average, the heart volume receiving 1Gy was 10.8% less for automated Ethos plans. All left sided Ethos plans were substantially below the right sided planning objective. While the ipsilateral breast DVH curves are similar for high doses, Ethos spares the breast below approximately 11Gy, with a reduction of 3% breast volume receiving 6Gy. Ethos automated plans had overall higher ipsilateral lung dose above 2.5Gy, but the discrepancy between plan types was at most 1.4%. The template presented here generated plans with generally inferior contralateral breast dose; 3.4% additional volume received 2.3Gy. Because the Ethos planning approach heavily spared the heart, automated planning also resulted in much lower contralateral lung dose, with 6.5% less volume receiving 1Gy on average.

## 3.4 Qualitative evaluation

The physician score summary for physicist-selected Ethos validation cohort plans is shown in Table 5. Physicians A and B considered 75% (15/20) and 90% (18/20) of plans clinically acceptable (scores of 4 or 5) without modification, respectively. 75% of the selected plans (15/20) received a clinically acceptable score from both reviewing physicians. The mode scores of physicians A and B are 4 and 5, respectively. When physicians scored the physicist-selected plan 3 or lower, they then evaluated the alternate plan geometries generated from the same treatment intent and scored the plan they favored most. The five plans receiving a score of 3 from physician A received one 4 and four 5s when alternate plans were evaluated. The two plans receiving a score of 3 from physician B received one 3 and one 4 when alternate plans were evaluated. Thus, at least one plan of treatable quality was automatically generated using the proposed planning approach for 100% of intents for physician A and 95% of intents for physician B.

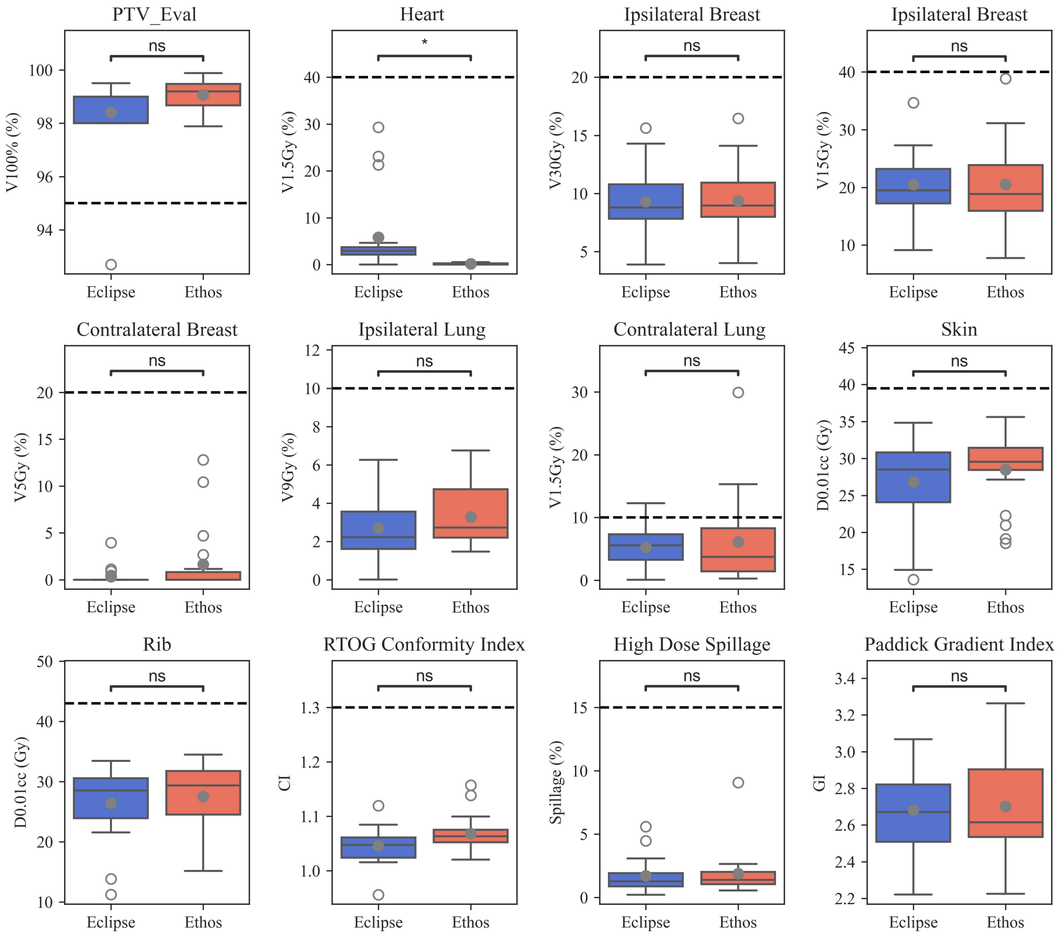


FIGURE 2
Boxplots summarizing manual Eclipse and automated Ethos validation cohort planning metrics. Open and closed circles indicate outlier and mean values, respectively. Significance values for the difference between TPS metric distributions were obtained *via* the Wilcoxon signed rank test and are stratified as follows: ns (not significant): (0.004, *p*, 1.00]; *: (0, *p*, 0.004].

Four plans received a score of 3 from physician A due to the lateral extent of 15Gy streaking prevalent in IMRT plans. One plan received a 3 because physician A preferred the contralateral breast and lung V5Gy be further reduced given favorable patient anatomy. Both plans receiving a score of 3 from physician B were penalized due to lateral extent of 15Gy streaking. However, physician B further specified that they would have considered whole breast treatment over APBI for the plan receiving a score of 3 even after alternate plan evaluation, primarily due to challenging anatomy and target location.

## 4 Discussion

In this work, we evaluated APBI plans automatically generated from a standard planning approach in the Ethos adaptive workspace; nine patients (ten plans) were iteratively re-planned until desired quality was achieved and twenty validation cohort patients were only planned once using the resulting template. 85% of selected validation cohort plans met all planning objectives with significant reduction in heart dose, and physicians A and B scored 75% and 90% of physicist-selected plans as clinically acceptable, respectively. Physicians A and B deemed at least one automatically generated plan clinically acceptable, without modification, 100% and 95% of the time, respectively. This study showed that high quality APBI treatment plans can be created in an automated process with a well-tuned template. Although we do not measure planning times prospectively, our team estimates that creation of a clinically acceptable treatment plan takes between 1 and 4 hours of active planner time for one case. With the proposed template, a plan can be created from scratch with around 5 minutes of active time to set up and approximately an hour of passive time for optimization and dose calculation running as a background process. Additionally, we have shown that automated plans were of similar quality to manual plans while simultaneously reducing heart dose.

Four patients in this study with PTV_Eval volumes > 124cc (two in the tuning cohort, two in the validation cohort) received APBI treatment despite failing to meet RAD 1802 inclusion criteria. The treating physician for these cases, who is also the RAD 1802 principal investigator, was comfortable exceeding this threshold due to personal APBI experience, and because these patients had larger breasts or were receiving re-irradiation. Ipsilateral breast V30Gy and V15Gy objectives were achieved for all four patients.

While Ethos contralateral lung dose was on average significantly less than Eclipse dose due to heart avoidance, there were three outlier plans with high contralateral lung dose. 1/13 right-sided plans and 2/7

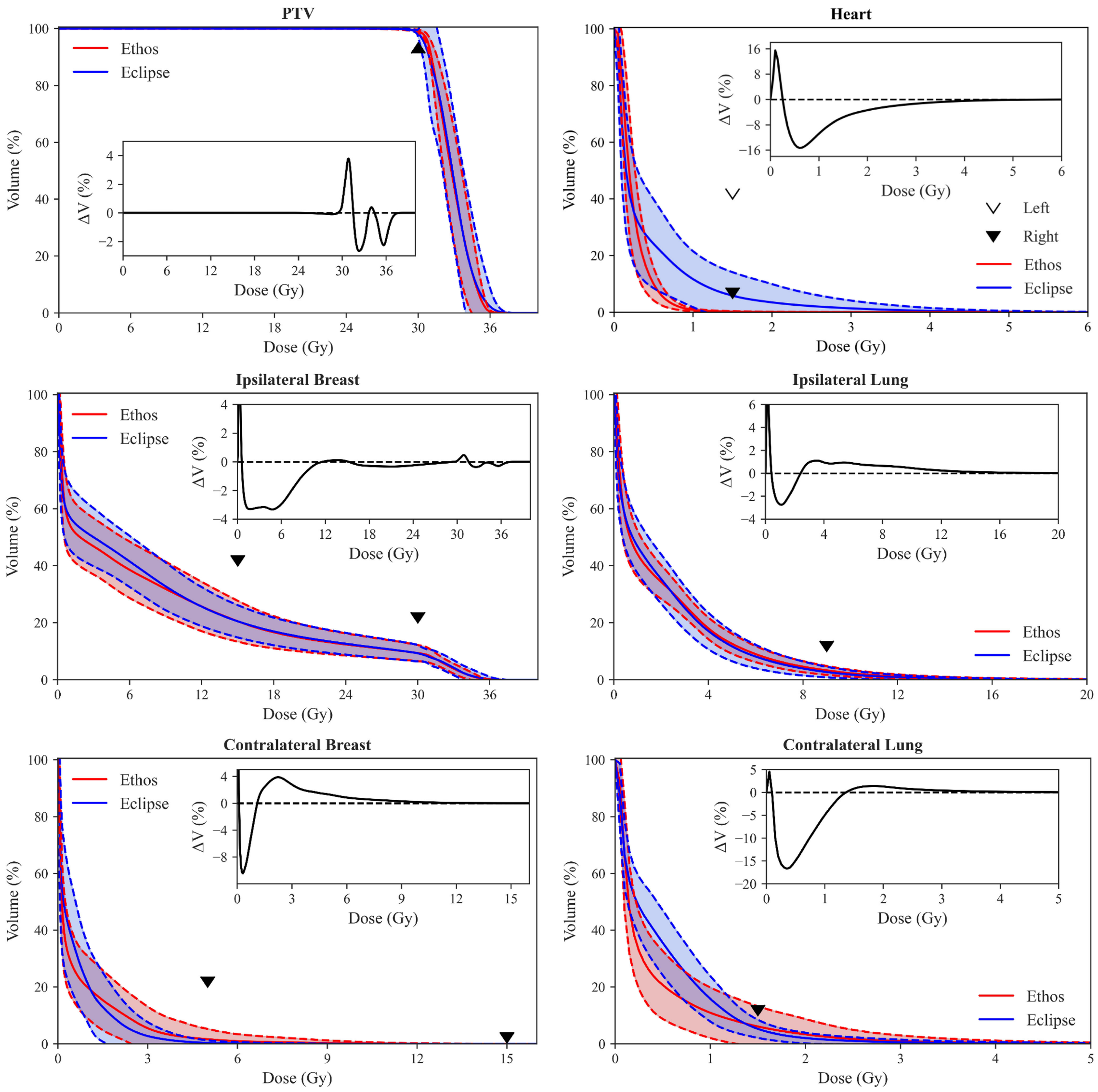


FIGURE 3
Population DVH comparison of Eclipse manual and Ethos automatic plans. Shaded areas show the mean ± standard deviation of all validation cohort data, and the inferior/superior point of triangles illustrate RAD 1802 planning objectives. Insets show the difference between mean population DVHs (i.e., Ethos mean volume minus Eclipse mean volume). Inset axes were sized for optimal visualization.

left-sided plans did not meet the V1.5Gy objective, suggesting that the template may be further improved by increasing the priority of contralateral lung planning goals, especially for the left-sided template. However, the effects of this change in priority require further dosimetric investigation and physician evaluation, as this may affect dose contribution to the ipsilateral lung or contralateral breast, or both. The priority adjustment described above should be considered if contralateral lung dose constraints are exceeded during clinical implementation. Additionally, as this template is being used for adaptive radiation therapy, it is possible that the dose the patient receives to the contralateral lung is lower than the initial plan based on changes in daily anatomy

Significant effort has been dedicated to sparing the heart in lung RT due to high levels of proximal dose (33, 34), but it has also been observed that breast RT induces cardiac toxicity linearly with no apparent dose threshold. Increased risk in major coronary events

TABLE 5 Qualitative scoring summary of plans selected by the physicist for physicians A and B.

| Physician | Score | | | | |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 |
| A | 0 | 0 | 5 | 11 | 4 |
| B | 0 | 0 | 2 | 6 | 12 |

between 7.4% and 19% per additional Gy of mean heart dose have been reported for breast RT (35–37). The initial clinically-treated plans for comparison in this work were of high-quality, with cardiac dose levels below the UAB RAD 1802 objectives for all 20 validation cohort plans. Drawing from the data published by Darby et al, which showed that risk of major coronary events increased linearly with mean heart dose with no apparent threshold (35), it becomes imperative that the planner continue to minimize heart dose, even well below acceptable levels (i.e., V1.5Gy< 5% and 40% for right and left-sided targets, respectively (8, 10)), so long as the net effect on target coverage and sparing of other OARs is not detrimental. To that end, the authors argue that leveraging the Ethos to spare even 1Gy is clinically meaningful, so long as other Ethos plan characteristics are similar in quality to manual Eclipse plans. As shown in Figure 3, the template-generated plans led to reductions in heart dose above 0.5 Gy relative to the Eclipse plans. It should also be mentioned that the OAR dose being spared would be greater were the 6Gy x 5 hypo-fractionated scheme converted to 2Gy equivalent fractions. Furthermore, the Ethos platform allows for RT plan adaption based on daily cone beam CT (CBCT) anatomy (38, 39), which could allow for further reduced doses due to daily re-optimization.

Ethos plans were slightly, but consistently, less conformal than Eclipse plans. While some of this discrepancy may be attributed to template design and optimizer differences, it is due at least in part to tertiary collimation width. The double banked, 10mm width Ethos MLC bank is staggered, effectively producing 5mm width MLCs. The Edge has 2.5mm central HDMLC leaves, resulting in twice the collimation resolution. It is reasonable to assume that Ethos plans would see some measurable reduction in CI and high-dose spillage were the MLC width halved. However, Automated Ethos plans had superior CI values (1.07± 0.05) compared to the 30Gy arm published by Timmerman et al. using the Cyberknife (1.22 ± 0.10) (7). It is also important to note that the mean Ethos validation cohort target volume was smaller than the mean 30Gy arm Cyberknife cohort target volume (Ethos: 77.6cc; Cyberknife: 80.9cc), and CI typically decreases with increasing target size. Thus, the authors argue that the automated plans presented here, while slightly less conformal than Edge plans, are still of high-quality. Further studies are required to deconflate the effects of the different collimators and optimization engine on Ethos plan quality.

The upper Ethos outlier for CI and high dose spillage originated from one plan. This plan presented challenging and abnormal patient anatomy which elucidates a fundamental limitation of this study: fixed beam geometries. The target of interest was the smallest PTV_Eval in the validation cohort and located medially in the upper, inner breast quadrant. The standard field geometries failed to address the patient-specific anatomy; the 2 partial arc and lateral IMRT field geometries span angles from 0° to 180°, clockwise, and the equidistant 9 and 12-field IMRT geometries only space fields every 40° and 30°, respectively. Given the very medial nature of this target, it would have benefitted from partial arcs or densely placed lateral IMRT fields ranging from -90° to 90°. This example highlights that the proposed template does not negate the need for dosimetrist involvement or patient-specific anatomy review; it is expected that abnormal target location or anatomy will require beam geometry modification prior to planning in some instances.

Both reviewing physicians performed a slice-by-slice evaluation of all validation cohort plans. Physicians considered disease extent and location, anatomy favorability, dose distribution shape, and PTV undercoverage in addition to verifying satisfactory DVH metrics. IMRT plan geometries tended to have comparable or even improved GI relative to VMAT, leading the reviewing physicist to select many IMRT plans for further evaluation. However, physician A strongly preferred the consolidated shape of VMAT 15Gy isodose lines compared to IMRT, which tended to exhibit greater lateral extent but similar volume. Physician B was not as opposed to 15Gy streaking, except in more serious cases. This highlights the role of personal preference when reviewing plans qualitatively. While we observed stylistic differences in plan evaluation between the two physician raters, the template provides a mechanism to standardize practices across practitioners, resulting in a large majority of evaluated plans considered acceptable during qualitative review. A future prospective analysis will elucidate if any changes are made after the proposed template is clinically commissioned for use outside of this study.

Artificial intelligence (AI) promises to revolutionize every aspect in radiation oncology care, and has already made a profound impact in enabling the clinical implementation of online adaptive radiotherapies (40, 41). From automated contouring (42–44) to radiotherapy dose estimations (45–47), AI applications are playing a key role increasing efficiency and, often times, improving quality of care through more consistent radiotherapy (48). For example, studies have shown that auto-contouring can significantly save contouring time, providing the critical time savings needed to minimize patient motion during online adaptive treatment design and delivery (49). While most clinical applications currently focus on efficiency improvements, we can expect that in the near future clinical teams will be supported by various AI-driven clinical support systems to compliment decision-making during adaptive treatment's design and delivery. In the current study, we evaluate radiotherapy treatment plans generated using Varian's IOE, which uses an artificial intelligence driven optimization process to automatically generate radiotherapy treatment plans. Our study shows that this novel optimization engine provides high-quality APBI treatment plans for a large majority of cases (with no planner interaction) after defining a robust planning template through a data-driven iterative approach.

APBI treatments were transitioned from the Edge to the Ethos in 2021 at our institution, and APBI treatment for 17 patients has been successfully completed in the Ethos adaptive workspace. During the first course of adaptive treatment on the Ethos, we noticed that the GTV location, volume, and shape changed from simulation to first fraction, and between each subsequent fraction. Consequently, adapted plans significantly spared OARs compared to scheduled plans (i.e., initial treatment-approved plans recalculated onto daily CBCT anatomy). Therefore, even though automated Ethos plans are overall similar in quality to manual Eclipse plans, the added benefit of daily CBCT based adaption vastly outweighs whatever slight deficiencies might exist in the proposed Ethos planning approach (i.e., higher Ethos contralateral breast dose). The impact of daily adaptation on both plan quality and patient outcomes warrants further investigation. Other future projects include implementing the APBI template presented here into our clinical workflow and continuing to generate planning templates for other sites.

The manuscript presented here, including study design and analysis, was developed for consistency with recently published RATING guidelines for generating high-quality planning studies (RAdiotherapy Treatment plannINg study Guidelines) (50). The authors' self-assessment score was 94% (195/207) and the accompanying grading template is added to the Supplementary Material.

Although APBI planning is challenging due to proximity of many OARs and the need for conformity and steep dose gradients, the Ethos templates investigated in this work automatically generate high-quality left- or right-sided APBI plans. Ethos plans had similar target coverage, reduced heart dose, and otherwise similar OAR dose to manual Eclipse plans. 85% of validation cohort plans met all planning objectives, and only the contralateral lung V1.5Gy objective was violated for any plan. Physicians A and B scored at least one plan from each intent of clinically acceptable quality, without reoptimization, 100% and 95% of the time, respectively. Therefore, the approach summarized here enables consistent and high-quality generation of Ethos APBI plans with a specific emphasis on minimizing heart dose.

## Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

## Ethics statement

The studies involving human participants were reviewed and approved by the Institutional Review Board 1207033005. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

## Author contributions

JP conceptualized the study, generated treatment plans, analyzed data, and prepared the manuscript. CC conceptualized the study, anonymized patient data, and prepared the manuscript. YC analyzed data and prepared the manuscript. RP conceptualized the study and prepared the manuscript. MS and DB conceptualized the study, scored treatment plans, and prepared the manuscript. DS conceptualized the study, generated treatment plans, and prepared the manuscript. JH conceptualized the study, analyzed data, and prepared the manuscript. All authors contributed to the article and approved the submitted version.

## Conflict of interest

RP's institution, University of Alabama at Birmingham, has product evaluation agreements and research grants with Varian Medical Systems. He has a patent licensed by UAB Research Foundation to Varian Medical Systems. He has received honoraria for presentations on behalf of Varian Medical Systems. He has received a stipend to speak at Sun Nuclear meetings. Varian Medical Systems provides equipment to UAB as a part of a product evaluation agreement. DB has received research support from Varian including for this study, speaker honoraria, and research support from Novocure. DS received honoraria from Varian medical systems to present on their behalf and is an educational consultant speaker for Varian Medical Systems; He received support through a clinical trial sponsored by Varian Medical Systems.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fonc.2023.1130119/full#supplementary-material

## References

1. Ahmad A. Breast cancer statistics: Recent trends. *Adv Exp Med Biol* (2019) 1152:1–7. doi: 10.1007/978-3-030-20301-6_1

2. Fisher B, Anderson S, Bryant J, Margolese RG, Deutsch M, Fisher ER, et al. Twenty-year follow-up of a randomized trial comparing total mastectomy, lumpectomy, and lumpectomy plus irradiation for the treatment of invasive breast cancer. *New Engl J Med* (2002) 347:1233–41. doi: 10.1056/NEJMoa022152

3. Vicini FA, Kestin LL, Goldstein NS. Defining the clinical target volume for patients with early-stage breast cancer treated with lumpectomy and accelerated partial breast irradiation: A pathologic analysis. *Int J Radiat Oncol Biol Phys* (2004) 60:722–30. doi: 10.1016/j.ijrobp.2004.04.012

4. Olivotto IA, Whelan TJ, Parpia S, Kim DH, Berrang T, Truong PT, et al. Interim cosmetic and toxicity results from RAPID: A randomized trial of accelerated partial breast irradiation using three-dimensional conformal external beam radiation therapy. *J Clin Oncol* (2013) 31:4038–45. doi: 10.1200/JCO.2013.50.5511

5. Vermeulen S, Cotrutz C, Buchanan C, Dawson P, Morris A, Porter B, et al. Accelerated partial breast irradiation: Using the CyberKnife as the radiation delivery platform in the treatment of early breast cancer. *Front Oncol* (2011) 1. doi: 10.3389/fonc.2011.00043

6. Vermeulen SS, Haas JA. CyberKnife stereotactic body radiotherapy and CyberKnife accelerated partial breast irradiation for the treatment of early breast cancer. *Transl Cancer Res* (2014) 3:295–302. doi: 10.3978/j.issn.2218-676X.2014.07.06

7. Rahimi A, Morgan HE, Kim DW, Zhang Y, Leitch M, Wooldridge R, et al. Cosmetic outcomes of a phase 1 dose escalation study of 5-fraction stereotactic partial breast irradiation for early stage breast cancer. *Int J Radiat Oncol Biol Phys* (2021) 110:772–82. doi: 10.1016/j.ijrobp.2021.01.015

8. Rahimi A, Thomas K, Spangler A, Rao R, Leitch M, Wooldridge R, et al. Preliminary results of a phase 1 dose-escalation trial for early-stage breast cancer using 5-fraction stereotactic body radiation therapy for partial-breast irradiation. *Int J Radiat Oncol Biol Phys* (2017) 98:196–205.e2. doi: 10.1016/j.ijrobp.2017.01.020

9. Livi L, Meattini I, Marrazzo L, Simontacchi G, Pallotta S, Saieva C, et al. Accelerated partial breast irradiation using intensity-modulated radiotherapy versus whole breast irradiation: 5-year survival analysis of a phase 3 randomised controlled trial. *Eur J Cancer* (2015) 51:451–63. doi: 10.1016/j.ejca.2014.12.013

10. Liu Y, Veale C, Hablitz D, Krontiras H, Dalton A, Meyers K, et al. Feasibility and short-term toxicity of a consecutively delivered five fraction stereotactic body radiation therapy regimen in early-stage breast cancer patients receiving partial breast irradiation. *Front Oncol* (2022) 12:901312. doi: 10.3389/fonc.2022.901312

11. Popple RA, Brown MH, Thomas EM, Willey CD, Cardan RA, Covington EL, et al. Transition from manual to automated planning and delivery of volumetric modulated arc therapy stereotactic radiosurgery: Clinical, dosimetric, and quality assurance results. *Pract Radiat Oncol* (2021) 11:e163–71. doi: 10.1016/j.prro.2020.10.013

12. Thomas EM, Phillips H, Popple RA, Fiveash JB. Development of a knowledge based model (RapidPlan) for brain metastasis stereotactic radiosurgery and validation with automated non-coplanar treatment planning (HyperArc). *Int J Radiat Oncol* (2017) 99:E727–8. doi: 10.1016/j.ijrobp.2017.06.2353

13. Kisling KD, Ger RB, Netherton TJ, Cardenas CE, Owens CA, Anderson BM, et al. A snapshot of medical physics practice patterns. *J Appl Clin Med Phys* (2018) 19:306–15. doi: 10.1002/acm2.12464

14. Petragallo R, Bardach N, Ramirez E, Lamb JM. Barriers and facilitators to clinical implementation of radiotherapy treatment planning automation: A survey study of medical dosimetrists. *J Appl Clin Med Phys* (2022) 23:e13568. doi: 10.1002/acm2.13568

15. Winkel D, Bol GH, van Asselen B, Hes J, Scholten V, Kerkmeijer LG, et al. Development and clinical introduction of automated radiotherapy treatment planning for prostate cancer. *Phys Med Biol* (2016) 61:8587–95. doi: 10.1088/1361-6560/61/24/8587

16. Moore KL, Brame RS, Low DA, Mutic S. Quantitative metrics for assessing plan quality. *Semin Radiat Oncol* (2012) 22:62–9. doi: 10.1016/j.semradonc.2011.09.005

17. Nelms BE, Robinson G, Markham J, Velasco K, Boyd S, Narayan S, et al. Variation in external beam treatment plan quality: An inter-institutional study of planners and planning systems. *Pract Radiat Oncol* (2012) 2:296–305. doi: 10.1016/j.prro.2011.11.012

18. Ge Y, Wu QJ. Knowledge-based planning for intensity-modulated radiation therapy: A review of data-driven approaches. *Med Phys* (2019) 46:2760–75. doi: 10.1002/mp.13526

19. Li N, Carmona R, Sirak I, Kasaova L, Followill D, Michalski J, et al. Highly efficient training, refinement, and validation of a knowledge-based planning quality-control system for radiation therapy clinical trials. *Int J Radiat Oncol Biol Phys* (2017) 97:164–72. doi: 10.1016/j.ijrobp.2016.10.005

20. Tambe NS, Pires IM, Moore C, Cawthorne C, Beavis AW. Validation of in-house knowledge-based planning model for advance-stage lung cancer patients treated using VMAT radiotherapy. *Br J Radiol* (2020) 93:20190535. doi: 10.1259/bjr.20190535

21. Olanrewaju A, Court LE, Zhang L, Naidoo K, Burger H, Dalvie S, et al. Clinical acceptability of automated radiation treatment planning for head and neck cancer using the radiation planning assistant. *Pract Radiat Oncol* (2021) 11:177–84. doi: 10.1016/j.prro.2020.12.003

22. Rhee DJ, Jhingran A, Kisling K, Cardenas C, Simonds H, Court L. Automated radiation treatment planning for cervical cancer. *Semin Radiat Oncol* (2020) 30:340–7. doi: 10.1016/j.semradonc.2020.05.006

23. Kierkels RG, Visser R, Bijl HP, Langendijk JA, van 't Veld AA, Steenbakkers RJ, et al. Multicriteria optimization enables less experienced planners to efficiently produce high quality treatment plans in head and neck cancer radiotherapy. *Radiat Oncol* (2015) 10:87. doi: 10.1186/s13014-015-0385-9

24. Naccarato S, Rigo M, Pellegrini R, Voet P, Akhiat H, Gurrera D, et al. Automated planning for prostate stereotactic body radiation therapy on the 1.5 T MR-linac. *Adv Radiat Oncol* (2022) 7:100865. doi: 10.1016/j.adro.2021.100865

25. Cilla S, Ianiro A, Romano C, Deodato F, Macchia G, Buwenge M, et al. Template-based automation of treatment planning in advanced radiotherapy: a comprehensive dosimetric and clinical evaluation. *Sci Rep* (2020) 10:423. doi: 10.1038/s41598-019-56966-y

26. Vanderstraeten B, Goddeeris B, Vandecasteele K, van Eijkeren M, De Wagter C, Lievens Y. Automated instead of manual treatment planning? a plan comparison based on dose-volume statistics and clinical preference. *Int J Radiat Oncol Biol Phys* (2018) 102:443–50. doi: 10.1016/j.ijrobp.2018.05.063

27. Han EY, Cardenas CE, Nguyen C, Hancock D, Xiao Y, Mumme R, et al. Clinical implementation of automated treatment planning for whole-brain radiotherapy. *J Appl Clin Med Phys* (2021) 22:94–102. doi: 10.1002/acm2.13350

28. Huang K, Das P, Olanrewaju AM, Cardenas C, Fuentes D, Zhang L, et al. Automation of radiation treatment planning for rectal cancer. *J Appl Clin Med Phys* (2022) 23:e13712. doi: 10.1002/acm2.13712

29. Lim TY, Dragojevic I, Hoffman D, Flores-Martinez E, Kim GY. Characterization of the Halcyon(TM) multileaf collimator system. *J Appl Clin Med Phys* (2019) 20:106–14. doi: 10.1002/acm2.12568

30. Shaw E, Kline R, Gillin M, Souhami L, Hirschfeld A, Dinapoli R, et al. Radiation therapy oncology group: radiosurgery quality assurance guidelines. *Int J Radiat Oncol Biol Phys* (1993) 27:1231–9. doi: 10.1016/0360-3016(93)90548-A

31. Paddick I, Lippitz B. A simple dose gradient measurement tool to complement the conformity index. *J Neurosurg* (2006) 105 Suppl:194–201. doi: 10.3171/sup.2006.105.7.194

32. Pogue JA, Cardenas CE, Harms J, Soike MH, Kole AJ, Schneider CS, et al. Design and validation of an automated radiation therapy treatment planning approach for locally advanced lung cancer. *medRxiv* (2022). doi: 10.1101/2022.09.30.22280584

33. Ferris MJ, Martin KS, Switchenko JM, Kayode OA, Wolf J, Dang Q, et al. Sparing cardiac substructures with optimized volumetric modulated arc therapy and intensity modulated proton therapy in thoracic radiation for locally advanced non-small cell lung cancer. *Pract Radiat Oncol* (2019) 9:e473–81. doi: 10.1016/j.prro.2019.04.013

34. Harms J, Zhang J, Kayode O, Wolf J, Tian S, McCall N, et al. Implementation of a knowledge-based treatment planning model for cardiac-sparing lung radiation therapy. *Adv Radiat Oncol* (2021) 6:100745. doi: 10.1016/j.adro.2021.100745

35. Darby SC, Ewertz M, McGale P, Bennet AM, Blom-Goldman U, Bronnum D, et al. Risk of ischemic heart disease in women after radiotherapy for breast cancer. *N Engl J Med* (2013) 368:987–98. doi: 10.1056/NEJMoa1209825

36. van den Bogaard VA, Ta BD, van der Schaaf A, Bouma AB, Middag AM, Bantema-Joppe EJ, et al. Validation and modification of a prediction model for acute cardiac events in patients with breast cancer treated with radiotherapy based on three-dimensional dose distributions to cardiac substructures. *J Clin Oncol* (2017) 35:1171–8. doi: 10.1200/JCO.2016.69.8480

37. Laugaard Lorenzen E, Christian Rehammar J, Jensen MB, Ewertz M, Brink C. Radiation-induced risk of ischemic heart disease following breast cancer radiotherapy in Denmark, 1977-2005. *Radiother Oncol* (2020) 152:103–10. doi: 10.1016/j.radonc.2020.08.007

38. Mao W, Riess J, Kim J, Vance S, Chetty IJ, Movsas B, et al. Evaluation of auto-contouring and dose distributions for online adaptive radiation therapy of patients with locally advanced lung cancers. *Pract Radiat Oncol* (2022) 12:e329–38. doi: 10.1016/j.prro.2021.12.017

39. Sibolt P, Andersson LM, Calmels L, Sjostrom D, Bjelkengren U, Geertsen P, et al. Clinical implementation of artificial intelligence-driven cone-beam computed tomography-guided online adaptive radiotherapy in the pelvic region. *Phys Imaging Radiat Oncol* (2021) 17:1–7. doi: 10.1016/j.phro.2020.12.004

40. Chapman JW, Lam D, Cai B, Hugo GD. Robustness and reproducibility of an artificial intelligence-assisted online segmentation and adaptive planning process for online adaptive radiation therapy. *J Appl Clin Med Phys* (2022) 23:e13702. doi: 10.1002/acm2.13702

41. El Naqa I, Haider MA, Giger ML, Ten Haken RK. Artificial intelligence: Reshaping the practice of radiological sciences in the 21st century. *Br J Radiol* (2020) 93:20190855. doi: 10.1259/bjr.20190855

42. Cardenas CE, Yang J, Anderson BM, Court LE, Brock KB. Advances in auto-segmentation. *Semin Radiat Oncol* (2019) 29:185–97. doi: 10.1016/j.semradonc.2019.02.001

43. Harms J, Lei Y, Tian S, McCall NS, Higgins KA, Bradley JD, et al. Automatic delineation of cardiac substructures using a region-based fully convolutional network. *Med Phys* (2021) 48:2867–76. doi: 10.1002/mp.14810

44. Wong J, Huang V, Wells D, Giambattista J, Giambattista J, Kolbeck C, et al. Implementation of deep learning-based auto-segmentation for radiotherapy planning structures: a workflow study at two cancer centers. *Radiat Oncol* (2021) 16. doi: 10.1186/s13014-021-01831-4

45. Gotz TI, Schmidkonz C, Chen S, Al-Baddai S, Kuwert T, Lang EW. A deep learning approach to radiation dose estimation. *Phys Med Biol* (2020) 65:035007. doi: 10.1088/1361-6560/ab65dc

46. 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. doi: 10.1002/mp.14827

47. Nguyen D, Long T, Jia X, Lu W, Gu X, Iqbal Z, et al. A feasibility study for predicting optimal radiation therapy dose distributions of prostate cancer patients from patient anatomy using deep learning. *Sci Rep* (2019) 9:1076. doi: 10.1038/s41598-018-37741-x

48. Cornell M, Kaderka R, Hild SJ, Ray XJ, Murphy JD, Atwood TF, et al. Noninferiority study of automated knowledge-based planning versus human-driven optimization across multiple disease sites. *Int J Radiat Oncol Biol Phys* (2020) 106:430–9. doi: 10.1016/j.ijrobp.2019.10.036

49. Vaassen F, Hazelaar C, Vaniqui A, Gooding M, van der Heyden B, Canters R, et al. Evaluation of measures for assessing time-saving of automatic organ-at-risk segmentation in radiotherapy. *Phys Imaging Radiat Oncol* (2020) 13:1–6. doi: 10.1016/j.phro.2019.12.001

50. Hansen CR, Crijns W, Hussein M, Rossi L, Gallego P, Verbakel W, et al. Radiotherapy treatment plannINg study guidelines (RATING): A framework for setting up and reporting on scientific treatment planning studies. *Radiother Oncol* (2020) 153:67–78. doi: 10.1016/j.radonc.2020.09.033

frontiers | Frontiers in Oncology

TYPE Original Research
PUBLISHED 15 March 2023
DOI 10.3389/fonc.2023.1124592

OPEN ACCESS

EDITED BY
Yang Sheng,
Duke University Medical Center,
United States

REVIEWED BY
Dongrong Yang,
Duke University Medical Center,
United States
Zhenyu Yang,
Duke University, United States

*CORRESPONDENCE
Kim Melanie Kraus
✉ kimmelanie.kraus@mri.tum.de

SPECIALTY SECTION
This article was submitted to
Radiation Oncology,
a section of the journal
Frontiers in Oncology

RECEIVED 15 December 2022
ACCEPTED 01 March 2023
PUBLISHED 15 March 2023

CITATION
Kraus KM, Oreshko M, Bernhardt D,
Combs SE and Peeken JC (2023)
Dosiomics and radiomics to predict
pneumonitis after thoracic stereotactic
body radiotherapy and immune
checkpoint inhibition.
*Front. Oncol.* 13:1124592.
doi: 10.3389/fonc.2023.1124592

# Dosiomics and radiomics to predict pneumonitis after thoracic stereotactic body radiotherapy and immune checkpoint inhibition

Kim Melanie Kraus[1,2,3]*, Maksym Oreshko[1,4], Denise Bernhardt[1,3], Stephanie Elisabeth Combs[1,2,3] and Jan Caspar Peeken[1,2,3]

[1]Department of Radiation Oncology, School of Medicine and Klinikum rechts der Isar, Technical University of Munich (TUM), Munich, Germany, [2]Institute of Radiation Medicine (IRM), Helmholtz Zentrum München (HMGU) GmbH German Research Center for Environmental Health, Neuherberg, Germany, [3]Partner Site Munich, German Consortium for Translational Cancer Research (DKTK), Munich, Germany, [4]Medical Faculty, University hospital, Ludwig-Maximilians-Universität (LMU) Munich, Munich, Germany

**Introduction:** Pneumonitis is a relevant side effect after radiotherapy (RT) and immunotherapy with checkpoint inhibitors (ICIs). Since the effect is radiation dose dependent, the risk increases for high fractional doses as applied for stereotactic body radiation therapy (SBRT) and might even be enhanced for the combination of SBRT with ICI therapy. Hence, patient individual pre-treatment prediction of post-treatment pneumonitis (PTP) might be able to support clinical decision making. Dosimetric factors, however, use limited information and, thus, cannot exploit the full potential of pneumonitis prediction.

**Methods:** We investigated dosiomics and radiomics model based approaches for PTP prediction after thoracic SBRT with and without ICI therapy. To overcome potential influences of different fractionation schemes, we converted physical doses to 2 Gy equivalent doses (EQD2) and compared both results. In total, four single feature models (dosiomics, radiomics, dosimetric, clinical factors) were tested and five combinations of those (dosimetric+clinical factors, dosiomics +radiomics, dosiomics+dosimetric+clinical factors, radiomics+dosimetric +clinical factors, radiomics+dosiomics+dosimetric+clinical factors). After feature extraction, a feature reduction was performed using pearson intercorrelation coefficient and the Boruta algorithm within 1000-fold bootstrapping runs. Four different machine learning models and the combination of those were trained and tested within 100 iterations of 5-fold nested cross validation.

**Results:** Results were analysed using the area under the receiver operating characteristic curve (AUC). We found the combination of dosiomics and radiomics features to outperform all other models with $\mathrm{AUC}_{\mathrm{radiomics+dosiomics,\ D}}$ = 0.79

(95% confidence interval 0.78-0.80) and $\mathrm{AUC}_{\mathrm{radiomics+dosiomics,\ EQD2}} = 0.77$ (0.76-0.78) for physical dose and EQD2, respectively. ICI therapy did not impact the prediction result (AUC $\leq$ 0.5). Clinical and dosimetric features for the total lung did not improve the prediction outcome.

**Conclusion:** Our results suggest that combined dosiomics and radiomics analysis can improve PTP prediction in patients treated with lung SBRT. We conclude that pre-treatment prediction could support clinical decision making on an individual patient basis with or without ICI therapy.

KEYWORDS

**pneumonitis, SBRT (stereotactic body radiation therapy), radiomics, dosiomics, immune checkpoint inhibition, model based prediction, lung cancer**

# 1 Introduction

High precision stereotactic body radiation therapy (SBRT) is common standard for treatment of early stage inoperable lung cancer as well as for pulmonary oligo-metastases with excellent local control and an acceptable toxicity profile (1–4). While immunotherapy including checkpoint inhibitors (ICIs) substantially improved the outcome for early lung cancer patients with regard to local tumor control and overall survival (5), the impact of combination with thoracic radiotherapy remains unclear with regard to the development of side effects. PTP is a rather frequent and dose limiting side effect of both, radiation and ICI therapy. As the development of PTP is dose dependent, the risk increases for high fractional doses as applied by SBRT (6). In contrast to the majority of data in the literature, there is also evidence of increased all grade pneumonitis rates (5, 7, 8) after combined radioimmunotherapy with ICIs. This might be of relevance for decision making with regard to further therapeutic options on a patient individual basis.

The applied radiation dose is the most important factor for radiation-dependent pneumonitis. Dose volume histograms (DVHs), however, cannot account for the spatial distribution of the dose and potential effects on the tissue. Thus, prediction of the risk for the development of PTP relying on the spatial distribution could gain clinical advantage for individual patient treatment. Apart from conventional dosimetric approaches, sophisticated methods such as machine learning gain more and more importance for radiation oncology. In recent years, it has been shown that spatial quantitative features assessing the image grey-level distribution extracted from medical imaging data (radiomics) allow for unprecedented predictions of clinical endpoints including patient survival, disease progression, tumor characterization, tumor response and tumor detection (9–17). Analysis using spatial features of the dose distribution or image grey-level distributions, referred to as dosiomics (18–22) or radiomics (23–25) and even the combination of both (26, 27) have also been successfully investigated for prediction of lung toxicity after thoracic radiotherapy in previous studies.

The radiomics features based on pretreatment computed tomography (CT) data showed improvement to predict high grade radiation pneumonitis after definitive radiotherapy (23, 25) and after SBRT (24). Several studies investigated lung toxicity prediction for normofractionated radio(chemo)therapy (RCT). Liang et al. compared dosiomics prediction of radiation pneumonitis after primary thoracic radiotherapy with dosimetric and normal tissue control possibility (NTCP) models and found dosiomics to surpass all other methods (20). In a similar approach, Bourbonne et al. also found dosiomics models to outperform clinical and dosimetric models for prediction of lung toxicity (18). Additionally, combination of radiomics and dosiomics models could even improve the prediction of radiation pneumonitis (26) and for SBRT, other studies support these findings. Jiang et al., additionally revealed improved prediction by machine learning models using dosiomics for different anatomical regions of interest (27), however only for normofractionated radiation schemes. Adachi et al. also tested dosiomics against dosimetric models and against a hybrid model of both resulting in best prediction of radiation pneumonitis achieved with the dosiomics model (19).

These studies investigated PTP prediction after normofractionated R(C)T or SBRT using radiomics and dosiomics combined or dosiomics, respectively. In addition to the above summarized findings, with this study, we aim to find the potential value for the occurrence of PTP after thoracic SBRT using the combination of radiomics and dosiomics analysis of 3D dose distributions and CT data. Additionally, we investigate the potential impact of combined radioimmunotherapy with ICIs.

# 2 Methods

## 2.1 Clinical factors

A total of 110 cases of primary lung cancer or pulmonary metastases received SBRT between 2010 and 2021. All patients provided written informed consent before enrollment. Dose and fractionation schemes varied with fraction doses ranging between

5 Gy and 15 Gy. Patient data involving patient age, sex, karnofsky performance index (KPI), tumor location and size, previous chemotherapy and ICI therapy within 50 days around SBRT. The occurrence of post-treatment pneumonitis (PTP) of all grades according to the Common Terminology Criteria for Adverse Events version 5.0 (28) was detected in follow-up CT scans and from corresponding clinical findings (e.g. dyspnea, cough, pain) during follow-up visits monitored in the patient files. An overview of the patient data is provided in Figure 1.

## 2.2 CT and dose data

Radiotherapy planning CTs, 3D dose distributions, lung and treatment volume segmentations as well as dose volume histogram (DVH) data were selected from the radiotherapy treatment planning system Eclipse (Varian, Paolo Alto). Patients received a 4D-CT prior to radiotherapy. A gross tumor volume (GTV) was delineated on ten phase CTs. Subsequently, an internal target volume was generated which encompasses the GTV across all ten 4D-CT phases. An additional margin of up to 5 mm was added to the internal target volume resulting in the planning target volume (PTV).

Dosimetric data for the total lung included mean dose, the volume receiving at least 5 Gy (V5) and V10, V15, V20, V30, V40, V50, accordingly (29). Required post processing of the segmentation data was performed using the open source platform 3D Slicer (30) and the Radiation Therapy toolkit (31). To take the impact of different fractionation schemes into account, physical dose distributions as extracted from Eclipse were converted into 2 Gy fractions equivalent doses (EQD2) on a voxel basis using an in-house developed Matlab tool (32) according to equation (1) where $D$ is the sum dose over all fractions, $d$ is the fraction dose, and $\frac{\alpha}{\beta}$ is equal to 3 for lung tissue. Dose outside the lung was not considered.

$$EQD2 = D\left[\frac{d+\frac{\alpha}{\beta}}{2+\frac{\alpha}{\beta}}\right] \tag{1}$$

## 2.3 Feature extraction

From each volume of interest (total lung minus GTV, ipsilateral lung minus GTV, PTV + 2cm isotropic margin) 104 radiomics and dosiomics features were extracted from the planning CT and 3D dose distributions using the open-source library Pyradiomics in Python (see Supplemental Table 1 for a list of all features) leading to 312 features, respectively (33). 3D dose maps were treated as images with Gy values as grey-levels. Feature reduction was performed within 1000-fold bootstrapping using pearson intercorrelation coefficient with a cut-off value of 0.7 (arbitrarily chosen to allow sufficient input features for all feature sets) and the Boruta algorithm as previously described (34). In brief, the Boruta algorithm iteratively removes features that appear unimportant for the prediction of the PTP in comparison to synthetic random features (35). The features were ranked according to the frequency of selection overall bootstrap runs. The final feature set was defined as the top-ranking features. The final feature number per model was defined as the median feature number selected over all bootstrap runs. For combined models, the preselected features from each group were used as input for the same procedure.

## 2.4 Machine learning models

The entire process flow is depicted in Figure 2. Three single predictive models (radiomics, dosiomics, clinical factors) and five combined models (dosiomics + radiomics, DVH + clinical factors, radiomics + DVH + clinical data, dosiomics + DVH + clinical data, all) were investigated for the physical dose and EQD2 dose distributions. Different machine learning models with in-built feature reduction including random forest (rf), logistic elastic net regression (glmnet), support vector machine (svmRadial), and logitBoost were trained and tested using 100 iterations of 5-fold nested cross validation in R according to Deist et al. (36). This led to training/test splits of 88:22 and 70:18 in the outer and

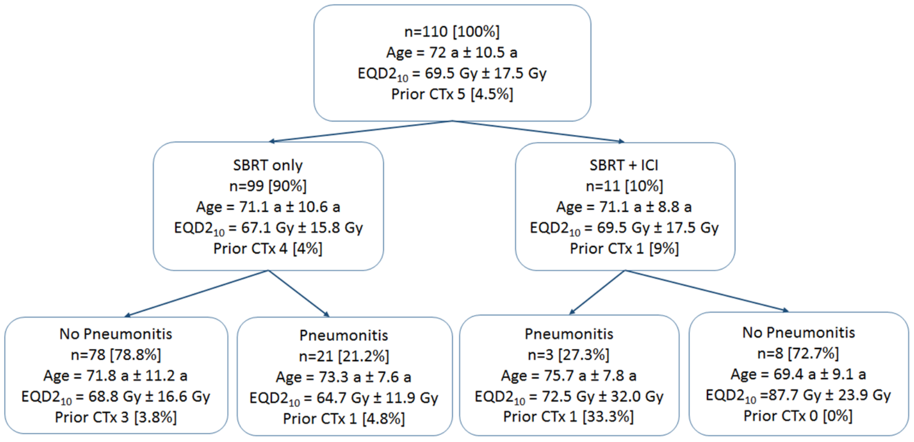


FIGURE 1
Patient data groups. Patient mean age and standard deviations are provided. Prescription doses are given in mean values and standard deviations of equivalent uniform doses for an α/β of 10 Gy (EQD2$_{10}$). The number of patients who received prior chemotherapy (CTx) is provided.

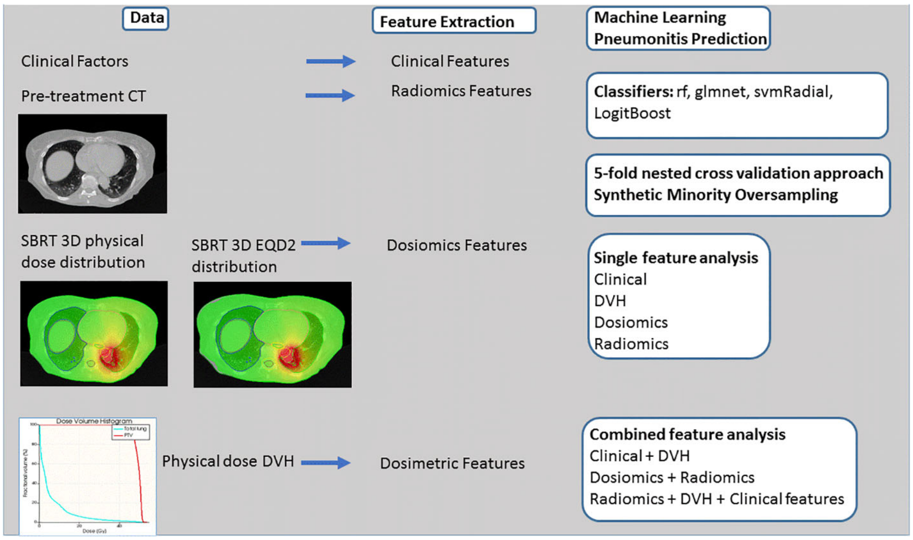


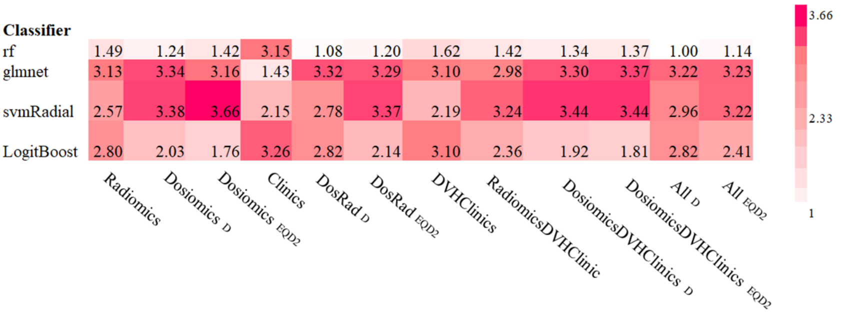
FIGURE 2
Process flow. Clinical, Computed Tomography (CT) and 3D dose volume and dose volume histogram (DVH) data is used for feature extraction. PTP prediction is performed testing different classifiers such as random forest (rf), logistic elastic net regression (glmnet), support vector machine (svmRadial), and logitBoost and 5-fold nested cross validation approach and Synthetic Minority Oversampling. Four single models and five combined models are analyzed.

inner folds, respectively. Due to class imbalance, Synthetic Minority Oversampling Technique (SMOTE) resampling was applied based on the R DMwR package (37) introducing data augmentation of the minority class *via* generation of synthetic samples using a k-nearest neighbor approach and undersampling of the majority class. Due to the small event number, a k-value of 3 was chosen for the k-nearest neighbor procedure. The ratio of oversampling and undersampling was empirically optimized leading to "perc.over" and "perc.under" equaling to the default value of 200%. For comparison, all machine learning models were also calculated without any weighting or SMOTE resampling (see Supplemental Table 3). Hyperparameter optimization was performed within the inner folds using grid search (see Supplemental Table 4 for Hyperparameter Space). Single feature models (e.g. ICI) were modeled using logistic regression. The entire process flow is depicted in Figure 2. Model performance was analysed using the area under the receiver operating characteristic curve (AUC) on the test sets of the outer folds. Data is presented as mean values and confidence intervals with a confidence level of 95%. For comparison of different classifiers used, AUC values were calculated for each dataset and repetition and were ranked by ordering between numbers ranging from 1 to 4 for the four different single predictive models. Data is presented in box and scatterplots as ranked AUC values with each point representing the result of one outer validation fold.

# 3 Results

## 3.1 Comparison of classifiers

Comparison of different classifiers revealed rf to perform best for all models tested resulting in a mean AUC rank value of 1.08 and 1.20 for physical dose and EQD2 analysis. Figure 3 shows the ranked AUC values for all applied classifiers. Based on these findings, for the following analyzes, we chose rf.



| Classifier | Radiomics | Dosiomics $_D$ | Dosiomics $_{EQD2}$ | Clinics | DosRad $_D$ | DosRad $_{EQD2}$ | DVHClinics | RadiomicsDVHClinic | DosiomicsDVHClinics $_D$ | DosiomicsDVHClinics $_{EQD2}$ | All $_D$ | All $_{EQD2}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| rf | 1.49 | 1.24 | 1.42 | 3.15 | 1.08 | 1.20 | 1.62 | 1.42 | 1.34 | 1.37 | 1.00 | 1.14 |
| glmnet | 3.13 | 3.34 | 3.16 | 1.43 | 3.32 | 3.29 | 3.10 | 2.98 | 3.30 | 3.37 | 3.22 | 3.23 |
| svmRadial | 2.57 | 3.38 | 3.66 | 2.15 | 2.78 | 3.37 | 2.19 | 3.24 | 3.44 | 3.44 | 2.96 | 3.22 |
| LogitBoost | 2.80 | 2.03 | 1.76 | 3.26 | 2.82 | 2.14 | 3.10 | 2.36 | 1.92 | 1.81 | 2.82 | 2.41 |

FIGURE 3
Ranked mean AUC values for all classifiers and models tested. Subscripted D and EQD2 refer to physical dose and EQD2, respectively.

## 3.2 Clinical factors

A summary of the clinical parameters collected and the patient groups is given in Table 1 and Figure 1. Most tumors occurred in the right upper lung (30 (27.3%)). A total of 10% of patients received additional ICI therapy. Five patients received primary lung cancer treatment, however, all in a metastasized stage, and six were treated due to metastases. Most of the patients (95%) did not receive previous chemotherapy. Pneumonitis occurred in 24 (21.8%) of all patients, 12.5% (3) of them received additional ICI therapy and 87.5% (21) did not receive additional ICI therapy.

TABLE 1 Clinical factors.

| Characteristic | Value | Value [%] |
|---|---|---|
| Age | | |
| Mean ± SD | 72 ± 10.48 | |
| Range | 33-90 | |
| Sex | | |
| Male | 69 | 62.7 |
| Female | 41 | 37.3 |
| KPI | | |
| Mean ± SD | 95 ± 5.90 | |
| Range | 80-100 | |
| Tumor size | | |
| Mean ± SD | 61162.1 cc ± 75582 cc | |
| Range | 4601.3 cc-524554 cc | |
| Location | | |
| RUL | 30 | 27.3 |
| RML | 2 | 1.8 |
| RLL | 25 | 22.7 |
| LUL | 36 | 32.7 |
| LLL | 11 | 10.0 |
| RC | 3 | 2.7 |
| LC | 3 | 2.7 |
| SBRT+ICI | | |
| Yes | 11 | 10.0 |
| No | 99 | 90.0 |
| Prior CTx | | |
| Yes | 5 | 4.5 |
| No | 105 | 95.5 |
| Pneumonitis | | |
| Yes | 24 | 21.8 |
| No | 85 | 77.3 |

## 3.3 Feature extraction

All features used for feature extraction are listed in Supplement Table 1. The reduced extracted features for all models tested are provided in Supplement Table 2. There was no correlation between ICI and the selected features within the model combining all features. In total, four clinical features were extracted and ranked as follows: tumor size, patient age, tumor location and patient sex. From dosimetric parameters, only V50 and V5 were selected for physical dose and EQD2 features, respectively. Combining both models resulted just in the combination of all single feature models.

Across all model analyzes, 17 to 33 features were found. The most relevant features are listed in Table 2.

## 3.4 Prediction model performance

### 3.4.1 Single feature models

For both, physical dose and EQD2, dosiomics models predicted PTP better than random with $\text{AUC}_{\text{dosiomics, EQD2}} = 0.68$ (0.67-0.70) and $\text{AUC}_{\text{dosiomics,D}} = 0.70$ (0.68-0.71), respectively. The radiomics model achieved the highest predictive value ($\text{AUC}_{\text{radiomics,D}} = 0.73$ (0.72-0.74)). Other classifiers resulted in worse predictive results depicted in Figure 4. DVH parameters achieved PTP prediction yielding no better than random (AUC = 0.43 (0.42-0.46)). Clinical data and ICI therapy status was not predictive for the development of PTP, independent from the applied classifier (AUC = 0.45 (0.44-0.47) and AUC = 0.46 (0.42-0.44)), respectively.

### 3.4.2 Combined feature models

For the combination of radiomics and dosiomics, PTP was predicted better than random with $\text{AUC}_{\text{radiomics+dosiomics, D}} = 0.79$ (0.78-0.80) and $\text{AUC}_{\text{radiomics+dosiomics, EQD2}} = 0.77$ (0.76-0.78) for both, physical dose and EQD2, respectively. Combination with other models including ICI therapy and clinical data did not improve the prediction model. Results are depicted in Figure 5.

# 4 Discussion

Our results indicate that additional ICI therapy has no impact on the prediction of PTP after thoracic SBRT. PTP prediction can be improved by combining radiomics and dosiomics features. This combination outperformed radiomics-only and dosiomics-only models as well as DVH and clinical parameters and can improve prediction of PTP after thoracic SBRT.

In our work, the dosiomics feature model surpassed all clinical and DVH models with an AUC of 0.70 and 0.68 for physical dose and EQD2. These results are well in line with findings in the current literature. For example, in the study of Liang et al. dosiomics analysis with an AUC of 0.78 also resulted in favorable results when compared to dosimetric and NTCP factors (20). Importantly, in our study, prediction of PTP after thoracic SBRT could even be improved when dosiomics features were combined with radiomics

TABLE 2 Features ranked in the order of frequency they have been selected after feature reduction for all models tested.

| Model | Number of reduced features | Ranked features |
|---|---|---|
| Radiomics | 21 | PTV_original_shape_Sphericity |
| | | Total_Lung_original_glcm_Idn |
| | | Ispilateral_Lung_original_glcm_InverseVariance |
| Dosiomics$_{\text{D}}$ | 17 | PTV_original_shape_Sphericity |
| | | Total_Lung_original_shape_Flatness |
| | | PTV_original_glcm_Idmn |
| Dosiomics$_{\text{EQD2}}$ | 17 | PTV_original_shape_Sphericity |
| | | Total_Lung_original_shape_Flatness |
| | | PTV_original_glcm_Idmn |
| Radiomics + Dosiomics$_{\text{D}}$ | 28 | PTV_original_shape_Sphericity |
| | | PTV_original_glszm_SmallAreaLowGrayLevelEmphasis |
| | | Ipsilateral_Lung_original_glcm_InverseVariance |
| Radiomics + Dosiomics$_{\text{EQD2}}$ | 28 | PTV_original_shape_Sphericity |
| | | PTV_original_glcm_Idmn |
| | | Ispilateral_Lung_original_glcm_InverseVariance |
| Radiomics + Clinical Factors + DVH | 27 | PTV_original_shape_Sphericity |
| | | Total_Lung_original_glcm_Idn |
| | | Ispilateral_Lung_original_glcm_InverseVariance |
| Dosiomics$_{\text{D}}$ + Clinical factors + DVH | 22 | PTV_original_shape_Sphericity |
| | | Total_Lung_original_shape_Flatness |
| | | Total_Lung_original_shape_Elongation |
| Dosiomics$_{\text{EQD2}}$ + Clinical factors + DVH | 22 | PTV_original_shape_Sphericity |
| | | Total_Lung_original_shape_Flatness |
| | | Total_Lung_original_shape_Elongation |
| Radiomics +Dosiomics$_{\text{D}}$ + Clinical factors + DVH | 33 | PTV_original_shape_Sphericity |
| | | PTV_original_glszm_SmallAreaLowGrayLevelEmphasis |
| | | Ispilateral_Lung_original_glcm_InverseVariance |
| Radiomics +Dosiomics$_{\text{DEQD2}}$ + Clinical factors + DVH | 33 | PTV_original_shape_Sphericity |
| | | PTV_original_glcm_Idmn |
| | | Ispilateral_Lung_original_glcm_InverseVariance |

Subscripted EQD2 refers to the equivalent dose in 2 Gy fractions and D to the physical dose.

features, which has not been previously shown for patients receiving lung SBRT. Two other works studying patients receiving lung RCT showed combined radiomics and dosiomics models to outperformed single feature class models with an AUC of 0.68 and 0.88 for radiomics and dosiomics combination models, respectively (26, 38). Jiang et al. found the combination of radiomics, dosimetrics, age and tumor T stage to result in a further increased AUC of 0.94.

The total performance of our model with a maximum AUC of 0.79 for the combined radiomics/dosiomics model is well in line with other studies on PTP prediction (20, 24, 26). A few studies, however, achieved larger predictive AUC values above 0.90. Several reasons may explain this fact: 1) The majority of other studies tested prediction of grade ≥ 2 pneumonitis, whereas we tested prediction of all grades of pneumonitis. The reason for this choice of data inclusion was triggered by unknown potential interfering effects associated with the combination of SBRT with immunotherapy that should not be overseen at this stage. Hence, we considered any detectable lung damage or symptom associated with pneumonitis worthwhile to include in our data set. 2) We

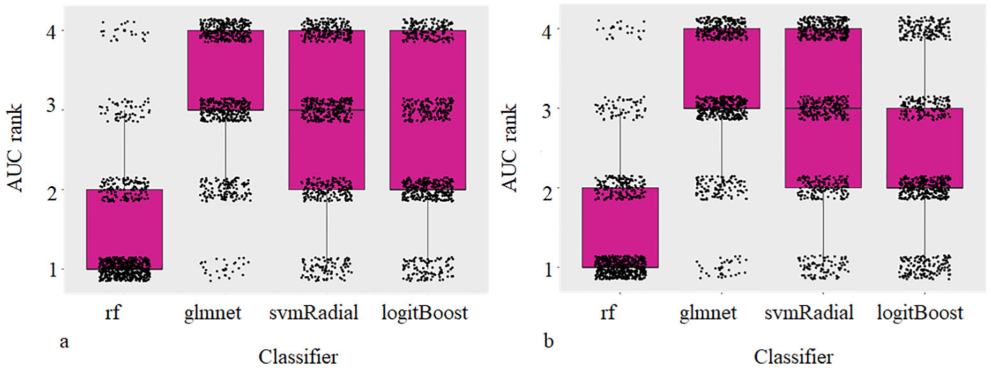


FIGURE 4
Box and Scatterplots showing area under the receiver operating characteristic curves (AUCs) rank values (lower being better) for different classifiers used over all datasets and repetitions for physical (a) and EQD2 dosiomics analysis (b).

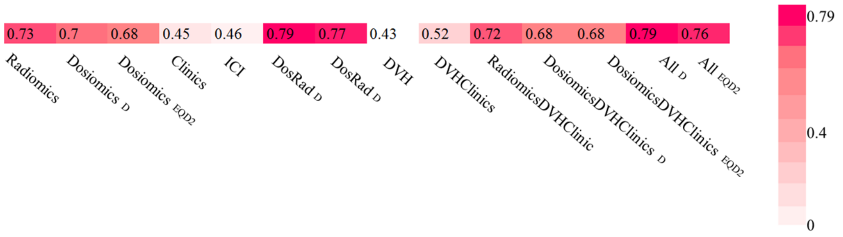


FIGURE 5
Area under the receiver operating characteristic curves (AUCs) heat maps as prediction substitute for PTP for physical Dose and EQD2 using random forest classifier and logistic regression for single feature models.

applied a sophisticated nested cross validation approach separating the validation cohorts for hyperparameter optimization from the actual testing cohort. By iterating the process 100 times, statistical robustness was achieved. This procedure reduces the risk of overly optimistic results that may derive from small test sets or simple cross validation approaches (18, 19, 27).

The DVH features extracted were expected to be comparable with commonly known dosimetric risk factors for radiation pneumonitis such as mean lung dose, the lung volume receiving a dose of 10 Gy and 20 Gy, V10 and V20, respectively. Palma et al. found V20 to be predictive for grade ≥ 2 radiation pneumonitis after radiochemotherapy (39). Tsujino et al. found V20 and Fay et al. V30 and mean lung dose to be most predictive for symptomatic radiation pneumonitis after radiotherapy (40, 41). However, in our study only V50 and V5 were selected by feature extraction and did not predict PTP better than random (AUC< 0.5) in contrast to previous works (18, 19, 24). Different from other studies, we included all grades of pneumonitis into our analysis which could lead to differing dosimetric parameters or even missing correlation of common dosimetric parameters and the development of PTP. In our study, the highest grade of PTP observed was grade 2 in three patient cases and out of these one received additional ICI therapy. Due to the retrospective character of this investigation, the probability of misgrading increases. Our SBRT fractionation schemes cover a rather large range including single doses with a minimum of 5 Gy and lower total doses addressed to treat metastatic disease less likely to cause PTP.

Addition of clinical factors did not improve the prediction of pneumonitis. Likewise, Krafft et al. observed clinical characteristics to not improve the prediction model for high grade pneumonitis after definitive radiotherapy with conventional fractionation (23).

We converted doses to 2 Gy equivalent doses in order to compare different fractionation schemes applied and compared prediction outcome for dosiomics models based on physical dose and biological dosiomics features. As expected, results were comparable with a mean AUC of 0.7 and 0.68 for single dosiomics features analysis using physical dose and EQDs, respectively. This is well in line with findings in the literature (42). However, EQD2 could not further improve the prediction leading to the conclusion that conversion into EQD2 might be unnecessary for PTP prediction.

Development of machine learning models in a dataset of 110 patients is a challenging task, especially when considering the observed imbalance of the predicted outcome. To be able to test our medical hypothesis with regard to the comparison of the predictive values of different feature sets, we decided for several technical steps to allow for optimal training and testing the limitations and reduce the risk of overfitting: 1) we compared multiple machine learning algorithms to determine the algorithm best suited to learn from the small dataset; 2) we applied a cross validation approach with 5 folds to ensure a minimum of samples in the patients subgroups; 3) we applied SMOTE to decrease the influence of the imbalanced outcome variable; 4) we applied multiple feature reduction steps to reduce the feature space to the most predictive features per feature set; 5) no assumption of the optimal number of features was made beforehand; 6) we applied a nested-cross validation approach allowing for repeated testing on unseen data, completely independent of the data used for hyperparameter optimization. Finally, our models achieved good predictive performances in the range of multiple previous works as discussed above. Comparison of the results calculated without any weighting or SMOTE resampling did not change the presented result. Thus, the choice of data augmentation did not alter the relevant comparison of the analyzed models. Importantly, all prediction models were trained and tested simultaneously using the same technical principles and patient subsets down to the internal cross validation folds, guaranteeing optimal comparability. As consequence, the limitations of the model development were the same for all models – allowing for a fair comparison of the predictive value of the underlying feature sets.

Obvious limitations of this study are the retrospective character of data collection. Prospective data could improve the data quality with regard to PTP definition. Patients in this study receiving ICI therapy where all in a metastasized tumor stage. Clearly, this could lead to an imbalance between the SBRT only and the SBRT plus ICI group with slightly enhanced PTP rates (27.3% *vs.* 21.2%) in the combined therapy group. Additionally, there is a lack of patients included in the ICI group resulting a paucity of PTP events. Very few patients were diagnosed with pneumonitis grade ≥ 2, which could limit the clinical relevance of the prediction results. In our study, we decided to include all grade pneumonitis. One reason for this choice was to account for unknown effects occurring during combined radioimmunotherapy, and another reason was the uncertainty of grading coming along with retrospective data collection. Further, we did not apply external test data. External validation, however, is necessary to demonstrate reproducibility of models which is planned in future.

## 5 Conclusions

We demonstrated the potential of combining radiomics and dosiomics features to improve the prediction of PTP after thoracic SBRT. Clinical factors and dosimetric features did not further improve the prediction in this study. Additional immunotherapy with ICIs did not impact the prediction of PTP after thoracic SBRT.

These results could contribute to the prevention of pneumonitis by improvement of clinical decision making prior to thoracic SBRT with and without immunotherapy with ICIs.

## Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material. Further inquiries can be directed to the corresponding author.

## Ethics statement

The studies involving human participants were reviewed and approved by 466/16S. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

## Author contributions

KK and JP designed the project. KK and JP wrote the paper. KK and MO collected and analysed the data. JP, DB, KK and SC provided expert clinical knowledge. All authors contributed to the article and approved the submitted version.

## Funding

KK received funding for this project from the German Cancer Consortium (DKTK).

## Conflict of interest

Authors KK, SC and JP was employed by Helmholtz Zentrum München HMGU GmbH German Research Center for Environmental Health.

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fonc.2023.1124592/full#supplementary-material

## References

1. Ettinger DS, Wood DE, Aisner DL, Akerley W, Bauman JR, Bharat A, et al. Non-small cell lung cancer, version 3.2022, NCCN clinical practice guidelines in oncology. *J Natl Compr Canc Netw* (2022) 20:497–530. doi: 10.6004/jnccn.2022.0025

2. Timmerman RD, Paulus R, Pass HI, Gore EM, Edelman MJ, Galvin J, et al. Stereotactic body radiation therapy for operable early-stage lung cancer: Findings from the NRG oncology RTOG 0618 trial. *JAMA Oncol* (2018) 4:1263–6. doi: 10.1001/jamaoncol.2018.1251

3. Chang JY, Senan S, Paul MA, Mehran RJ, Louie AV, Balter P, et al. Stereotactic ablative radiotherapy versus lobectomy for operable stage I non-small-cell lung cancer: a pooled analysis of two randomised trials. *Lancet Oncol* (2015) 16:630–7. doi: 10.1016/S1470-2045(15)70168-3

4. Chang JY, Mehran RJ, Feng L, Verma V, Liao Z, Welsh JW, et al. Stereotactic ablative radiotherapy for operable stage I non-small-cell lung cancer (revised STARS): long-term results of a single-arm, prospective trial with prespecified comparison to surgery. *Lancet Oncol* (2021) 22:1448–57. doi: 10.1016/S1470-2045(21)00401-0

5. Antonia SJ, Villegas A, Daniel D, Vicente D, Murakami S, Hui R, et al. Overall survival with durvalumab after chemoradiotherapy in stage III NSCLC. *New Engl J Med* (2018) 379:2342–50. doi: 10.1056/NEJMoa1809697

6. Yamashita H, Takahashi W, Haga A, Nakagawa K. Radiation pneumonitis after stereotactic radiation therapy for lung cancer. *World J Radiol* (2014) 6:708–15. doi: 10.4329/wjr.v6.i9.708

7. Shaverdian N, Lisberg AE, Bornazyan K, Veruttipong D, Goldman JW, Formenti SC, et al. Previous radiotherapy and the clinical activity and toxicity of pembrolizumab in the treatment of non-small-cell lung cancer: a secondary analysis of the KEYNOTE-001 phase 1 trial. *Lancet Oncol* (2017) 18:895–903. doi: 10.1016/S1470-2045(17)30380-7

8. Anscher MS, Arora S, Weinstock C, Amatya A, Bandaru P, Tang C, et al. Association of radiation therapy with risk of adverse events in patients receiving immunotherapy: A pooled analysis of trials in the US food and drug administration database. *JAMA Oncol* (2022) 8:232–40. doi: 10.1001/jamaoncol.2021.6439

9. Peeken JC, Wiestler B, Combs SE. Image-guided radiooncology: The potential of radiomics in clinical application. In: Schober O, Kiessling F, Debus J, editors. *Molecular imaging in oncology*. Cham: Springer International Publishing (2020). p. 773–94. doi: 10.1007/978-3-030-42618-7_24

10. Leger S, Zwanenburg A, Leger K, Lohaus F, Linge A, Schreiber A, et al. Comprehensive analysis of tumour Sub-volumes for radiomic risk modelling in locally advanced HNSCC. *Cancers* (2020) 12:3047. doi: 10.3390/cancers12103047

11. Peeken JC, Shouman MA, Kroenke M, Rauscher I, Maurer T, Gschwend JE, et al. Combs SE. a CT-based radiomics model to detect prostate cancer lymph node metastases in PSMA radioguided surgery patients. *Eur J Nucl Med Mol Imaging* (2020) 47:2968–77. doi: 10.1007/s00259-020-04864-1

12. Peeken JC, Bernhofer M, Spraker MB, Pfeiffer D, Devecka M, Thamer A, et al. CT-based radiomic features predict tumor grading and have prognostic value in patients with soft tissue sarcomas treated with neoadjuvant radiation therapy. *Radiother Oncol* (2019) 135:187–96. doi: 10.1016/j.radonc.2019.01.004

13. Peeken JC, Spraker MB, Knebel C, Dapper H, Pfeiffer D, Devecka M, et al. Tumor grading of soft tissue sarcomas using MRI-based radiomics. *eBioMedicine* (2019) 48:332–40. doi: 10.1016/j.ebiom.2019.08.059

14. Shahzadi I, Zwanenburg A, Lattermann A, Linge A, Baldus C, Peeken JC, et al. Analysis of MRI and CT-based radiomics features for personalized treatment in locally advanced rectal cancer and external validation of published radiomics models. *Sci Rep* (2022) 12:10192. doi: 10.1038/s41598-022-13967-8

15. Lang DM, Peeken JC, Combs SE, Wilkens JJ, Bartzsch S. Deep learning based HPV status prediction for oropharyngeal cancer patients. *Cancers* (2021) 13:786. doi: 10.3390/cancers13040786

16. Navarro F, Dapper H, Asadpour R, Knebel C, Spraker MB, Schwarze V, et al. Development and external validation of deep-Learning-Based tumor grading models in soft-tissue sarcoma patients using MR imaging. *Cancers* (2021) 13:2866. doi: 10.3390/cancers13122866

17. Llorián-Salvador O, Akhgar J, Pigorsch S, Borm K, Münch S, Bernhardt D, Rost B, Andrade-Navarro M, Combs S, Peeken J. Machine Learning based Prediction of Pain Response to Palliative Radiation Therapy - is there a Role for Planning CT-based Radiomics and Semantic Imaging Features? *Preprints* (2022). doi: 10.20944/preprints202212.0195.v1

18. Bourbonne V, Da-Ano R, Jaouen V, Lucia F, Dissaux G, Bert J, et al. Radiomics analysis of 3D dose distributions to predict toxicity of radiotherapy for lung cancer. *Radiother Oncol* (2021) 155:144–50. doi: 10.1016/j.radonc.2020.10.040

19. Adachi T, Nakamura M, Shintani T, Mitsuyoshi T, Kakino R, Ogata T, et al. Multi-institutional dose-segmented dosiomic analysis for predicting radiation pneumonitis after lung stereotactic body radiation therapy. *Med Phys* (2021) 48:1781–91. doi: 10.1002/mp.14769

20. Liang B, Yan H, Tian Y, Chen X, Yan L, Zhang T, et al. Dosiomics: Extracting 3D spatial features from dose distribution to predict incidence of radiation pneumonitis. *Front Oncol* (2019) 9:269. doi: 10.3389/fonc.2019.00269

21. Palma G, Monti S, Xu T, Scifoni E, Yang P, Hahn SM, et al. Spatial dose patterns associated with radiation pneumonitis in a randomized trial comparing intensity-modulated photon therapy with passive scattering proton therapy for locally advanced non-small cell lung cancer. *Int J Radiat Oncol Biol Phys* (2019) 104:1124–32. doi: 10.1016/j.ijrobp.2019.02.039

22. Palma G, Monti S, Thor M, Rimner A, Deasy JO, Cella L. Spatial signature of dose patterns associated with acute radiation-induced lung damage in lung cancer patients treated with stereotactic body radiation therapy. *Phys Med Biol* (2019) 64:155006. doi: 10.1088/1361-6560/ab2e16

23. Krafft SP, Rao A, Stingo F, Briere TM, Court LE, Liao Z, et al. The utility of quantitative CT radiomics features for improved prediction of radiation pneumonitis. *Med Phys* (2018) 45:5317–24. doi: 10.1002/mp.13150

24. Hirose T-A, Arimura H, Ninomiya K, Yoshitake T, Fukunaga J-I, Shioyama Y. Radiomic prediction of radiation pneumonitis on pretreatment planning computed tomography images prior to lung cancer stereotactic body radiation therapy. *Sci Rep* (2020) 10:20424. doi: 10.1038/s41598-020-77552-7

25. Kawahara D, Imano N, Nishioka R, Ogawa K, Kimura T, Nakashima T, et al. Prediction of radiation pneumonitis after definitive radiotherapy for locally advanced non-small cell lung cancer using multi-region radiomics analysis. *Sci Rep* (2021) 11:16232. doi: 10.1038/s41598-021-95643-x

26. Puttanawarut C, Sirirutbunkajorn N, Tawong N, Jiarpinitnun C, Khachonkham S, Pattaranutaporn P, et al. Radiomic and dosiomic features for the prediction of radiation pneumonitis across esophageal cancer and lung cancer. *Front Oncol* (2022) 12:768152. doi: 10.3389/fonc.2022.768152

27. Jiang W, Song Y, Sun Z, Qiu J, Shi L. Dosimetric factors and radiomics features within different regions of interest in planning CT images for improving the prediction of radiation pneumonitis. *Int J Radiat Oncol Biol Phys* (2021) 110:1161–70. doi: 10.1016/j.ijrobp.2021.01.049

28. *Common terminology criteria for adverse events (CTCAE) | protocol development*. CTEP. Available at: https://ctep.cancer.gov/protocoldevelopment/electronic_applications/ctc.htm (Accessed March 28, 2022).

29. Kong F-MS, Moiseenko V, Zhao J, Milano MT, Li L, Rimner A, et al. Organs at risk considerations for thoracic stereotactic body radiation therapy: What is safe for lung parenchyma? *Int J Radiat Oncol Biol Phys* (2021) 110:172–87. doi: 10.1016/j.ijrobp.2018.11.028

30. *3D slicer image computing platform* . 3D Slicer. Available at: https://slicer.org/ (Accessed October 7, 2021).

31. Pinter C, Lasso A, Wang A, Jaffray D, Fichtinger G. SlicerRT: Radiation therapy research toolkit for 3D slicer. *Med Phys* (2012) 39:6332–8. doi: 10.1118/1.4754659

32. Matlab. MATLAB, Version R2020a. (2020). Natick, Massachusetts: The MathWorks Inc.

33. *Radiomics*. Available at: https://www.radiomics.io/pyradiomics.html (Accessed November 23, 2022).

34. Peeken JC, Asadpour R, Specht K, Chen EY, Klymenko O, Akinkuoroye V, et al. MRI-Based delta-radiomics predicts pathologic complete response in high-grade soft-tissue sarcoma patients treated with neoadjuvant therapy. *Radiother Oncol* (2021) 164:73–82. doi: 10.1016/j.radonc.2021.08.023

35. Kursa MB, Rudnicki WR. Feature selection with the boruta package. *J Stat Soft* (2010) 36:1–13. doi: 10.18637/jss.v036.i11

36. Deist TM, Dankers FJWM, Valdes G, Wijsman R, Hsu I-C, Oberije C, et al. Machine learning algorithms for outcome prediction in (chemo)radiotherapy: An empirical comparison of classifiers. *Med Phys* (2018) 45:3449–59. doi: 10.1002/mp.12967

37. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic minority over-sampling technique. *jair* (2002) 16:321–57. doi: 10.1613/jair.953

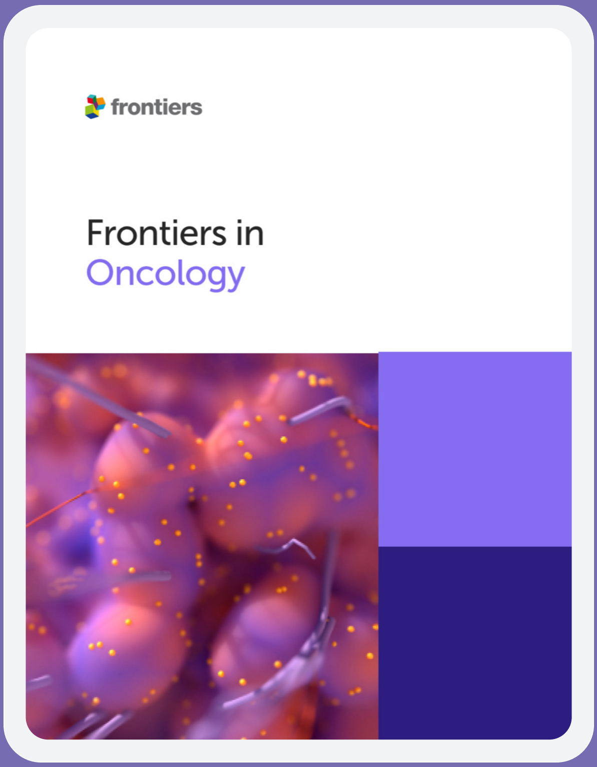
38. Li B, Zheng X, Zhang J, Lam S, Guo W, Wang Y, et al. Lung subregion partitioning by incremental dose intervals improves omics-based prediction for acute radiation pneumonitis in non-Small-Cell lung cancer patients. *Cancers* (2022) 14:4889. doi: 10.3390/cancers14194889

39. Palma DA, Senan S, Tsujino K, Barriger RB, Rengan R, Moreno M, et al. Predicting radiation pneumonitis after chemoradiotherapy for lung cancer: An international individual patient data meta-analysis. *Int J Radiat Oncol Biol Phys* (2013) 85:444–50. doi: 10.1016/j.ijrobp.2012.04.043

40. Tsujino K, Hirota S, Endo M, Obayashi K, Kotani Y, Satouchi M, et al. Predictive value of dose-volume histogram parameters for predicting radiation pneumonitis after concurrent chemoradiation for lung cancer. *Int J Radiat Oncol Biol Phys* (2003) 55:110–5. doi: 10.1016/s0360-3016(02)03807-5

41. Fay M, Tan A, Fisher R, Mac Manus M, Wirth A, Ball D. Dose-volume histogram analysis as predictor of radiation pneumonitis in primary lung cancer patients treated with radiotherapy. *Int J Radiat Oncol Biol Phys* (2005) 61:1355–63. doi: 10.1016/j.ijrobp.2004.08.025

42. Puttanawarut C, Sirirutbunkajorn N, Khachonkham S, Pattaranutaporn P, Wongsawat Y. Biological dosiomic features for the prediction of radiation pneumonitis in esophageal cancer patients. *Radiat Oncol* (2021) 16:220. doi: 10.1186/s13014-021-01950-y



Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact