Aus der
Klinik und Poliklinik für Strahlentherapie und Radioonkologie
Klinik der Ludwig-Maximilians-Universität München
Direktor: Prof. Dr. Claus Belka

Prior knowledge-aware deep learning auto-segmentation for MRI-guided radiotherapy

Dissertation
zum Erwerb des Doktorgrades der Naturwissenschaften
an der Medizinischen Fakultät der
Ludwig-Maximilians-Universität München

vorgelegt von
Maria Rädler, geb. Kawula

aus
Limanowa, Polen

Jahr
2025

Mit Genehmigung der Medizinischen Fakultät
der Ludwig-Maximilians-Universität München

Betreuer:	PD Dr. Christopher Kurz
Zweitgutachter:	Prof. Dr. Olaf Dietrich
Dekan:	Prof. Dr. med. Thomas Gudermann
Tag der mündlichen Prüfung:	19. Dezember 2025

Affidavit

Rädler, Maria

Surname, first name

I hereby declare, that the submitted thesis entitled:

Prior knowledge-aware deep learning auto-segmentation for MRI-guided radiotherapy

is my own work. I have only used the sources indicated and have not made unauthorised use of services of a third party. Where the work of others has been quoted or reproduced, the source is always given.

I further declare that the dissertation presented here has not been submitted in the same or similar form to any other institution for the purpose of obtaining an academic degree.

Kraków, 03.01.2026

Place, Date

Maria Rädler

Signature doctoral candidate

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Dekanat Medizinische Fakultät
Promotionsbüro

Confirmation of congruency between printed and electronic version of the doctoral thesis

Rädler, Maria

name, first name

I hereby declare that the electronic version of the submitted thesis, entitled:

Prior knowledge-aware deep learning auto-segmentation for MRI-guided radiotherapy

is congruent with the printed version both in content and format.

Kraków, 03.01.2026
Place, Date

Maria Rädler
Signature doctoral candidate

Table of Contents

Abstract

Thanks to combined magnetic resonance (MR) linear accelerators (LINACs) called MR-LINACs, the integration of image guidance and online plan adaptation into radiotherapy became feasible. Dose-free magnetic resonance imaging (MRI), daily plan adjustments, and gating were shown to improve treatment outcomes and reduce side effects. However, treatment adaptation entails time-consuming segmentation of organs at risk (OARs) and target volumes, not only on the pre-treatment planning MRI but also on all daily fraction images. Quick segmentation of fraction images is crucial to minimize patient waiting time for irradiation. To improve and speed up fraction segmentation compared to conventional auto-segmentation models, expert knowledge from the planning phase could be leveraged.

The aim of this work was to investigate deep learning (DL) auto-segmentation methods for cancer patients receiving conventional radiotherapy or magnetic resonance-guided radiation therapy (MRgRT). The main focus was on personalized models to enhance the segmentation in fractionated online adaptive treatments.

The first paper analyzed the impact of DL contours on dose optimization in conventional radiotherapy of prostate cancer patients. It investigated the possible correlation between contour quality and the quality of treatment plans optimized using these contours. The study concluded that networks achieving state-of-the-art segmentation performance predict contours that lead to satisfactory dose distribution in most investigated cases. No strong correlations were found between the geometric and dosimetric metrics.

The second study explored personalized models generated by fine-tuning conventional DL population models with the manually delineated planning MRI of a patient. The target group were prostate cancer patients undergoing MRgRT. Personalized models effectively learned organ shapes as defined on the planning image for each patient. They were particularly beneficial for target volume segmentation and for patients with unusual anatomies.

The third study investigated networks for combined image registration and segmentation as an alternative to personalized models from the second paper. These networks were trained to register the planning and fraction MRIs and propagate the planning expert contours to the daily anatomy. The registration-based networks were successful in prostate clinical target volume (CTV) segmentation. The latter is difficult to segment with population models due to the individual shape but does not change significantly throughout the weeks. Personalized models performed better than registration networks for OARs undergoing larger changes.

The last study explored further options for personalized training. It analyzed the

impact of population models on patient-specific segmentation networks for abdominal OARs. The study investigated the adjustment of personalized models to the patient's anatomy from the planning day and optionally from additional fractions. It also explored whether training from scratch (i.e., without the population model) using only the segmented planning MRI is sufficient to create a personalized model. The study showed that by fine-tuning population models with expert delineations of a given patient (planning or planning plus previous fractions), the models predict clinically usable contours with little to no corrections needed. Using single patient data was insufficient to develop robust personalized models. Regardless, all personalized models improved with updates to the prior fraction anatomies.

Summarizing, personalized models created by fine-tuning population models with expert-segmented images of a given patient performed best among all investigated alternatives. Training times of personalized models were short enough for clinical implementation. By reducing the necessity of manual contour corrections, personalized models have the potential to shorten treatment adaptation and reduce inter and intra-observer segmentation variability in MRgRT at MR-LINACs.

Zusammenfassung

Die Integration von Bildführung und Online-Plananpassung in die Strahlentherapie wude dank der kombinierten Magnetresonanz-(MR)-Linearbeschleuniger (LINACs), den so-genannten MR-LINACs, ermöglicht. Bilgebung ohne Strahlungsbelastung durch die Magnetresonanztomographie (MRT), tägliche Plananpassungen und Gating können nachgewiesen die Behandlungsergebnisse verbessern und die Nebenwirkungen verringern. Die Anpassung des Bestrahlungsplans erfordert jedoch eine zeitaufwändige Segmentierung der Risikoorgane und der zu bestrahlenden Volumen, und das nicht nur bei den MRT-Bildern der Bestrahlungsplanung, sondern auch bei allen täglich aufgenommenen Fraktionsbildern. Eine schnelle Segmentierung von Fraktionsbildern ist von entscheidender Bedeutung, um die Wartezeit der Patienten zu minimieren, sowie die Zahl der Patienten am MR-LINAC zu maximieren. Zur Verbesserung und Beschleunigung der Segmentierung der Fraktionsbilder im Vergleich zu herkömmlichen Auto-Segmentierungsmethoden könnte die ärztlich verifizierte Segmentierung aus der Planungsphase genutzt werden.

Ziel dieser Arbeit war die Untersuchung von deep learning (DL) Autosegmentierungsmethoden für Krebspatienten, die eine konventionelle Strahlentherapie oder Magnetresonanz-gestützte Radiotherapie (MRgRT) erhalten. Das Hauptaugenmerk lag dabei auf personalisierten Modellen zur Verbesserung der Segmentierung bei fraktionierten online-adaptiven Behandlungen.

In der ersten Arbeit wurden die Auswirkungen der DL-Konturen auf die Dosisoptimierung in der konventionellen Strahlentherapie von Prostatakrebspatienten analysiert. Eine mögliche Korrelation zwischen der Qualität der Konturen und der Qualität der Behandlungspläne, die anhand dieser Konturen optimiert wurden, wurde untersucht. Die Studie kam zu dem Schluss, dass DL-Netzwerke, die eine state-of-the-art Segmentierungsleistung erreichen, in den meisten untersuchten Fällen Konturen vorhersagen, die zu einer zufriedenstellenden Dosisverteilung führen. Es wurden jedoch keine starken Korrelationen zwischen den geometrischen und dosimetrischen Metriken festgestellt.

Die zweite Studie untersuchte personalisierte Modelle, die durch Feinabstimmung konventioneller DL-Populationsmodelle mit den manuell erstellten Planungs-MRT-Segmentierungen eines Patienten optimiert wurden. Die Zielgruppe waren Prostatakrebspatienten, die mit MRgRT behandelt wurden. Personalisierte Modelle konnten effektiv die spezifischen Organstrukturen der Patienten anhand der Planungsbilder und Konturen erlernen. Sie waren besonders vorteilhaft für die Segmentierung von Tumorvolumen und für Patienten mit ungewöhnlicher Anatomie.

In der dritten Studie wurden neuronale Netzwerke für die kombinierte Bildregi-

strierung und -segmentierung als Alternative zu den personalisierten Modellen aus der zweiten Studie untersucht. Diese Netzwerke wurden darauf trainiert, die Planungs- und Fraktions-MRTs aufeinander zu registrieren um die Expertenkonturen der Planungsphase auf die aktuelle Anatomie zu übertragen. Die Bildregistrierungs-basierten Netzwerke waren bei der Segmentierung von Prostata-Zielvolumen erfolgreich. Letztere sind aufgrund der individuellen Form schwer mit Populationsmodellen zu segmentieren, ändern sich aber im Laufe der Behandlungsfraktionen nicht wesentlich. Personalisierte Modelle schnitten besser ab als Bildregistrierungsnetzwerke für Risikoorgane, die größere Veränderungen durchlaufen.

Die letzte Studie untersuchte weitere Varianten des personalisierten Trainings. Sie analysierte die Auswirkungen von Populationsmodellen auf patientenspezifische Segmentierungsnetzwerke für abdominale Risikoorgane. Die Studie untersuchte die Anpassung der personalisierten Modelle basierend auf der Anatomie des Patienten am Planungstag und optional an zusätzlichen Fraktionen. Es wurde zudem untersucht, ob ein eigenständiges Training (d. h. nicht auf dem Populationsmodell basierend) nur unter Verwendung des segmentierten Planungs-MRT ausreicht, ein personalisiertes Modell zu erstellen. Die Studie zeigte, dass die Modelle durch Feinabstimmung von Populationsmodellen mithilfe von Expertensegmentierungen eines gegebenen Patienten (Planungbild oder Planungs- und vorherige Fraktionsbilder) klinisch akzeptable Konturen vorhersagen, wobei nur wenige bis keine Korrekturen erforderlich sind. Die Verwendung von Daten eines einzelnen Patienten war nicht ausreichend, um robuste personalisierte Modelle zu erstellen. Unabhängig davon verbesserten sich alle personalisierten Modelle unter Verwendung zusätzlicher Fraktionsbilder und Konturen.

Zusammenfassend lässt sich sagen, dass personalisierte Modelle, die durch Feinabstimmung von Populationsmodellen mithilfe von ärztlich verifizierten Segmentierungen der jeweiligen Patienten erstellt wurden, unter allen untersuchten Alternativen am besten abschnitten. Die Trainings der personalisierten Modelle sind ausreichend schnell für die klinische Anwendung. Indem sie die Notwendigkeit manueller Konturkorrekturen verringern, haben personalisierte Modelle das Potenzial, die Behandlungsanpassung zu verkürzen und die Segmentierungsvariabilität bei der MRgRT am MR-LINAC zu reduzieren.

List of publications and conference contributions

First author publications

1. **M. Kawula**, D. Purice, M. Li, G. Vivar, S. A. Ahmadi, K. Parodi, C. Belka, G. Landry, C. Kurz: Dosimetric impact of deep learning-based CT auto-segmentation on radiation therapy treatment planning for prostate cancer
Radiation Oncology 2022;17(1):21.

2. **M. Kawula**, I. Hadi, L. Nierer, M. Vagni, D. Cusumano, L. Boldrini, L. Placidi, S. Corradini, C. Belka, G. Landry, C. Kurz: Patient-specific transfer learning for auto-segmentation in adaptive 0.35 T MRgRT of prostate cancer: a bi-centric evaluation
Medical Physics 2023;50:1573–85.

3. **M. Kawula**, M. Vagni, D. Cusumano, L. Boldrini, L. Placidi, S. Corradini, C. Belka, G. Landry, C. Kurz: Prior knowledge-based deep learning auto-segmentation in magnetic resonance imaging-guided radiotherapy of prostate cancer
Physics and Imaging in Radiation Oncology 2023;28:100498.

4. **M. Kawula**, S. Marschner, C. Wei, S. Corradini, C. Belka, G. Landry, C. Kurz: Personalized deep learning auto-segmentation models for adaptive fractionated magnetic resonance-guided radiation therapy of the abdomen
Medical Physics, 2024

Co-authored publications

1. Y. Xiong, M. Rabe, L. Nierer, **M. Kawula**, S. Corradini, C. Belka, M. Riboldi, G. Landry, C. Kurz: Assessment of intrafractional prostate motion and its dosimetric impact in MRI-guided online adaptive radiotherapy with gating
Strahlentherapie und Onkologie 2023;199(6), 544-553.

2. H. Schmitz, A. Thummerer, **M. Kawula**, E. Lombardo, K. Parodi, C. Belka, F. Kamp, C. Kurz, G. Landry: ScatterNet for projection-based 4D cone-beam computed tomography intensity correction of lung cancer patients
Physics and Imaging in Radiation Oncology 2023;27,100482.

3. M. F. Ribeiro, S. Marschner, **M. Kawula**, M. Rabe, S. Corradini, C. Belka, M. Riboldi, G. Landry, C. Kurz: Deep learning based automatic segmentation of organs-at-risk for 0.35 T MRgRT of lung tumors
Radiation Oncology 2023;27,100482.

4. Y. Xiong, M. Rabe, C. Rippke, **M. Kawula**, L. Nierer, S. Klüter, C. Belka, M. Niyazi, J. Hörner-Rieber, S. Corradini, G. Landry, C. Kurz: Impact of daily plan adaptation on accumulated doses in ultra-hypofractionated magnetic resonance-guided radiation therapy of prostate cancer
Physics and Imaging in Radiation Oncology 2024;100562.

5. M. Vagni, H. E. Tran, A. Romano, G. Chiloiro, L. Boldrini, K. Zormpas-Petridis, **M. Kawula**, G. Landry, C. Kurz, S. Corradini, C. Belka, L. Indovina, M. A. Gambacorta, L. Placidi, D. Cusumano: Auto-segmentation of pelvic organs at risk on 0.35 T MRI using 2D and 3D Generative Adversarial Network models
Physica Medica 2024;119,103297.

6. M. Vagni, H. E. Tran, F. Catucci, G. Chiloiro, A. D'Aviero, A. Re, A. Romano, L. Boldrini, **M. Kawula**, E. Lombardo, C. Kurz, G. Landry, C. Belka, L. Indovina, M. A. Gambacorta, D. Cusumano, L. Placidi: Impact of bias field correction on 0.35 T pelvic MR images: evaluation on generative adversarial network-based OARs' auto-segmentation and visual grading assessment
Frontiers in oncology 2024;14,1294252.

First author oral presentations

1. **M Kawula**, I. Hadi, L. Nierer, M. Vagni, D. Cusumano, L. Boldrini, L. Placidi, S. Corradini, C. Belka, G. Landry, C. Kurz: Patient-Specific Transfer Learning to Enhance the Performance of Deep Learning Auto-Segmentation in 0.35 T MRgRT for Prostate Cancer
AAPM Annual Meeting 2022

2. **M. Kawula**, I. Hadi, L. Nierer, M. Vagni, D. Cusumano, L. Boldrini, L. Placidi, S. Corradini, C. Belka, G. Landry, C. Kurz: Patientenspezifisches Transfer-Lernen zur Verbesserung Deep-Learning-basierter Autosegmentierung in der 0,35 T MR-geführten Strahlentherapie von Prostatakrebspatienten
DGMP Jahrestagung 2022

3. **M. Kawula**, I. Hadi, D. Cusumano, L. Boldrini, L. Placidi, S. Corradini, C. Belka, G. Landry, C. Kurz: AI auto-segmentation for MRgRT of prostate cancer: evaluating 269 MR images from two institutes
ESTRO 2022

4. **M. Kawula**, I. Hadi, L. Nierer, M. Vagni, D. Cusumano, L. Boldrini, L. Placidi, S. Corradini, C. Belka, G. Landry, C. Kurz: Comparison of AI-based autosegmentation techniques exploiting prior knowledge at a 0.35 T MR-Linac
ESTRO 2023

5. **M. Kawula**†, M. F. Ribeiro, S. Marschner, S. Corradini, C. Belka, G. Landry, C. Kurz: Vergleich patientenspezifischer AI-Modelle für die Autosegmentierung von Risikostrukturen in der 0.35 T MR-geführten Strahlentherapie von abdominalen Läsionen
DGMP Jahrestagung 2023
†**Winner of the Young Investigator Award 2023**

6. **M. Kawula**, S. Marschner, M. F. Ribeiro, S. Corradini, C. Belka, G. Landry, C. Kurz: Personalized auto-segmentation models for adaptive fractionated MRgRT
ESTRO 2024

Co-authored oral presentations

1. H. Schmitz, E. Lombardo, **M. Kawula**, K. Parodi, C. Belka, F. Kamp, C. Kurz, G. Landry: ScatterNet for 4D cone-beam CT intensity correction of lung cancer patients
ESTRO 2023

2. H. Schmitz, **M. Kawula**, E. Lombardo, A. Thummerer, K. Parodi, C. Belka, F. Kamp, G. Landry, C. Kurz: Deep learning basierte Intensitätskorrektur von 4D-Conebeam Computertomographie Bildern
DGMP Jahretagung 2023

3. M. Vagni, H.E. Tran, A. Romano, L. Boldrini, G. Chiloiro, G. Landry, C. Kurz, S. Corradini, **M. Kawula**, E. Lombardo, K. Zormpas Petridis, M.A. Gambacorta, L. Indovina, C. Belka, V. Valentini, L. Placidi, D. Cusumano: A comparison between 2D and 3D GAN for rectum and bladder auto-segmentation on 0.35 T MR images
Physica Medica 2023;115,102770.

First author poster presentations

1. **M. Kawula**, D. Purice, M. Li, G. Vivar, S. Ahmadi, K. Parodi, C. Belka, G. Landry, C. Kurz: Dosimetric impact of auto segmentation on treatment planning in IMRT for prostate patients
ESTRO 2021

2. **M. Kawula**, D. Purice, M. Li, G. Vivar, S. A. Ahmadi, K. Parodi, C. Belka, G. Landry, C. Kurz: Impact of automatic organ segmentation on dose optimization in IMRT for prostate patients.
Dreiländertagung der Medizinischen Physik 2021

Co-authored poster presentations

1. M. Vagni, H.E. Tran, A. Romano, L. Boldrini, G. Chiloiro, G. Landry, C. Kurz, S. Corradini, **M. Kawula**, E. Lombardo, M.A. Gambacorta, L. Indovina, C. Belka, V. Valentini, L. Placidi, D. Cusumano: A comparison between 2D and 3D GAN as a supporting tool for rectum segmentation on 0.35 T MR images
   ESTRO 2023

2. F. Thiele, **M. Kawula**, S. Katzendobler, R. Bodensohn, C. Kurz, G. Landry, J. Weller, M. Niyazi: AI-based assessment of brain volume decrease after SRS and WBRT
   Annual Meeting of the SNO 2023

3. Y. Xiong, **M. Kawula**, M. Ribeiro, S. Marschner, S. Corradini, C. Belka, G. Landry, M. Rabe, C. Kurz: Impact of patient specific deep learning OAR segmentation on accumulated online adapted MRgRT dose
   ESTRO 2024

Own contributions to publications included in this dissertation

1. **Dosimetric impact of deep learning-based CT auto-segmentation on radiation therapy treatment planning for prostate cancer**.
   Maria Kawula, Dinu Purice, Minglun Li, Gerome Vivar, Seyed-Ahmad Ahmadi, Katia Parodi, Claus Belka, Guillaume Landry, Christopher Kurz
   Radiation Oncology 2022;17(1):21.
   https://doi.org/10.1186/s13014-022-01985-9

   The first author (author of this thesis) performed data preprocessing, trained all the networks (optimal hyperparameters and code implementation were provided by co-authors), calculated doses, analyzed the results, and wrote the manuscript.

2. **Patient-specific transfer learning for auto-segmentation in adaptive 0.35 T MR-gRT of prostate cancer: a bi-centric evaluation**
   Maria Kawula, Indrawati Hadi, Lukas Nierer, Marica Vagni, Davide Cusumano, Luca Boldrini, Lorenzo Placidi, Stefanie Corradini, Claus Belka, Guillaume Landry, Christopher Kurz.
   Medical Physics 2023;50:1573–85.
   https://doi.org/10.1002/mp.16056

   The first author contributed to the study design, collected data, performed data pre-processing, performed hyperparameter optimization, trained all networks, performed data analysis, and wrote the manuscript.

3. **Prior knowledge based deep learning auto-segmentation in magnetic resonance imaging-guided radiotherapy of prostate cancer**
   Maria Kawula, Marica Vagni, Davide Cusumano, Luca Boldrini, Lorenzo Placidi, Stefanie Corradini, Claus Belka, Guillaume Landry, Christopher Kurz
   Physics and Imaging in Radiation Oncology 2023;28:100498.
   https://doi.org/10.1016/j.phro.2023.100498

   The first author contributed to the study design, collected data, performed data pre-processing, code implementation, and hyperparameter optimization, trained all networks, performed data analysis, and wrote the manuscript.

4. **Personalized deep learning auto-segmentation models for adaptive fractionated magnetic resonance-guided radiation therapy of the abdomen**
   Maria Kawula, Sebastian Marschner, Chengtao Wei, Marvin Fernando Ribeiro, Stefanie Corradini, Claus Belka, Guillaume Landry, Christopher Kurz
   Medical Physics 2024
   `https://doi.org/10.1002/mp.17580`

   The first author contributed to the data collection and study design. The first author performed data pre-processing, code implementation, and hyperparameter optimization, trained all networks, performed data analysis, and wrote the manuscript.

Acronyms

AI artificial intelligence

BCE binary cross entropy
bSSFP balanced steady-state free precession

CBCT cone-beam computed tomography
CC correlation coefficient
CNN convolutional neural network
CT computed tomography
CTV clinical target volume

DDF dense displacement field
DIR deformable image registration
DL deep learning
DNA deoxyribonucleic acid
DSC dice similarity coefficient
DVH dose-volume histogram

EBRT external beam radiation therapy

FBP filtered back projection
FE frequency encoding
FID free induction decay
FL focal loss
FOV field-of-view

GPU graphics processing unit
GRE gradient echo
GTV gross tumor volume

HD Hausdorff distance
HU Hounsfield units

ICRU International Commission on Radiation Units and Measurements
IMRT intensity-modulated radiation therapy

LINAC linear accelerator

MC Monte Carlo
MI mutual information

ML machine learning
MLC multileaf collimator
MR magnetic resonance
MRgRT magnetic resonance-guided radiation therapy
MRI magnetic resonance imaging
MSD mean squared difference

NMR nuclear magnetic resonance
NN neural network

OAR organ at risk

pCT pseudo CT
PE phase encoding
PET positron emission tomography
PS patient-specific
PTV planning target volume

ReLU rectified linear unit
RF radio frequency
RTT radiation therapy technologist

SE spin echo
STN spatial transformer network

TPS treatment planning system

VMAT volumetric-modulated arc therapy

Chapter 1

Introduction

Cancer has been a significant global public health concern for many years [1]. The International Agency for Research on Cancer estimated nearly 20 million new incidences and 9.7 million deaths caused by cancer in 2022 [2]. Despite an increase in incidence, the mortality rate has been decreasing [1]. That can be attributed to the continuous advancements in diagnostics and therapy. The five most important techniques of cancer treatment, the so-called pillars, include surgery, chemotherapy, radiotherapy, immunotherapy, and molecularly targeted therapy [3]. These pillars are frequently combined rather than used in isolation, and about half of the patients receive radiotherapy in the course of their treatment [4].
Radiation therapy can be classified into two types based on the location of the radiation source. In brachytherapy, radioactive sources are placed inside the patient's body close to or inside the tumor, in order to minimize the healthy tissue exposure [5]. The second type is external beam radiation therapy (EBRT). It is a non-invasive technique that uses therapeutic radiation coming from outside the patient. It employs photons, electrons, protons, or heavy ions [6]. The most prevalent form of EBRT is intensity-modulated radiation therapy (IMRT) with photons at medical LINACs. In IMRT, the intensity of beams passing through the primary target volume is increased, while the intensity of beams passing through sensitive organs is decreased [7]. IMRT allows for high tumor volume coverage and more effective healthy tissue sparing. However, it is sensitive to setup uncertainties (e.g., patient positioning), anatomical changes (weight loss or gain), or physiological motions (breathing or peristaltic movements) and requires considerable safety margins around the target volumes [8].
A recent development addressing the shortcomings of IMRT is the implementation of image guidance and online treatment adaptation at combined MR-LINACs [9, 10]. They introduce daily MRI acquisition followed by quick plan adaptation to the current patient's anatomy. Additionally, the implementation of fast MRI sequences enables gated treatment, that is, pausing irradiation when the beam misses its target. MR-LINACs allow for smaller safety margins, which leads to two main benefits. Firstly, normal tissue is spared more efficiently, resulting in lower toxicity and fewer side effects [11]. Secondly, higher doses can be safely delivered to the tumor without exceeding limits in OARs. Thus, fewer irradiation sessions, i.e., fractions, are necessary while improving tumor control [12]. However, image-guided treatments also have some disadvantages that need to be considered. Besides the high costs of combined

MR-LINACs, the radiotherapy workflows are more laborious and time-consuming [13]. There is also an intrinsic trade-off involved in online adaptation. More precise adjustments improve treatment plan quality but prolong the time the patient has to spend on the treatment couch. In turn, the longer patients wait for the irradiation, the higher the likelihood of anatomical changes.
Accurate segmentation of organs and target volumes is an important step in the radiotherapy workflow. Yet, if performed fully manually, it can be very time-consuming. Long segmentation times are not the only concern of manual contouring. The inter- and intra-observer variability and the need for trained experts are other frequently reported issues [14]. Motivated by these disadvantages, many auto-segmentation algorithms have been developed over the years. There are methods based on image intensities, organ shape modeling, or using previously segmented images, i.e., atlases [15].
Several years ago, DL algorithms entered the stage of auto-segmentation. Two important breakthroughs significantly impacted this field: the convolutional neural networks (CNNs) [16] and the U-Net [17] architecture. By introducing convolutional layers, information encoded in the pixel intensities could be combined with their spatial location, which is crucial for image processing. The strength of the U-Net is that thanks to its encoder-decoder structure with skip connections, it captures both the coarse and fine image features. The DL algorithms were shown to be significantly faster and more accurate than previous segmentation methods [15, 18].
The research topic of this dissertation combines the fields of radiation therapy and deep learning. DL algorithms were studied for auto-segmentation of OARs and target volumes in medical imaging, especially for daily plan adaptation in MRgRT. The data in the study originated from prostate cancer patients and patients diagnosed with tumors in the abdominal area. An important new aspect of this work was the departure from the following typical DL assumption: neural networks must be trained on large and diverse datasets. This assumption is to ensure that the networks learn generic data features that enable them to perform well on new unseen examples. However, in fractionated MRgRT, the images on consecutive fraction days are highly correlated, and the same anatomical structures of the same patient must be repeatedly contoured. Therefore, adjusting the model to a given patient using segmented data from the beginning of the treatment may outperform the population models on subsequent treatment days. This has the potential to reduce the contouring workload for the physicians while ensuring that patient-specific delineation choices are consistently preserved, shorten treatment adaptation time, and consequently improve treatment outcomes.
This cumulative dissertation is structured as follows. In Chapter 2, the principles of computed tomography (CT) and MR imaging, as well as image registration, are presented. Chapter 3 describes the role of radiation therapy in cancer treatment and outlines the workflows of conventional EBRT and MRgRT. The main topic of Chapter 4 is DL and its applications in image segmentation and registration. Chapter 5 lists studies carried out in the scope of this thesis. The first study investigated CT auto-segmentation models for prostate cancer patients treated at conventional LINACs. Moreover, it examined the impact of DL-generated contours on optimizing dose distributions. In the remaining studies, the focus was on MRgRT at the MRIdian MR-LINAC installed at the LMU University Hospital. The second study compared two auto-segmentation strategies for prostate cancer patients: population-based and per-

sonalized DL models. The third study investigated an alternative approach to DL auto-segmentation relying on deformable registration of a previously segmented image followed by contour propagation. The fourth study brought new ideas to personalized auto-segmentation models. It investigated the impact of population models on personalized networks, analyzed progressive learning, and explored the possibility of single-image training. Chapter 6 summarizes the thesis's main findings and provides an outlook on possible future implementations and studies.

Chapter 2

Medical imaging

CT and MRI imaging play a crucial role in this thesis. This chapter introduces both modalities together with methods commonly used for medical image registration.

2.1 Computed tomography

This section introduces the basics of CT imaging. First, the photon-matter interactions and methods for X-ray generation are described. Subsequently, the main aspects of CT in the view of radiation therapy are presented. Finally, the filtered back projection is explained, which is the most common algorithm for CT image reconstruction.

2.1.1 X-ray interactions with matter

X-rays are a type of electromagnetic radiation commonly used in diagnostics and radiation therapy [19]. When X-ray photons are sent through a patient's body, they interact with the tissues and transfer their energy through photon-matter interactions. The three processes most relevant for therapeutic and diagnostic X-ray energies are the photoelectric effect, Compton scattering, and pair production [20, 21]. Figure 2.1 presents the regions where each of the three interactions dominates depending on the atomic number of the absorber and the initial photon energy.

In the photoelectric effect, a bound atomic electron absorbs a photon. A fraction of the photon's energy is used to ionize the atom, while the released electron carries away its remaining part in the form of kinetic energy. The photoelectric effect is the dominant form of interaction in biological tissues for photon energies up to tens of keV (see Figure 2.1). Therefore, it is more relevant for diagnostic imaging, where X-ray energies range from 70 to 140 keV [21]. Typical therapeutic beam energies lie between 6 and 15 MeV [24], with an average of approximately one-third of the maximum energy achieved at a given acceleration voltage. In this range, X-rays interact predominantly through the Compton interaction. In the Compton effect, photons passing through the medium scatter on electrons loosely bound to their atoms (i.e., the valence electrons) and transfer a fraction of their energy to these electrons. The third major process is pair production. It is particularly relevant for materials with high atomic numbers and energies above 10 MeV. The pair-production effect can occur when a photon with

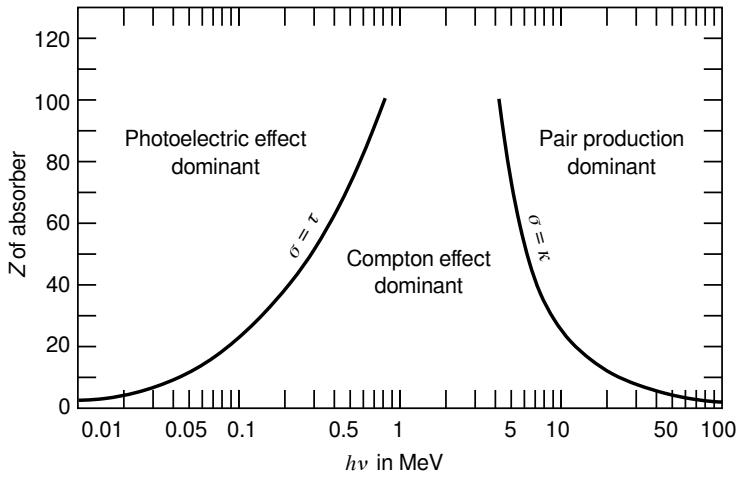

Figure 2.1: Three most relevant photon-matter interaction processes for diagnostic and therapeutic photon energies. The curves divide regions where the photoelectric effect, Compton scattering, or pair production are dominant depending on the photon energies $h\nu$ and the atomic number of the absorber Z. Typical Z values for human body tissues are below 10 [22]. Figure adapted from [23]

energy above 1.02 MeV passes close to an atomic nucleus. In the process, the photon is converted into an electron-positron pair.

2.1.2 Diagnostic X-rays generation

X-rays for medical purposes are produced by fast electrons colliding with materials of high atomic number. For energies used in diagnostics (mostly between 70 and 140 keV [21]), they are generated inside X-ray tubes, as shown in Figure 2.2. The X-ray tube cathode is heated to temperatures that allow the electrons to overcome their binding energy and be emitted from the filament [24]. The electrons released from the cathode are accelerated by the tube voltage and hit the anode at a high speed. Within the anode, X-rays are produced in two main processes that contribute to the X-ray tube's continuous and discrete energy spectrum. First, the electrons slow down within the anode by interacting with atomic nuclei, which leads to a continuous spectrum of bremsstrahlung radiation. Second, fast electrons interact with the inner-shell electrons of the anode material and kick them out of the atom. Electrons from the outer shells almost instantaneously fill up the created vacancy. The transition to the lower shell is accompanied by an emission of a photon with a characteristic energy. They appear on the X-ray tube spectrum as discrete sharp peaks, the so-called K, L, etc., lines. An exemplary X-ray tube spectrum is presented in Figure 2.3. The maximum energy of the X-ray spectrum is determined by the tube voltage, while its intensity scales with the cathode filament current.

2.1.3 Computed tomography imaging

CT is a quantitative volumetric imaging technique of crucial importance for radiation therapy [19, 27]. It is essential for dose calculation during treatment planning and is

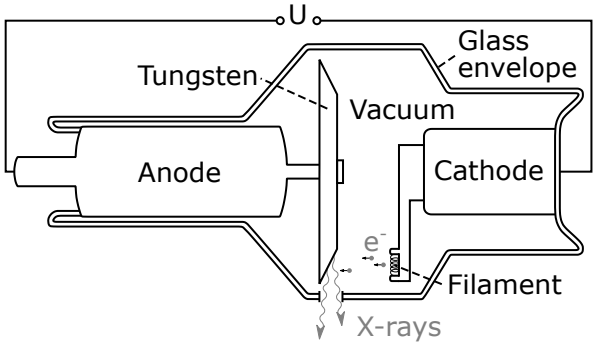

Figure 2.2: Schematic drawing of an X-ray tube. Electrons released from the heated filament accelerate thanks to the applied voltage U and hit the anode, generating X-rays. Figure adapted from [25].

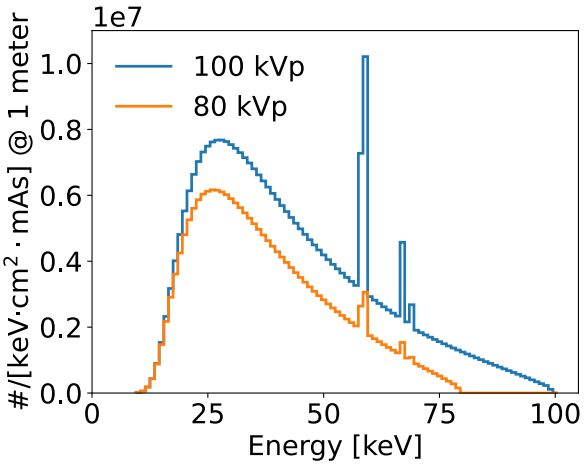

Figure 2.3: Energy spectra of an X-ray tube. Data points have been calculated using SpekCalc [26] setting peak energies to 80 and 100 keV, air thickness to 1 m, aluminum filter thickness to 1 mm, and a tungsten target tilt angle to 30°.

widely used for diagnostics as a fast and cost-effective imaging technique.

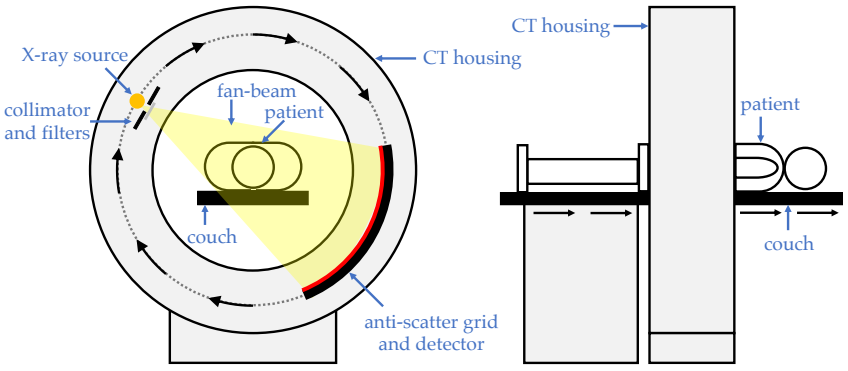

Figure 2.4: Scheme of a typical CT scanner with a patient couch. The X-ray source, collimator, and filters are on one side of the gantry, while the detector and anti-scatter grid are on the other side. As the gantry rotates around the patient the couch moves along the scanner axis. This results in a helical path of the X-ray source with respect to the patient. Figure courtesy of Dr. Moritz Rabe [28].

Figure 2.4 shows the design of a typical CT scanner. It consists of a patient couch and a circular gantry with an X-ray tube and an arc-shaped detector. A collimator and a beam hardening filter are installed in front of the X-ray tube. The filter aims to eliminate low-energy photons from the spectrum, which would otherwise be attenuated within the first centimeters of the patient. This would increase the imaging dose without improving image quality. The detector array with an anti-scatter grid is mounted on the opposite side of the gantry. Each detector element consists of a scintillator coupled with a photodiode. This combination converts the scintillation photons generated

Table 2.1: Typical CT values for several organs and tissues in the human body for photon energies of 120 kV [21].

Tissue/Organ	CT value [HU]
Air	-1000
Lungs	-900 to -500
Fat	-100 t0 -70
Water	0
Kidneys	20 to 40
Pancreas	20 to 50
Blood	30 to 60
Liver	40 to 70
Bones (marrow)	70 to 350
Bones (cortical)	350 to 2000

by the impinging X-rays into an electrical signal.
During a typical CT scan, the gantry rotates around the patient, and the couch moves continuously along the scanner axis, to cover the region of interest. The measured quantity is the amount of X-ray attenuation caused by a patient. To find it, the initial (I_0) and final (I) beam intensities are compared. According to Lambert-Beer's law:

$$I = I_0 \exp\left(-\int_\Gamma \mu(\mathbf{r})\mathrm{d}s\right), \tag{2.1}$$

where $\mu(\mathbf{r})$ is the attenuation coefficient, and Γ is the beam path. Since the measurement is repeated from different angles, i.e., the X-ray source rotates around the patient, it is possible to reconstruct the spatial distribution of $\mu(\mathbf{r})$. By convention, the resulting μ-values are represented in Hounsfield units (HU), which are defined as follows:

$$\mathrm{CT}(\mathbf{r}) = \frac{\mu(\mathbf{r}) - \mu_{\mathrm{water}}}{\mu_{\mathrm{water}}} \cdot 1000\ \mathrm{HU}, \tag{2.2}$$

where μ_{water} is the attenuation coefficient of water. The scale is chosen such that water is assigned a value of 0 HU and air a value of -1000 HU. Table 2.1 lists typical CT numbers for several organs and tissues in the human body. For historical reasons (using 12-bit integers), CT numbers typically range from -1024 to 3071 HU and cover all human tissues. With the help of a calibration curve, the HUs are converted to the tissue electron densities. This information is then used by the dose calculation algorithms [29].

CT provides quantitative information, short acquisition times, cost-effectiveness, high geometric accuracy, and spatial resolution. Yet, it also has disadvantages. The first drawback is the poor soft-tissue contrast. The CT value ranges overlap for organs such as kidneys, pancreas, or blood. The second drawback is the imaging dose in the order of several mSv per scan. Despite being relatively low, especially if compared to doses used in radiotherapy, each scan increases the probability of secondary cancer.
Like any other imaging technique, CT is prone to artifacts. High-Z materials like dental fillings, metal implants, fiducial markers, and calcifications can cause them. The

artifacts are visible in images as white stars covering the material location and the neighboring region. In addition, photon scattering, beam hardening, patient movement, and noise can further degrade the image quality.

2.1.4 Filtered back projection

The goal of CT image reconstruction is to find the object of interest based on the measured projections [21]. There are two main classes of reconstruction algorithms: iterative and analytical. The analytical methods are more widespread due to easier implementation and shorter computation times. The following paragraph describes the basis of the most commonly used analytical reconstruction method, i.e., filtered back projection (FBP) for a parallel beam geometry. Dedicated modifications for fan-beam or cone-beam geometries are necessary but beyond the scope of this work. They can be found in [19].

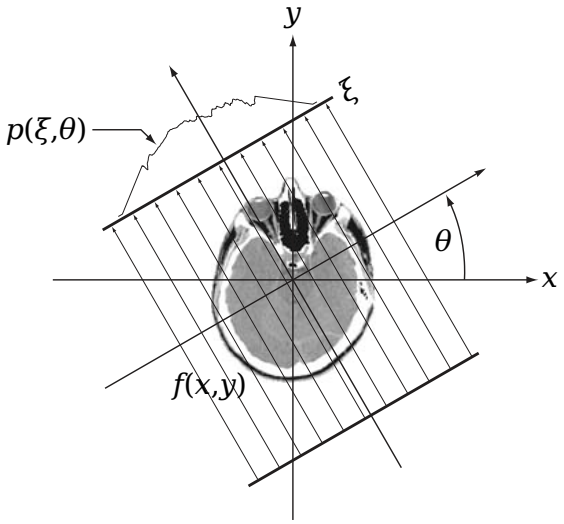

Figure 2.5: CT scan geometry. Figure adapted from [30]

In its simplest 2D form, the FBP assumes a parallel beam geometry and reconstructs the object slice by slice. Figure 2.5 shows a typical system geometry. Every beam path through the patient is represented by a line parameterized via angle θ and the distance from the rotation center $\xi = x\cos\theta + y\sin\theta$. The projection $p(\xi,\theta)$ can be calculated

$$p(\xi,\theta) = \iint \mathrm{d}x\mathrm{d}y f(x,y)\delta(x\cos\theta + y\sin\theta - \xi), \tag{2.3}$$

where $f(x,y)$ represents the image to be reconstructed. The one dimensional Fourier transform $P(u,\theta)$ of $p(\xi,\theta)$ with respect to ξ can be calculated as follows

$$P(u,\theta) = \int \mathrm{d}\xi p(\xi,\theta)e^{-2\pi i u\xi} = \iint \mathrm{d}x\mathrm{d}y f(x,y)e^{-2\pi i u(x\cos\theta + y\sin\theta)} = F(u_x,u_y) \tag{2.4}$$

where $u_x \equiv u\cos\theta$ and $u_y \equiv u\sin\theta$. It can be observed that the one-dimensional Fourier transform of projections w.r.t. ξ is equivalent to the two-dimensional Fourier transform of $f(x,y)$ in polar coordinates. This relationship is known as the Fourier

slice theorem.
In principle, it is possible to calculate $f(x,y)$ as the inverse Fourier transform of $F(u_x,u_y)$ using Equation (2.4). However, it would require resampling F data from the polar to Cartesian coordinates system in the Fourier domain, which could introduce strong artifacts to the image. To avoid these artifacts, an additional analytical step is carried out, i.e., Equation (2.4) is transferred back to the spatial domain:

$$f(x,y) = \int_0^{\pi} \mathrm{d}\theta \int_{-\infty}^{\infty} \mathrm{d}u|u|P(u,\theta)e^{2\pi iu(x\cos\theta+y\sin\theta)} = \int_0^{\pi} \mathrm{d}\theta \int_{-\infty}^{\infty} \mathrm{d}uK(u)P(u,\theta)e^{2\pi iu\xi} \tag{2.5}$$

where $K(u) = |u|$ is the ramp kernel. The final equation can be re-written by applying the convolution theorem, which states that the Fourier transform of the product of two functions is equivalent to the convolution of their Fourier transforms

$$f(x,y) = \int_0^{\pi} \mathrm{d}\theta p(\xi,\theta) * k(\xi) \tag{2.6}$$

where $k(\xi)$ is the inverse Fourier transform of the filter $K(u)$.

2.2 Magnetic resonance imaging

This section introduces the fundamental concepts of nuclear magnetic resonance (NMR) and MRI. First, the physics background on the spin, net magnetization, and their interaction with an external magnetic field is described. Following that, the focus is on the basic elements of MRI sequences and the description of the related Fourier space. This section ends with a brief description of common imaging artifacts.
Although NMR can occur for all nuclei with a non-zero spin, the focus in this work is on the hydrogen nucleus (a single proton) as it is commonly used in clinical MRI. A Cartesian coordinate system is used throughout this chapter. Following common convention, the main magnetic field $\vec{B}$ is assumed to be aligned with the z-axis, i.e., $\vec{B} = B_0\hat{z}$, where B_0 is its absolute field strength.

2.2.1 Spin

Spin angular momentum $\vec{S}$ or simply *spin*, is an intrinsic property of particles that plays a crucial role in the NMR phenomenon [21, 31]. It is a vector quantity with a magnitude S, which is connected to the quantum spin number s through the formula:

$$S = \sqrt{s(s+1)}\hbar, \tag{2.7}$$

where $\hbar \equiv h/2\pi$ denotes the Planck constant h divided by 2π. The quantum spin number s can only take a multiple of half-integer values. For the hydrogen nucleus, H^1, $s = 1/2$ and the z-component of the spin vector is $S_z = m_s\hbar$ where $m_s = \pm 1/2$.

All particles having a non-zero spin interact with an external electromagnetic field. To describe this interaction, a vector quantity called *magnetic dipole moment* $\vec{\mu}$ is introduced:

$$\vec{\mu} = \gamma\vec{S} \qquad\qquad z\text{-component: } \mu_z = \gamma S_z = \pm\frac{1}{2}\gamma\hbar, \tag{2.8}$$

where $\gamma = 2.675 \times 10^8 \mathrm{s}^{-1}\mathrm{T}^{-1}$ [32] is the proton gyromagnetic ratio, and the plus/minus sign indicates the parallel ("spin-up")/anti-parallel ("spin-down") orientation of the spin with respect to the magnetic field vector. The potential energy E associated with a spin placed in an external magnetic field $\vec{B}$ is

$$E = -\vec{\mu} \cdot \vec{B} \qquad \text{for } \vec{B} = B_0\hat{z} : \; E = -\mu_z B_0. \tag{2.9}$$

Combining Equation (2.8) and Equation (2.9) reveals an energy difference ΔE between the spin-up and spin-down configurations

$$\Delta E = \frac{1}{2}\gamma\hbar B_0 - \left(-\frac{1}{2}\gamma\hbar B_0\right) = \gamma\hbar B_0. \tag{2.10}$$

This energy level split is known as the *Zeeman effect*. The transition between the energy levels is accompanied by the emission or absorption of a photon with energy $\Delta E = \hbar\omega_0$ and frequency ω_0

$$\omega_0 \equiv \gamma B_0. \tag{2.11}$$

The frequency ω_0 is known as the *Larmor precession frequency* [33] and will be discussed further in the subsequent sections.

2.2.2 Magnetization, relaxation, and the Bloch equation

Despite NMR being a quantum mechanical phenomenon, it is possible to explain most of the concepts using classical mechanics [31, 34]. Moreover, it is more practical not to think of individual nuclei but rather consider their sum over a given volume V, which is referred to as *magnetization* $\vec{M}$

$$\vec{M} = \frac{1}{V}\sum_{\mu_i \in V} \vec{\mu}_i. \tag{2.12}$$

The volume V should be large enough to contain many protons but small enough to approximate any external fields to be constant within.
Let us consider a magnetization vector $\vec{M}$ placed in an external magnetic field $\vec{B}$. Assuming no additional external forces and no spin interactions, the evolution of the magnetization can be described as follows:

$$\frac{\mathrm{d}\vec{M}}{\mathrm{d}t} = \gamma\vec{M} \times \vec{B}. \tag{2.13}$$

Setting $\vec{B} = B_0\hat{z}$ leads to two decoupled equations for the longitudinal M_z and transversal $\vec{M}_\perp$ components of the magnetization vector [31]:

$$\begin{aligned} \frac{\mathrm{d}M_z}{\mathrm{d}t} &= 0 \Rightarrow M_z = M_z(t_0), \\ \frac{\mathrm{d}\vec{M}_\perp}{\mathrm{d}t} &= \gamma\vec{M}_\perp \times (B_0\hat{z}) \Rightarrow M_\perp = M_\perp(t_0)e^{-i\omega_0 t}, \end{aligned} \tag{2.14}$$

where t_0 is the initial time and $\omega_0 \equiv \gamma B_0$ is the Larmor precession frequency. According to Equation (2.14) the longitudinal magnetization component remains constant, while the transverse component precesses around the z-axis with the Larmor frequency.
In order to model interactions between nuclear spins and their environment modifications of Equation (2.14) are necessary. Let us assume that at time $t_0 = 0$ the magnetization vector (with initially $\vec{M} = M_0\hat{z}$) was entirely tipped to the transversal plane. To return to the state of minimal energy (alignment with the z-axis), protons transfer their excess energy to the environment (*lattice*). The speed of the process is directly proportional to the number of protons remaining in the transversal plane $M_0 - M_z$ and can be described as follows

$$\frac{\mathrm{d}M_z}{\mathrm{d}t} = \frac{1}{T_1}(M_0 - M_z) \Rightarrow M_z(t) = M_0(1 - e^{-t/T_1}), \tag{2.15}$$

where T_1 is the experimentally determined *spin-lattice relaxation time* [35].
There is another mechanism causing modifications to Equation (2.14). Each spin generates a small magnetic field, which alters the main magnetic field and leads to spatially dependent Larmor frequencies. Consequently, spins precess with slightly different frequencies and acquire different phases. The net transversal magnetization component diminishes over time according to the equation

$$\frac{\mathrm{d}\vec{M}_\perp}{\mathrm{d}t} = \gamma\vec{M}_\perp \times (B_0\hat{z}) - \frac{1}{T_2}\vec{M}_\perp \Rightarrow \vec{M}_\perp = \vec{M}_\perp(t_0)e^{-i\omega_0 t}e^{-t/T_2}, \tag{2.16}$$

where T_2 is the experimentally determined *spin-spin relaxation time* [35]. The two, Equation (2.15) and Equation (2.16) are known as the *Bloch equations* [36] that describe the evolution of the magnetization vector in the presence of an external magnetic field.
The dephasing process characterized by the T_2 constant is random in nature and thus irreversible. Apart from the T_2 processes, spin dephasing is accelerated by the magnetic field inhomogeneities characterized by the T_2' constant. Their sources include the main magnet imperfections or susceptibility-induced local field distortions. The total relaxation time T_2^* can be determined as follows

$$\frac{1}{T_2^*} = \frac{1}{T_2} + \frac{1}{T_2'}. \tag{2.17}$$

The hydrogen nuclei in various biological tissues exhibit different values of T_1 and T_2. Liquids such as blood and cerebrospinal fluid tend to have longer relaxation times than solid tissues like muscle, fat, or tendon. Thanks to these variations tissues in the human body can be differentiated.

2.2.3 MRI signal

When a magnetization vector $\vec{M}_0$ located in an external magnetic field $\vec{B}$ is tipped out of an equilibrium state it starts to precess [31, 34]. The precession results in a changing magnetic field that induces a voltage in a nearby conductor according to *Faraday's*

Law of Induction. The properties of the generated signal are impacted by human body composition, the strength of the external field B_0, and additional interactions with the magnetization.
The magnetization can be tipped with an additional oscillating magnetic field $\vec{B_1}$ perpendicular to $\vec{B_0}$ lasting a short time t_p, the so-called radio frequency (RF) pulse. If its frequency matches the Larmor precession frequency $\omega_0 = \gamma B_0$ (on-resonance condition), $\vec{M_0}$ is rotated away from the z-axis by an angle α

$$\alpha = \gamma B_1 t_{\mathrm{p}}. \tag{2.18}$$

The flip angle often gives the name to the applied pulse. For instance, a pulse that flips the entire net magnetization to the perpendicular plane is referred to as a $\pi/2$-pulse while the one that causes a 180° flip is called a π-pulse. Although $\pi/2$ and π pulses are frequently used to explain principles of MRI, other flip angles are possible. In fact, smaller flip angles are more common as they can speed up signal acquisition or increase its amplitude [37].
Three basic signal types can be obtained by an application of one or two RF-pulses: free induction decay (FID), gradient echo (GRE), and spin echo (SE) [38].

Free induction decay FID is one of the most fundamental signals in MRI [38]. After an RF-pulse flips the magnetization vector, its transversal component $\vec{M}_\perp$ starts to precess around the main magnetic field axis. The generated signal is strong initially but decays as the spins lose their coherence due to the T_2^* processes. The resulting FID signal is a sinusoidal wave with ω_0 frequency and exponentially decreasing amplitude as described by Equation (2.16) (with T_2 replaced by T_2^*).

Gradient echo The GRE is a modification of the FID [38]. First, a $\pi/2$-pulse is used to flip the magnetization vector. Second, a dephasing gradient is switched on to introduce known changes to the main magnetic field $\vec{B_0}$. This change in the main magnetic field alters the local frequency at which the spins rotate and accelerates the loss of coherence on the transversal plane. In the second step, a reverse gradient is applied to undo the previous dephasing. As a result, the spins get re-focused and a GRE is created.

Spin echo The SE begins with a $\pi/2$-pulse [38]. The transversal magnetization decays quickly due to the dephasing as observed in the FID signal. However, many of the T_2^* processes are reversible. Flipping the magnetization by a π-pulse effectively ”inverts” the time. The spins can refocus again and generate an SE signal.

2.2.4 Spatial encoding

To reconstruct an MR image the acquired signal must be connected to the voxels that produced it. Two primary techniques make it possible, namely frequency encoding (FE) and phase encoding (PE) [31].

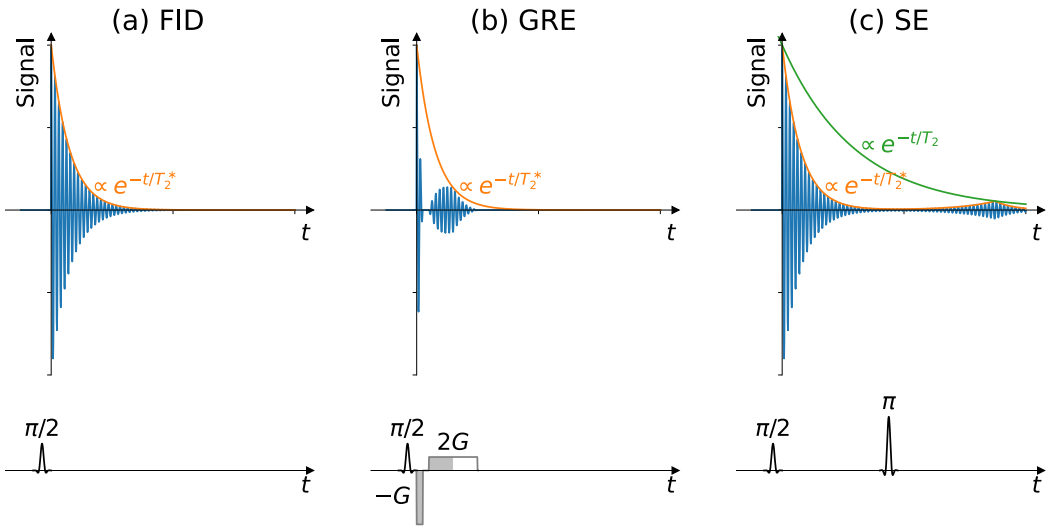

Figure 2.6: Free induction decay (FID), gradient echo (GRE), and spin echo (SE) signals. (a): After a $\pi/2$ excitation pulse, the FID signal decays as $\exp(-t/T_2^*)$. (b): Shortly after the $\pi/2$ excitation pulse, a dephasing gradient G is applied to diminish the transversal magnetization component in a controlled way quickly. After reversing the gradient polarity, a GRE is formed. (c): The SE is generated at the echo time T_E by two RF pulses: a $\pi/2$- pulse applied at time t=0, and a π-pulse applied at time $t = T_\mathrm{E}/2$. All signals were plotted using the same T_2 and T_2^* values. T_E was set to approximately $3T_2$. Sketch inspired by figure 11.10 in [21].

Frequency encoding FE is implemented by adding a magnetic field gradient $\vec{G}_f$ to the main magnetic field, which changes $\vec{B}_0$ by a few percents. As a result, the spin resonance frequency ω depends on the spatial position. Assuming that the gradient varies linearly with position $\vec{r}$ the precession frequency is given by [31,34]

$$\omega(\vec{r}) = \underbrace{\gamma B_0}_{\omega_0} + \gamma\vec{r}\cdot\vec{G}_f. \tag{2.19}$$

This method is utilized for slice selection (i.e., z-coordinate encoding) by applying an FE-gradient along the $\vec{B}_0$ simultaneously with the first RF-pulse. By setting the frequency of the RF-pulse to $\omega(z_i)$, only spins located at z_i are flipped to the transversal plane contributing to the acquired MRI signal.
The same method can be used for in-plane localization (x- or y-position within the slice; let us assume the former without loss of generality). Applying a gradient along the x-axis during the signal readout alters the spin precession frequency according to Equation (2.19). As a result, spins at position x_i produce a signal of frequency $\omega(x_i)$. To determine the third coordinate (y-position) a method called PE is typically used.

Phase encoding The phase and frequency encoding gradients both alter the local spin precession frequency. The PE gradient G_p is turned on along the y-axis for duration t_p only. After this gradient is switched off protons return to their original precession frequency with different accumulated phases. For a linearly varying PE gradient, the phase difference is [31,34]

$$\Phi(y) = \gamma G_\mathrm{p}(y) t_\mathrm{p}. \tag{2.20}$$

Repeating the PE step multiple times with different gradient strengths G_p leading to different phase shifts can differentiate between y-positions in the image.

2.2.5 Imaging sequences

An MRI sequence is a combination of RF pulses and magnetic field gradients applied in a particular order [21,31,38]. Various imaging contrasts can be achieved by manipulating the timing, duration, and order of the RF pulses and gradients. Two frequently used parameters are the echo time T_E and repetition time T_R. The echo time specifies the time between the excitation pulse and the signal's peak generated in the receiver coils, while the repetition time is the period between two successive excitation pulses.

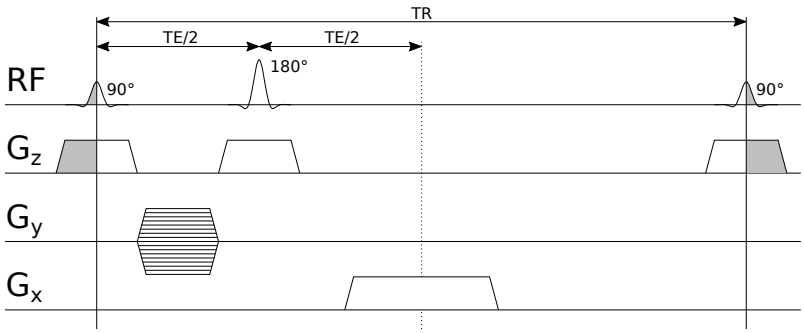

Figure 2.7: Schematic representation of the spin echo sequence. During the application of two radio frequency (RF) pulses, 90° and 180° at times 0 and TE/2, respectively, the slice selection gradient G_z along the z-axis is switched on. The phase encoding gradient G_y is switched on between the RF pulses. To read out the signal at time TE, the frequency encoding gradient G_x is switched on. The sequence repeats after the repetition time TR.

For an exemplary SE sequence shown in Figure 2.7 proper selection of T_E and T_R determines the contrast or "weighting" of the resulting image. The dependence of the SE signal intensity S on time constants and proton density [H] can be approximated by [21]:

$$S \propto [\mathrm{H}](1 - e^{-T_R/T_1})e^{-T_E/T_2}. \tag{2.21}$$

Long T_R allows all spins to recover between excitation pulses, minimizing T_1 effects. Short T_E reduces the impact of T_2 processes since there is not enough time for T_2 decay to differentiate between tissues. Consequently, setting both T_R and T_E to be short/long results in T_1-weighted/T_2-weighted images. Choosing a long TR and short TE can suppress the impact of both T1 and T2, resulting in proton density weighting.

2.2.6 Fourier space in image processing

According to the Fourier series theorem, any function $f(t)$ that is periodic with period T can be represented as a sum of sinusoidal terms having fundamental frequency ω

and complex amplitude a_n [30]

$$f(t) = \sum_{n\in\mathbb{Z}} a_n e^{i\omega_n t}, \qquad a_n = \frac{1}{T}\int_{-T/2}^{T/2} f(t)e^{-i\omega_n t}\mathrm{d}t \tag{2.22}$$

where $\omega_n \equiv 2\pi n/T$. Taking the limit of infinite periodicity $T \to \infty$ leads to:

$$f(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} F(\omega)e^{-i\omega t}\mathrm{d}\omega, \qquad F(\omega) = \int_{-\infty}^{\infty} f(t)e^{i\omega t}\mathrm{d}t. \tag{2.23}$$

The equation on the left is known as the *inverse Fourier transform*, while the one on the right is the *Fourier transform* [30].
In 2D imaging, the time t and the temporal frequency ω are replaced by the position (x, y) and spatial frequency (k_x, k_y), respectively, while $f(x, y)$ is a function describing pixel intensities [21, 31, 34]. A common representation of a two-dimensional Fourier space, also known as *k-space*, is an array with x and y axes representing the k_x and k_y frequencies. Each array entry represents the contribution of a given frequency to the final image. In the k-space, the low spatial frequencies correspond to the rough shapes of the imaged objects, while the high frequencies encode fine-grain details and sharper edges.
Let us consider a simple FID signal from a two-dimensional object of proton density $\rho(x, y)$ while a FE gradient G_x is switched on and the PE gradient was applied prior to that for a short time t_{pe}. At time t, the signal comes from all spins [31, 34]

$$S(t) = \iint \rho(x,y)e^{i\gamma(G_x t x + G_y t_{\mathrm{pe}} y)}\mathrm{d}x\mathrm{d}y = \iint \rho(x,y)e^{i(k_x x + k_y y)}\mathrm{d}x\mathrm{d}y, \tag{2.24}$$

where $k_x = \gamma G_x t$, and $k_y = \gamma G_y t_{\mathrm{pe}}$. A comparison of Equation (2.23) and the measured signal shows that the latter is the Fourier transform of the proton density across the object, which is often the quantity of interest in MRI. The sampled signal data can directly fill one line (for a given k_y) of the k-space matrix.

2.2.7 MR artifacts

MRIs are susceptible to various artifacts that degrade their quality [31, 34]. These can be grouped into three categories: tissue, motion, and acquisition-related distortions.

Tissue-related artifacts The first group includes the chemical shift and susceptibility artifacts. The former is attributed to slight variations in the local resonance frequencies caused by different nuclei being shielded differently by their environment from the externally applied magnetic field (e.g., fat and water molecules). As a consequence, the signal from some protons gets mislocalized. The susceptibility artifact is caused by substances altering the local magnetic field (dia, para, and ferromagnetic substances), leading to mismapping or even loss of the MRI signal [39].

Motion-related artifacts Motion artifacts in the form of discrete "ghosts" appear when the imaging structures move periodically in the field-of-view (FOV) [34]. Non-periodic movements result in diffused noise rather than ghosts. Motion artifacts are more severe along the phase encoding direction as the time necessary to cover the entire phase encoding space might take several minutes. A single echo (encoding one line in the frequency encoding) is much shorter than the movement itself.

Technique-related artifacts Many technique-related artifacts are caused by the fact that the acquired signal is discretized and the image to reconstruct is pixelated. Artifacts may occur when the Nyquist sampling theorem [30] is violated, the FOV is set smaller than the object dimensions, or because only a finite number of frequencies can be sampled from the Fourier domain [34]. They can also be related to hardware issues like insufficient shimming, coil heating, or eddy currents. Low-frequency intensity variations throughout the image, i.e., bias fields, occur when the signal is partially attenuated by tissues on its way to the coil. Due to gradient nonlinearities, the MRI signal generated by elements located further from the scanner center is assigned to more central grid voxels. As a result, the reconstructed object appears smaller, as shown in Figure 2.8 [40].

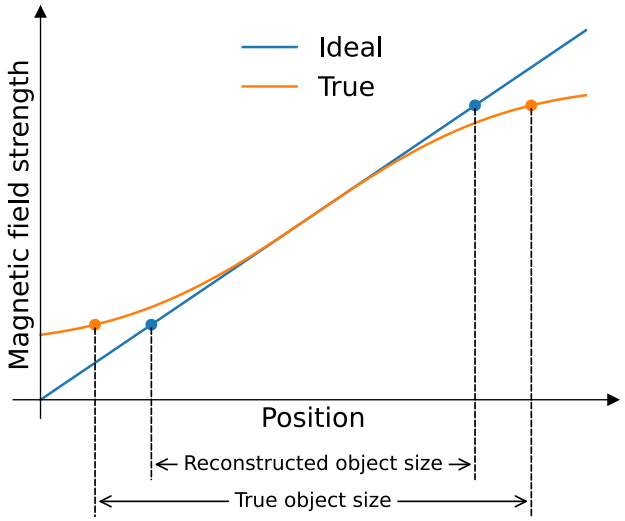

Figure 2.8: An illustration depicting the nonlinearity of the magnetic field gradient within an MRI scanner. A linear function can well approximate the gradient at the scanner center. However, the gradient coils generate non-linear magnetic fields close to the edges of the scanner. If not taken into account, it might lead to incorrect encoding of the edge pixels and make the reconstructed object appear smaller.

2.3 Image registration

The aim of image registration is to find a geometric transformation that aligns two images of interest. Registration has a wide range of applications in radiotherapy. Exemplary applications include combining information from different modalities, moni-

toring anatomical changes over time, dose accumulation, pseudo CT (pCT) generation, or using templates/prior images for contour propagation [41,42].

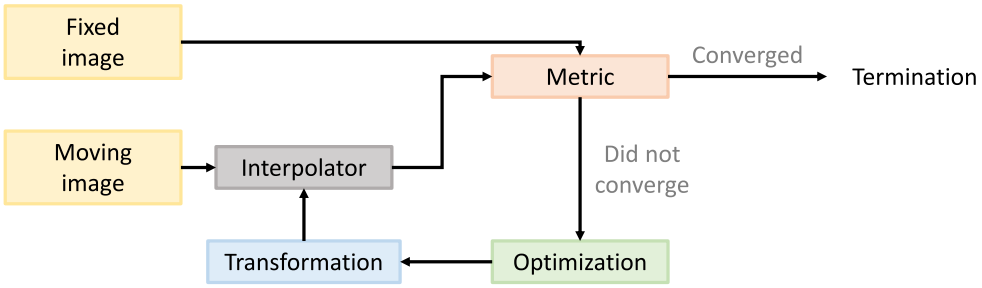

Figure 2.9: Typical workflow of iterative image registration. Figure adapted from [43].

Figure 2.9 presents a general scheme of iterative image registration. The inputs are the two images to be registered: a fixed (also known as static, reference, or target) and a moving (also known as floating) image. In the first step, the moving image is transformed and interpolated onto the static image grid, and a similarity metric is calculated between them. Subsequently, the optimizer updates the transformation parameters, and the process repeats until the metric reaches a certain limit or stops improving.
One distinguishes between parametric and non-parametric image registration [44]. In the non-parametric case, a dense displacement field that contains per-point information where the point should move in 3D is predicted directly. This method comes with a high computational cost due to the high number of degrees of freedom. In parametric image registration, a deformation field is described by a parametric model. The optimizer only needs to optimize the parameters of the model. This thesis used only conventional parametric or DL-based registrations; therefore, the non-parametric types are not described further.

2.3.1 Transformation

Rigid registration is a simple transformation model with only six degrees of freedom in 3D: three for translation and three for rotation. Rigid registration is used in radiotherapy to position the patient as the treatment couch can be shifted accordingly. Adding three degrees of freedom for scaling and three for shearing results in an affine transformation [45].
Models more complex than affine transformation are needed to describe anatomical changes. A common deformable image registration method uses B-splines. B-splines are basis functions of minimal support. They can effectively model local changes but are rather inefficient in predicting global changes. For that reason, B-spline registration is commonly preceded by an affine rigid alignment [43]. To set-up a B-spline registration, a uniform grid of $n_x \times n_y \times n_z$ control points $\phi_{i,j,k}$ and (s_x, s_y, s_z) spacing is placed over the image, where $i = \lfloor x/s_x \rfloor - 1$, $j = \lfloor y/s_y \rfloor - 1$, and $k = \lfloor z/s_z \rfloor - 1$. The transformation T_{DIR} [43] is defined as a tensor product of 1D cubic B-splines:

$$T_{\text{DIR}}(x,y,z) = \sum_{l=0}^{3}\sum_{m=0}^{3}\sum_{n=0}^{3} B_3^l(u) B_3^m(v) B_3^n(w) \phi_{i+l,j+m,k+n} \tag{2.25}$$

where $u = x/s_x - i + 1$, $v = y/s_y - j + 1$, $w = z/s_z - k + 1$, and

$$\begin{aligned} B_3^0(t) &= (1-t)^3/6, & B_3^1(t) &= (3t^3 - 6t^2 + 4)/6, \\ B_3^2(t) &= (-3t^3 + 3t^2 + 3t + 1)/6, & B_3^3(t) &= t^3/6. \end{aligned} \tag{2.26}$$

The $\phi_{i,j,k}$ (in total $3 \cdot n_x n_y n_z$ parameters) are optimized during the iterative registration process.

2.3.2 Metric

Metrics quantify the alignment of two images and can serve as objective functions during registration. Metrics can be feature- or intensity-based. Feature-based metrics require defining a set of points, lines, structures, etc., on both images prior to the registration. Conversely, intensity-based metrics require no pre-processing [43]. The most frequently used intensity-based metrics include mean squared difference (MSD), correlation coefficient (CC), also known as normalized cross-correlation, and mutual information (MI). If the fixed (F) and moving (M) images have intensities in the same range, a MSD metric [43] can be used

$$\mathrm{MSD}(F, M) = \frac{1}{N} \sum_i^N [F(\vec{x}_i) - T(M(\vec{x}_i))]^2 \tag{2.27}$$

where, T is the registration transformation, and N the number of pixels. If the intensities are in different ranges but are related by a linear mapping, CC [43] can be employed

$$\mathrm{CC}(F, M) = \frac{\sum_i^N (F(\vec{x}_i) - \bar{F})(T(M(\vec{x}_i)) - \bar{M})}{\sqrt{\sum_i^N (F(\vec{x}_i) - \bar{F})^2 \sum_i^N (T(M(\vec{x}_i)) - \bar{M})^2}} \tag{2.28}$$

where $\bar{M}$ and $\bar{F}$ are average intensities of M and F. MSD and CC are commonly used for intra-modality registration. For inter-modality registration MI [43] can be employed, that is based on probability theory and statistics:

$$\mathrm{MI}(F, M) = \sum_f \sum_m p_{FM}(f, m) \log \frac{p_{FM}(f, m)}{p_F(f) p_M(m)} \tag{2.29}$$

where $p_I(i)$ is the probability of the i^{th} intensity value occurring in image I, and $p_{FM}(f, m)$ is the joint probability of pairs of image values occurring together [43]. More information on registration metrics can be found in [30, 42, 43].

2.3.3 Regularization

When performing non-rigid image registration, transformations that yield accurate similarity metrics may result in deformations that are physiologically impossible, such as tissue tearing or folding [43]. To address this problem, a regularization term $\mathcal{L}_{\text{reg}}$ is

added to the objective function $\mathcal{L}$ to restrict the allowed transformation to physiologically meaningful solutions:

$$\mathcal{L}(u) = \mathcal{L}_{\text{sim}}(u) + \lambda \mathcal{L}_{\text{reg}}(u) \tag{2.30}$$

where u is the deformation field, and λ is a weighting parameter. The regularization term takes into account the existing knowledge of the problem. For example, to favor smooth deformations $\mathcal{L}_{\text{reg}}(u) = ||\nabla u||^2$, could be employed. To favor small local deformations, elastic regularization based on the Navier equation [43, 46, 47] from elasticity theory could be used. Another commonly used regularization term is the Jacobian determinant. Enforcing its positivity ensures no unphysical folding of the deformation field.

2.3.4 Optimization

The optimization problem can be stated in the following way: for a given transformation model T and objective function $\mathcal{L}(u)$, find parameter values of T that, by convention, minimize the objective function [43]. A typical problem in radiotherapy image registration requires iterative estimation of optimal values. Optimization algorithms suitable for rigid registration include the (conjugate) gradient descent [48], quasi-Newton [49], or Nelder and Mead methods [50].

Chapter 3

Radiation therapy

This chapter outlines the basic principles of the EBRT. It begins with a description of the physical and biological processes that make radiation useful for cancer treatment. Subsequently, a general structure of conventional and MRgRT workflows is presented. Both workflows depend on the department infrastructure, available treatment planning system (TPS) software, institutional guidelines, staff experience, and tumor location. Therefore, there might be slight variations in them among different institutions.

3.1 Physics and biology of radiation

The term *radiation* describes the emission and propagation of energy through space. Typical therapeutic photon beams have energies of 6MV [20, 21, 24], and their interactions with matter have been described in Section 2.1. Most of the beam energy is not deposited directly by the photons but is transferred to the electrons of the medium (i.e., human body tissues). The secondary electrons deposit their energy in multiple inelastic scattering events with the medium electrons. In medical practice, it is desirable to correlate the energy absorbed by the human body with a clinically observed effect. Therefore, the concept of dose D has been introduced, as the energy imparted by ionizing radiation to matter per unit mass [21]:

$$D = \frac{\mathrm{d}\bar{\mathcal{E}}}{\mathrm{d}m} \tag{3.1}$$

where $\bar{\mathcal{E}}$ represents the mean energy of ionizing radiation absorbed by a medium with mass m. The unit of absorbed dose is the gray (Gy), defined as the absorption of one joule of radiation energy per kilogram of matter: $1\,\mathrm{Gy} = 1\,\mathrm{J{\cdot}kg^{-1}}$. Figure 3.1 shows several depth-dose curves for photon beams with therapeutic energies.

One of the most common explanations for radiation-induced cell death is double-strand breaks in the deoxyribonucleic acid (DNA). Double-strand breaks are frequently impossible for the cell to repair. Without complete genetic information, cells cannot proliferate properly and die [51]. Conventional radiation therapy is typically divided into a number of irradiation sessions, i.e., fractions, over several weeks [21]. This fractionation is motivated by radiobiology's four "R's": repair, redistribution, repopulation, and reoxygenation [52, 53]. Repair of non-critical radiation damages is usually

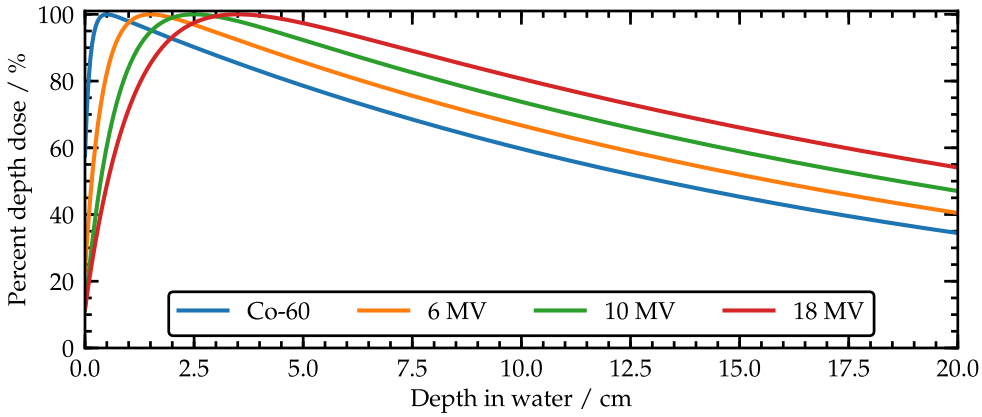

Figure 3.1: Depth-dose curves of photon beams in water with a field size of 10 cm×10 cm and source-to-surface distance of 100 cm. Each line demonstrates the percentage of dose deposited by a beam of photons (cobalt-60 beam and three photon beams produced by LINACs) relative to its maximum value as a function of depth in water. Figure courtesy of Dr. Moritz Rabe [28].

faster in normal tissues than in tumors. Therefore, fractionation provides a recovery time that the former can use more efficiently. Redistribution occurs for fast-cycling cells (mostly tumor cells) that have enough time between fractions to enter mitosis, which is a radiation-sensitive phase. On the other hand, slow-cycling cells (mostly in healthy tissues) remain longer in the initial cycle phases, which are more radiation-resistant. Intervals between fractions provide time for oxygen-deprived cells to re-oxygenate, enhancing their radiosensitivity. The last process, repopulation, causes tumor cells to proliferate faster after irradiation and is therefore considered one of the major factors leading to the failure of the treatment.

3.2 Conventional radiotherapy

3.2.1 Linear accelerators (LINACs)

Figure 3.2 and Figure 3.3 show the design of a typical LINAC used in radiotherapy [21]. In the electron gun, electrons are emitted from a heated anode and are pre-accelerated in a static electric potential. The electrons are accelerated further inside a waveguide by absorbing electromagnetic energy delivered by the RF power generator. When reaching the LINAC head, the electrons are directed toward the target to produce X-rays via bremsstrahlung. Subsequently, the generated beam passes through the primary collimator, (optional) flattening filter, ionization chamber, and secondary collimator. Before reaching the patient, the last component is the multileaf collimator (MLC), which shapes the beam into small segments.

3.2.2 Conventional radiotherapy workflow

Planning imaging A conventional radiotherapy workflow begins with the acquisition of planning images. They might include CT, MRI, and positron emission tomography

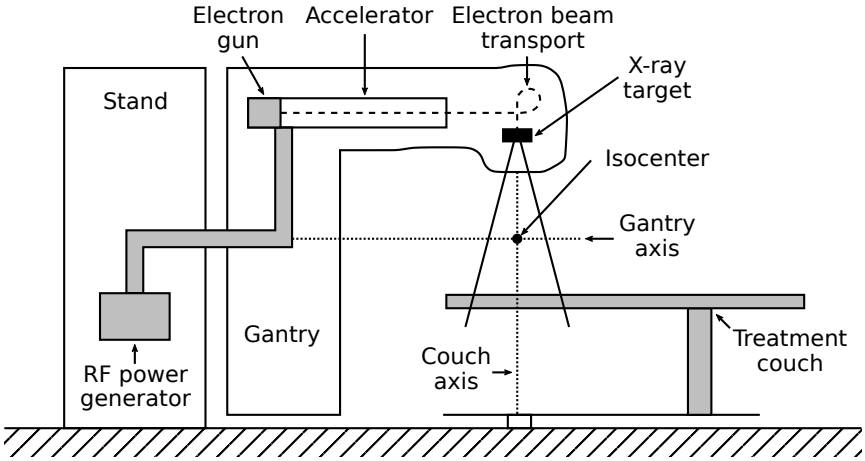

Figure 3.2: Schematic drawing of a linear accelerator (LINAC). Electrons are emitted from the electron gun, accelerated in the electromagnetic field, and transported towards the LINAC head. As electrons collide with the X-ray target, they generate therapeutic radiation. The LINAC arm can rotate around the treatment couch to deliver radiation from any angle. Figure re-drawn from [20].

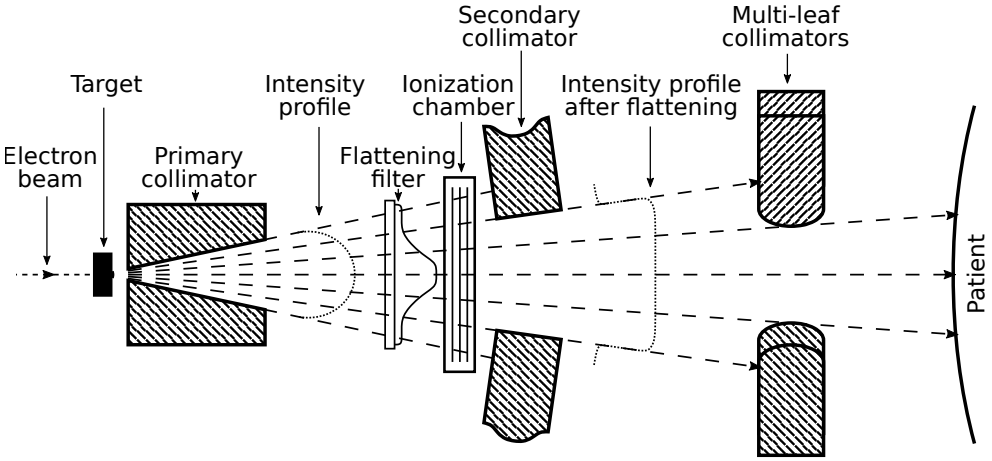

Figure 3.3: Schematic drawing of a LINAC head. A photon beam is produced by fast electrons that decelerate inside the X-ray target. The beam passes through the primary collimator, flattening filter, ionization chamber, secondary collimator, and multi-leaf collimator to finally reach the patient. Figure re-drawn from [21].

(PET) modalities, each of which offers distinct benefits [20,21]. The CT is mandatory to obtain the electron density in the human body, which is essential for dose calculation [19]. MRI is useful for segmenting tumor volumes and OARs due to its superior soft-tissue contrast. Finally, PET [54] images help to localize target volumes by highlighting regions of higher metabolic activity.

Target volume and organs at risk contouring Structures necessary for treatment planning can be segmented as soon as the planning images are available. In the following paragraph, the most relevant target volumes and the concept of OARs are described based on the International Commission on Radiation Units and Measurements (ICRU) guidelines described in Report 83 [55].
The gross tumor volume (GTV) is defined as the gross demonstrable extent and location of the tumor. Multiple GTVs might be defined simultaneously for one patient to cover the primary tumor, metastatic nodes, and metastasis.
The CTV is defined as a volume including the GTV and its surrounding areas with a certain probability of containing tumor cells not visible on the acquired images. In cases where irradiation occurs after complete surgical tumor resection, only the CTV encompassing the tumor bed is delineated. Conversely, benign tumors might not require a CTV due to the low probability of tumor cell infiltration into neighboring tissues. There is no universal definition of probability thresholds for delineating the CTV. The contouring is based on institutional guidelines and is influenced by physicians' experience, which leads to significant inter-observer variability. In addition, the same person delineating the same structure repeatedly does not typically deliver the same contour (intra-observer variability). Figure 3.4 shows an example of inter-observer variability in CTV segmentation for pancreatic cancer.
The planning target volume (PTV) is a geometrical concept introduced to ensure sufficient dose coverage of the CTV while accounting for possible positioning and treatment delivery uncertainties, patient motion, and anatomical changes. The PTV contour is created by expanding the CTV by a safety margin. The work described in [56] investigated the minimum margin size necessary to ensure a certain probability of CTV coverage. The recommendations derived from this study were adopted by many clinics and adjusted later based on the tools available for patient positioning, dose delivery techniques, the presence or absence of image guidance, motion compensation techniques, quality assurance, and the feasibility of plan adaptation.
The OARs are defined as structures that could suffer significant damage if excessively irradiated and for that reason should be considered during treatment planning. The selection of OARs for a given treatment depends on the prescribed dose and the PTV location. For particularly large structures, such as the intestines, spinal canal, or aorta, only their sections located in the PTV proximity might be considered as OARs.

Dose prescription and fractionation A higher total dose can be administered to the tumor thanks to fractionation while keeping the risk of normal tissue complications reasonably low [21]. A typical fractionation scheme at a conventional LINAC assumes administering 2 Gy five times per week over several weeks [58]. To assess the effectiveness of a given fractionation scheme, the empirical linear-quadratic model is frequently

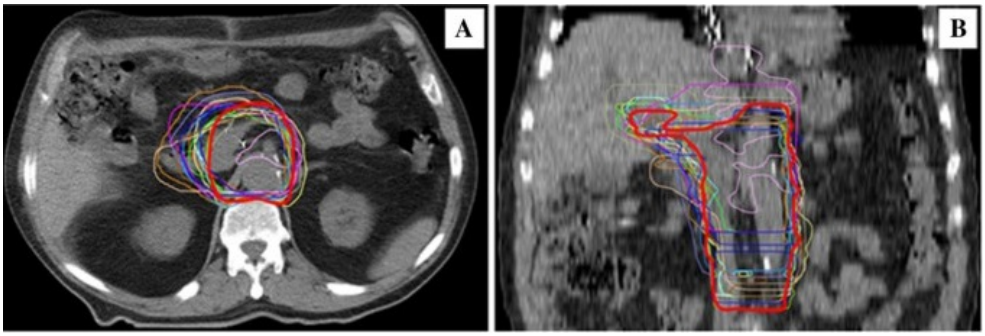

Figure 3.4: Graphical representation of inter-observer variability in CTV segmentation. Each of the 18 thin lines corresponds to a single-observer delineation, while the thick red line represents the delineation defined by a national reference radiation oncology center. Figure and caption adapted from [57]

used [59]:

$$S(D) = e^{-\alpha D - \beta D^2} \tag{3.2}$$

where $S(D)$ is the fraction of cells that survive a given dose D, while α and β are cell-dependent parameters describing a cell's radiosensitivity. The model can be intuitively explained by two mechanisms causing DNA damage [60]. The linear term is attributed to lethal damages caused by single-hit events, while the quadratic term describes lethal damages occurring after multiple sub-lethal hits. Sub-lethal damages do not accumulate effectively for tumors with a high α/β ratio, resulting in a low increase in cell killing per unit dose. High α/β ratio tumors are excellent candidates for fractionation, as delivering the dose over several days reduces the risk of normal-tissue complications without compromising tumor control. Low α/β ratios characterize tumors where sub-lethal hits accumulate more effectively and the therapy efficiency is enhanced with an increased fraction dose [60, 61].
Numerous empirical studies have been conducted to formulate guidelines recommending PTV doses necessary to achieve successful local tumor control. Moreover, each OAR was assigned dose values that it could tolerate [62]. Constraints may include the maximum dose to an OAR or the maximum acceptable relative OAR volumes receiving a dose above certain thresholds. These constraints are determined based on the OAR radiosensitivity and its type. For serial organs, like the spinal cord or rectum, an excessive dose delivered to a single point might cause serious injuries and, therefore, should be avoided. Parallel organs (e.g., lungs) would tolerate a single-point excessive dose but suffer more from a too-high dose delivered to a certain relative volume of the organ [55].

Treatment plan optimization The next step in the radiation therapy workflow is treatment plan optimization. There are two possibilities: forward and inverse planning. In the forward planning approach, the operator selects the beam angles and collimator settings through trial and error until a desired dose distribution is achieved. The inverse-planning approach begins with defining an objective function containing dose constraints and objectives indicated by a physician [63]. The goal is to optimize the beam properties and collimator settings that lead to a dose distribution that sat-

isfies the physician's prescription. This is the principle of IMRT [64] and its variation volumetric-modulated arc therapy (VMAT) [65]. In VMAT a continuous beam is delivered as the LINAC arm rotates around the patient and the MLC shape is continuously adapted. IMRT can also be realized in a step-and-shoot fashion, where the dose is only delivered at certain angles, pausing the radiation when the LINAC moves to the next position. Plan optimization is computationally intensive and requires experience to balance competing objectives and constraints. Certain parameters, such as gantry angles, collimator settings, and couch angles, are set beforehand. Other parameters need to be optimized. The latter include the MLC configurations (called segments) and the number of monitor units to be delivered in each of these configurations.
To reduce the complexity of dose optimization, it can be calculated on a coarser grid than the acquired planning CT. Methods used to calculate dose distribution include kernel-based (e.g., collapsed cone [66]) or Monte Carlo (MC)-based algorithms [67]. The kernel-based methods consider the transport of the primary photons separately from the interactions of the secondary electrons. In MC methods, particle paths and physical processes are simulated based on interaction cross-sections. Therefore, it is possible to determine the amount of energy deposited in each interaction at a given point. The deposited energies can be integrated to obtain the desired dose distribution. The MC techniques are more accurate but computationally more intensive. Many modern treatment planning systems use graphics processing unit (GPU)-based fast MC engines that produce accurate results in just a few minutes [20, 21].
To evaluate treatment plan quality or compare multiple plans graphically, the dose-volume histograms (DVHs) can be calculated [20]. These are cumulative dose-volume frequency distribution plots, i.e., they show the percentage of each organ's volume receiving at least a given dose. It is also beneficial to overlay the calculated dose with the CT to search for cold and hot spots that might indicate under- and over-dosage of certain regions.

Dose delivery The irradiation can begin once a radiation oncologist approves the treatment plan. The patient is placed in a treatment position that is the same as the one from the simulation day. For optimal positioning, a laser system is installed in the treatment room, and various immobilization devices (thermoplastic masks, abdominal compression tools, etc.) are applied. A cone-beam computed tomography (CBCT) scan can be performed before the first treatment and periodically thereafter for increased positioning accuracy [68]. For sites where tumor shrinkage or weight loss could affect the dose distribution significantly (e.g., head and neck region), a CBCT analysis might indicate a need for a new simulation scan and treatment plan (offline treatment adaptation) [68]. Finally, the dose can be administered according to the treatment plan. Various techniques might be employed to account for patient movements during irradiation. The methods include surface guidance, stereoscopic X-ray imaging [69], and, ideally, gated beam delivery techniques [70]. In the latter, the radiation is only delivered when the target or a surrogate is at a known location, e.g., only during breath holds.

3.3 Online adaptive image-guided radiation therapy

The treatment plan for the entire course of conventional radiation therapy is based on the planning image acquired several days prior to the first irradiation. However, there is a high likelihood of inter- and intra-fractional anatomical changes, positioning errors, or patient motions during treatment. More precisely, they might include tumor shrinkage, weight gain or loss, differences in organ fillings (stomach, bowel, rectum, bladder), respiratory, and gastrointestinal movements. In conventional radiotherapy without image guidance, the treatment plan must account for the uncertainties, i.e., the CTV-to-PTV margin must be large enough to ensure full CTV coverage [56]. However, the likelihood of normal tissue complications increases for large PTV margins [71,72]. The daily dose must be small enough to allow the OARs to repair radiation-induced damages. This increases the number of fractions required to deliver the ablative dose.
Introducing image guidance with treatment adaptation can alleviate many conventional radiotherapy problems. Firstly, daily pre-treatment imaging allows for accurate patient positioning and dose tailoring to the current anatomy (online treatment plan adaptation). Secondly, imaging during irradiation captures breathing and gastrointestinal motions, allowing the treatment to be paused when the beam misses the target (gating) [73]. Image guidance and treatment adaptation enable the reduction of the CTV-to-PTV margin and dose escalation in a hypo-fractionated setting [9,74].
Commercially available approaches to online adaptive image-guided radiation therapy with LINACs include X-ray [75] and MR-based methods. Although X-rays are more economical and easier to incorporate technically, they have some disadvantages compared to MRI. Firstly, an MRI can be frequently performed with no dose delivered to the patient. Secondly, the superior soft tissue contrast allows for more accurate tumor localization. There are three commercially available MR-LINAC systems [76]: the Elekta Unity (Stockholm, Sweden), the ViewRay MRIdian (MRIdian, ViewRay Inc, Cleveland, Ohio), and the MagnetTx Aurora RT (MagnetTx Oncology Solutions Ltd., Edmonton, AB, Canada). The next section describes the technical design and radiotherapy workflow of the MRIdian, which is the main focus of the presented thesis.

3.3.1 ViewRay MRIdian system

The ViewRay MRIdian combines a 0.35 T split-bore superconducting magnet with a 6 MV LINAC [9]. In the gap between the magnet halves and gradient coils, only thin fiberglass connectors are present. This allows the photon beam to pass through little attenuating material on its way to the patient. All LINAC components are placed inside ferromagnetic buckets that shield them from the magnetic field and avoid any interference with MR imaging. The photon beam is shaped by a double stack of MLCs made of a tungsten alloy.
The MR system enables volumetric imaging for patient positioning and on-table plan adaptation as well as 2D cine MRI acquisition for beam gating. The volumetric images typically have an in-plane resolution of 1.5 mm $\times$ 1.5mm and a slice thickness of up to 3 mm. The 2D cine images are acquired with a resolution of 3.5 mm $\times$ 3.5 mm, slice thickness of 5, 7, or 10 mm, and a frame frequency of 4 Hz. An alternative radial k-space readout is faster than the Cartesian one and allows for 2D imaging with a res-

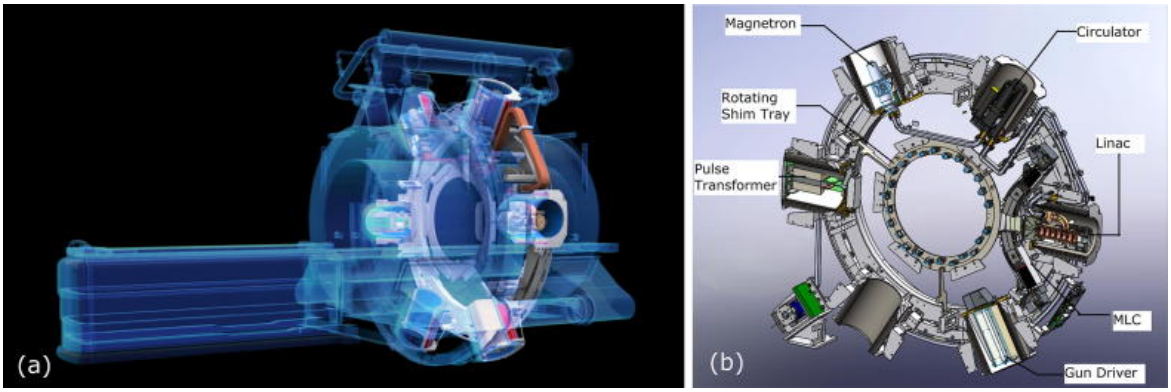

Figure 3.5: MRIdian MR-LINAC system with the main hardware components shown on the left panel and the gantry on the right panel. Figure source [9]

olution of 2.77 mm and a frame frequency of 8 Hz [77].
The MRIdian TPS can generate 3D conformal and step-and-shoot IMRT plans. The inverse dose calculation is performed by a fast GPU MC algorithm considering the static magnetic field.
The MRIdian system uses the balanced steady-state free precession (bSSFP) sequence

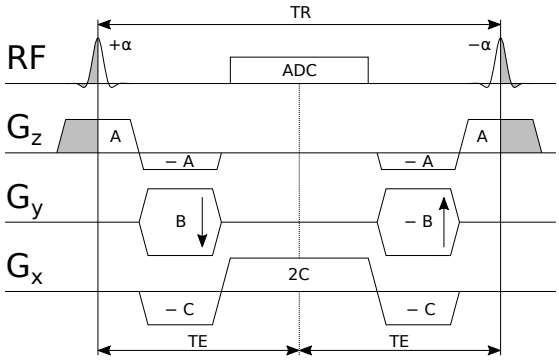

Figure 3.6: Balanced steady-state free precession sequence (bSSFP) implemented at the MRIdian MR-LINAC system. Figure adapted from [78].

schematically depicted in Figure 3.6 [78–81]. The dynamic equilibrium, i.e., the steady state, is achieved by a train of equidistant RF pulses. The arrival at the steady state can take several seconds and is often achieved with dedicated preparation pulses. Using a short repetition time $T_R \leq T_2^* < T_1$ and echo time of $T_E \approx T_R/2$ prevents the longitudinal magnetization from full recovery and interrupts relaxation in the transversal plane. The flip angle must be constant throughout the sequence to achieve identical magnetization at the beginning of each T_R. The acquired images show a T_2/T_1 contrast, characterized by a strong contrast between fluids or fat (bright) and solid tissues (dark).
All gradients applied between two consecutive RF pulses are fully balanced, i.e., the net gradient-induced dephasing is zero within one repetition time. Therefore, any phase accumulation within one T_R is caused by differences in local precession frequencies. This makes the sequence prone to susceptibility artifacts and poor shimming. The sequence offers fast 2D imaging, a high signal-to-noise ratio, and robustness against

motion artifacts.

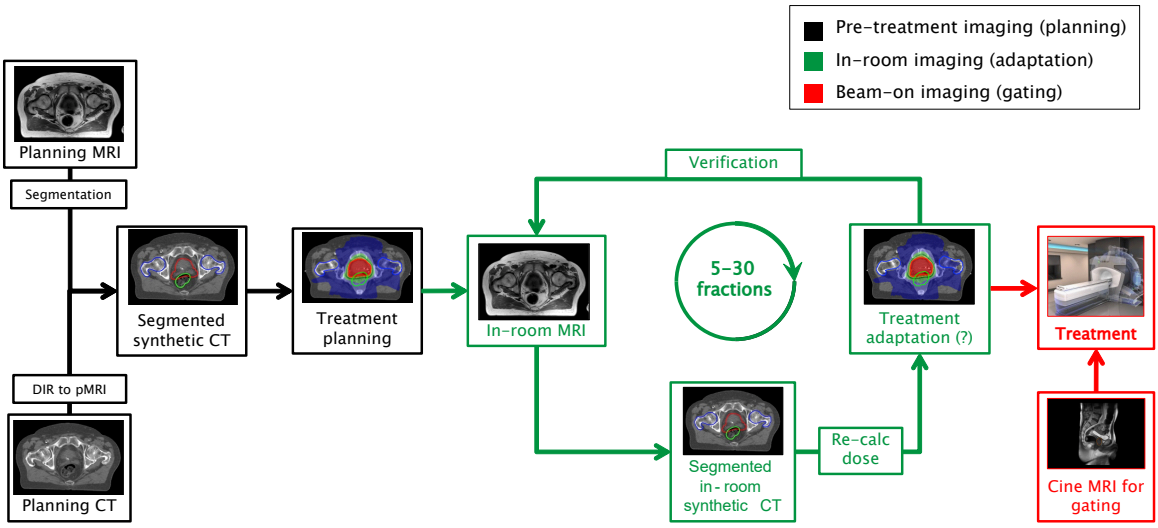

Figure 3.7: Online adaptive workflow used for MRgRT at the ViewRay MRIdian MR-LINAC. The pre-treatment phase, online adaptation during fractions, and treatment delivery are presented. Figure courtesy of Dr. Christopher Kurz.

3.3.2 MRIdian Workflow

The MRIdian radiotherapy workflow is heavily influenced by ViewRay's TPS design [9]. Nevertheless, institutions may differ in how certain steps are automated or who performs them. Figure 3.7 and the following paragraph outline the MRIdian workflow at the Radiation Oncology Department of the LMU University Hospital in Munich [82, 83].

The workflow at the MRIdian is in some parts similar to that at conventional LINACs described in Section 3.2.2. It begins with an acquisition of the planning CT and MR images that are registered to map the electron density from the CT to the planning MRI. The MRI is used by radiation oncologists for defining target volumes and manual contouring of OARs. Subsequently, a step-and-shoot IMRT plan is prepared by medical physicists.

A series of irradiations can follow once a radiation oncologist approves the plan. Each fraction begins with an MRI of the patient positioned the same way as on the planning day. A radiation therapy technologist (RTT) manually registers the planning and fraction MRIs, paying special attention to the PTV region. This is followed by a deformable image registration (DIR) of the planning and fraction MRIs in the TPS. The resulting vector field is used to deform the planning OAR contours and propagate the electron densities from the planning to the daily MRI. The target volumes are copied rigidly to prevent their shrinkage or shape change. The contours obtained that way are inaccurate and must be manually adjusted. This makes the contour preparation one of the most time-consuming steps in the adaptive workflow [13]. Subsequently, regions requiring electron density corrections (e.g., bowel gases that were not on the planning MRI) are marked as they cannot be accounted for by DIR. After the contours

and electron densities are adjusted, the original plan is recalculated based on the daily anatomy. The plan is ready for delivery if all constraints and objectives are met and the dose distribution is satisfactory. In the opposite case, a re-optimization is carried out. A new adapted plan is administered after quality assurance by an independent dose calculation algorithm.
The MRIdian uses real-time 2D cine imaging in the sagittal plane to gate the irradiation [84]. The gating target and boundary are chosen in one of the first frames. The gating target is usually the CTV or GTV, while the gating boundary is defined as a certain expansion of the gating target. The contours are propagated from one frame to the next using the optical flow algorithm. When the overlap of the gating target with the window is below a certain threshold, the beam delivery is paused. For relatively stable anatomies like the pelvis region, the pauses in the delivery are rare and might be caused by patient movement or passage of gases. The irradiation is delivered in breath-hold for tumors affected by respiratory motion. The patient is instructed to hold their breath for several seconds so the beam can hit the target. The beam delivery is paused automatically when the target is out of the gating window.

Chapter 4

Deep learning

This chapter introduces artificial intelligence, deep learning, and neural networks in the context of radiotherapy research. It describes the basics of deep learning for medical image auto-segmentation and registration, which is the main topic of this dissertation.

4.1 Artificial intelligence, machine learning, and deep learning

Artificial intelligence (AI), machine learning (ML), and DL have become current buzzwords. Although these three terms are often used interchangeably, DL is actually a subcategory of ML, and ML is a subcategory of AI. AI is a broad term describing the ability of digital computers to perform tasks commonly associated with intelligent beings [85]. ML focuses on implementing computer software that can learn autonomously [85]. Lastly, DL is a subcategory of ML that uses the so-called artificial neural networks (NNs). The latter consist of many layers of non-linear operations that successively extract and process higher-level features of the input data [86]. The following paragraphs focus on the design, training, and applications of NNs that are relevant to this dissertation.

4.2 Basics of neural networks for image segmentation

NNs are typically depicted as interconnected nodes organized in layers [87]. This work primarily utilized fully connected and convolution layers. In fully connected layers, every node in one layer is linked to every node in the consecutive layer. A convolution/deconvolution layer is a product of convolution/transposed convolution between the previous layer and a kernel, also known as a filter. Convolution combines pixel values with their spatial locations, which makes it particularly useful for image processing. CNNs frequently include max pooling layers. They employ the sliding filter approach, similar to convolution, but instead of conducting matrix multiplication, they output the maximum value of a given patch. The layers of NNs are frequently interleaved by activation and normalization functions. Activation functions aim to introduce non-linearities into the model, while normalization keeps the feature map

values in a reasonable range [88]. Frequently used activation functions include rectified linear unit (ReLU) defined as $f_{\text{ReLU}}(x) = \max(0, x)$ or its variation, a parametric ReLU (PReLU) defined as $f_{\text{PReLU}}(x) = x$ for $x \geq 0$ and $f_{\text{PReLU}}(x) = ax$ for $x < 0$ where a is a trainable parameter. Normalization types used throughout this work include batch and instance normalization. Batch normalization normalizes a feature map using the mean and standard deviation across a batch of training examples [89]. An instance normalization normalizes each feature map separately using its own mean and standard deviation [90].

NNs have thousands to billions of parameters [91]. Due to the high dimensionality of the problem, optimizing network parameters is done iteratively during training. The parameters include hyperparameters set before each training and learnable network parameters updated during training. Hyperparameters describe how the network is structured and the training carried out, while the learnable parameters include, e.g., weights and biases within the layers [92].

Iterative optimization requires a suitable cost or loss function and a tool to update parameters, the so-called optimizer. The choice of the loss function depends on the expected output. In medical image segmentation, the network output is typically a probability map p that predicts for each pixel if it belongs to the foreground $y = 1$ or to the background $y = 0$. Common choices include binary cross entropy (BCE) [17], focal loss (FL) [93], and dice similarity coefficient (DSC) [94]. Currently, the most frequently used one is based on the DSC, which is defined as

$$\text{DSC} = \frac{2\sum_{i=0}^{N} p_i y_i}{\sum_{i=0}^{N} p_i^2 + \sum_{i=0}^{N} y_i^2}, \tag{4.1}$$

where N is the number of image pixels, p_i is the network prediction for the i-th pixel, and y_i is the ground truth value for the i-th pixel. DSC performs well in cases of class imbalance between foreground and background, making it a suitable choice for medical image segmentation where the structures of interest, i.e., organs, are significantly smaller than the imaged region. Despite its benefits, DSC is a purely volumetric function that may underperform in predicting edge pixels. Several ways of estimating Hausdorff distance (HD) between two contours have recently been suggested as a potential loss function [95]. However, since the relatively recent implementation of HD-based loss, its impact on medical image segmentation is yet to be seen. The optimizer's role is to update the network parameters to minimize the loss function and improve prediction accuracy. Common choices include stochastic gradient descent [96], Adagrad [97], or Adam [98]. The latter was the optimizer of choice for most trainings described in this thesis.

After setting up the network architecture, selecting the loss function, and choosing the optimizer, iterative training can be carried out. At each iteration, a subset of training examples, i.e., a (mini-)batch, is fed into the network. A loss is computed for each training example and the mean loss of the mini-batch is calculated. The mini-batch size is a hyperparameter that should be large enough to represent the loss over the entire dataset but small enough to enable fast and memory-efficient training. After calculating the loss, the backpropagation algorithm computes gradients for all learnable network parameters [99]. Subsequently, the optimizer updates network parameters

based on these gradients. Another hyperparameter called learning rate controls the magnitude of parameter adjustments at each step. Large learning rates accelerate the convergence but simultaneously increase the risk of missing the optimal values. Conversely, small learning rates decelerate the convergence and increase the likelihood of falling into local minima only. Certain optimizers like Adam adapt the learning rate dynamically as the training progresses.
The size and variety of the data set are crucial for developing robust, well-generalizing models. Networks that were trained on too few or too homogeneous data are prone to overfitting and underperforming on new examples. Since assembling large data sets is not always possible, data augmentation is implemented to increase the variety of training examples. In this thesis focusing on image processing, augmentation was realized by introducing geometric or intensity variations to the existing images, e.g., applying affine or deformable deformations, adding noise, or blurring the image.

4.3 Deep learning in auto-segmentation

Organ and target segmentation is an integral part of radiation therapy (see Chapter 3). Yet if performed fully manually, it is a tedious, repetitive, time-consuming task, prone to inter- and intra-observer variability [57]. All these factors encouraged the development of automatic and semi-automatic segmentation tools. The next paragraph describes non-AI segmentation algorithms and shows their shortcomings, which drive the development of a new generation of auto-segmentation DL methods described later.

4.3.1 Non-AI auto-segmentation methods

Non-AI segmentation methods include thresholding [100], edge detection, boundary gradient tracking [101], or more advanced algorithms such as deformable shape models or level sets [102]. However, the usage of these algorithms in clinical practice is limited due to their insufficient performance [43]. The next generation of registration algorithms are atlas-based methods. An expert-segmented reference volume (called an atlas) is deformably registered to the image of interest. The resulting vector field is used to deform contours from the atlas to the new image. Therefore, atlas segmentation relies on image intensities as well as expert knowledge of organ position, shape, and size. A variation of this algorithm using several segmented reference volumes simultaneously is called a multi-atlas image segmentation [103]. For several years, it was state of the art and is now implemented in commercial systems such as RayStation (RaySearch Laboratories, Stockholm, Sweden), Elekta ABAS (Elekta, Stockholm, Sweden), or VelocityAI (Velocity Medical Systems, Atlanta, Georgia). The main limitation of multi-atlas segmentation is the relatively long run time due to the underlying registration.

4.3.2 DL auto-segmentation with U-Net architecture

New medical image segmentation algorithms started emerging with the development of more powerful GPUs and deep NNs [16]. In particular, CNNs gained popularity for their ability to combine pixel intensity values with their spatial location while reducing the number of parameters compared to fully connected networks. First CNNs were two-dimensional and worked on image patches. Larger input sizes became possible with the rapid increase in available GPU memory, and ultimately, 3D imaging data could be processed at once.
The most commonly used architecture for medical image segmentation is the so-called U-Net. The 2D fully convolutional U-Net was first proposed in 2015 for biomedical image segmentation [17]. It consisted of contracting (encoder) and expanding (decoder) paths connected at each resolution level via skip connections as shown in Figure 4.1. The encoder and decoder had a typical CNN structure, i.e., convolutional layers interleaved by max pooling or deconvolution operations, with ReLU [104] activation. Each time the max pooling/deconvolution halved/doubled the resolution, the number of feature channels was doubled/halved. The final 1×1 convolution layer transferred each component of the feature map to one of the possible output classes.
Over time, many changes have been introduced to this architecture while maintaining

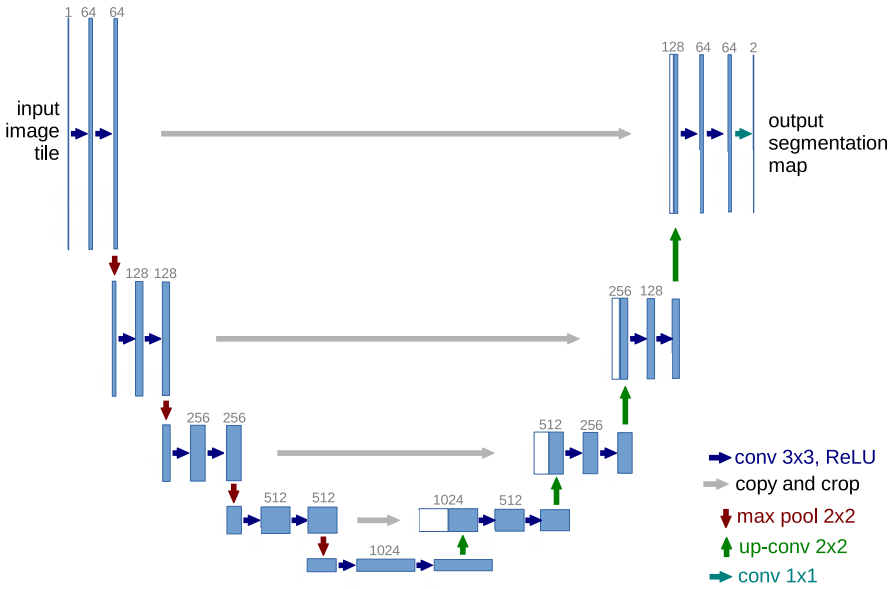

Figure 4.1: U-Net architecture as introduced by Ronneberger et al. [17]. The operations illustrated include 2D 3×3 convolutions, application of ReLU activation, 2×2 max pooling, 2×2 transverse (up) convolution, and a final 1×1 convolution. Gray numbers above the layers indicate the number of channels in each layer. Figure adapted from [17].

its primary encoder-decoder structure with skip connections. The network has been expanded to three dimensions [94], regular convolutional layers have been supplemented with residual connections (see Figure 4.2) [105], various types of normalization have been added [106], and max pooling was replaced by convolution with a stride of 2. Other U-Net variants incorporated fully connected layers, attention mechanisms [107], and recently, vision transformers [108] have been integrated into the U-Net architec-

ture [109]. Finally, the authors of the so-called nnU-Net introduced a training method that automatically configures itself, including pre- and post-processing, network architecture, and training parameters [110].

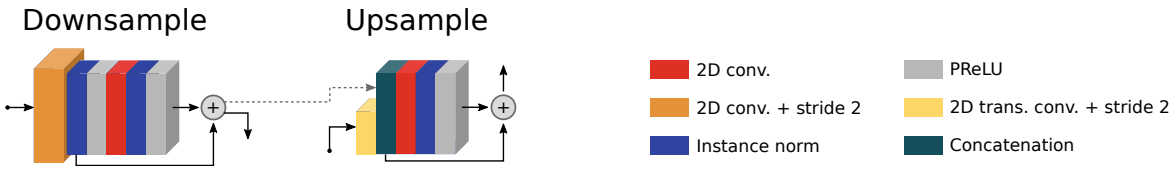

Figure 4.2: Residual units as introduced by Kerfoot et al. [105]. The operations illustrated include 2D convolution with an optional stride of 2, application of instance normalization and PReLU activation, transverse (up) convolution, and concatenation. The gray dashed arrow represents a skip connection used in U-Nets. Figure adapted from [105].

The first U-Net introduced by Ronneberger et al. [17] was trained using the BCE loss. Since the BCE tended to underperform in the presence of class imbalance between foreground and background, the focal loss and finally DSC loss were introduced [80]. DSC is a widely used segmentation loss in medical image segmentation and has been extensively utilized throughout this thesis.

4.3.3 Metrics for evaluating medical image segmentation

In order to evaluate the accuracy of medical image segmentation, different metrics are utilized [111]. They might be quantitative and describe the geometrical similarity between the predicted and ground truth contours, as well as qualitative and describe how well a given contour is suitable for treatment adaptation and dose optimization.
DSC was introduced in the previous subsection as a basis for a loss function during training. However, it is also commonly used to evaluate the overlap between two binary contours A and B:

$$\mathrm{DSC} = \frac{2\sum_{i=0}^{N} a_i b_i}{\sum_{i=0}^{N} a_i^2 + \sum_{i=0}^{N} b_i^2} \tag{4.2}$$

where $a_i \in A$, $b_i \in B$, and $a_i, b_i = 0, 1$. Alternative metrics measuring contour overlaps, such as the Jaccard index, the true/false positive/negative rates, or intersection over union, were not used in this work but can be frequently found in the literature [111].
While DSC is a volumetric metric that considers all voxels inside a contour, HD considers only the borders ∂A and ∂B of the investigated structures. HD is defined as:

$$\mathrm{HD} = \max(\mathrm{hd}(A,B), \mathrm{hd}(B,A)) \quad \text{and} \quad \mathrm{hd}(A,B) = \max_{a \in \partial A} \min_{b \in \partial B} ||a - b||_2. \tag{4.3}$$

The HD is sensitive to outliers. Therefore, it is common in medical image segmentation to report the average or 95$^{\text{th}}$ percentile HD instead of the maximum.
DSC and HD are purely geometric metrics and do not have a direct relation to their usefulness in radiotherapy. Therefore, it is beneficial to complement the quantitative analysis with a qualitative contour evaluation. An example is the Likert scale [112], where an expert evaluates the quality of a contour regarding its value for dose optimization. The commonly used grades in radiotherapy range from ”contours can be

used directly" through "contours require minor/major corrections" to "contours are not useful".
Dosimetric metrics compare radiotherapy plans generated using manual (ground truth) contours and contours predicted by the method under investigation, using, for example, the gamma index [113]. Moreover, DVHs might be created to compare the dose delivered to OARs or to evaluate target volume coverage [114]. Frequently used parameters include the maximum and average dose delivered to a structure or the volume of a structure that received at least a given dose.

4.4 Deep learning segmentation models with prior knowledge

DL auto-segmentation methods incorporating prior knowledge have become more desirable with the introduction of MR-LINAC systems like ViewRay MRIdian or Elekta Unity since these systems necessitate daily image contouring. Online adaptive radiation therapy at MR-LINACs entails a lot of repetitive and time-consuming contouring (see Chapter 3): target volumes and OARs have to be defined on the planning MRI, and then the same structures must be recontoured before every irradiation on the daily MRI. While there is more time for planning segmentation, fraction contours are prepared under time pressure. An effective auto-segmentation specifically for fraction images would benefit MR-LINAC treatments. However, conventional population models may not be optimal for fraction images since they do not include manually segmented planning images. The following paragraphs present two methods of incorporating prior knowledge into DL auto-segmentation models: patient-specific and registration-based networks.

4.4.1 Patient-specific auto-segmentation models

Patient-specific (PS) or personalized networks are defined in this work as models trained using segmented images of a patient of interest. The primary objective of PS models is to prioritize performance for an individual patient over generalizability. Figure 4.3 illustrates four types of PS models that will be described in this paragraph.

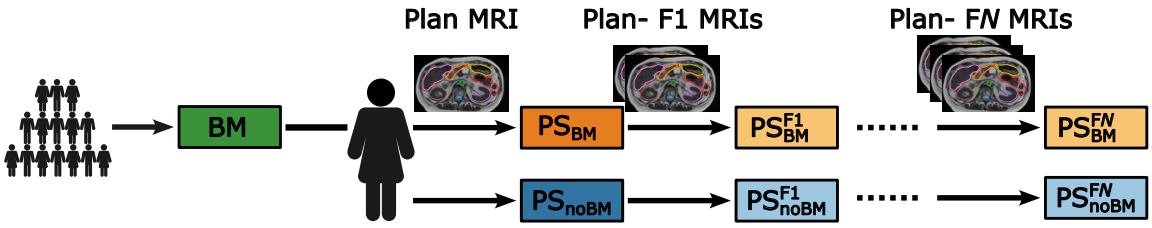

Figure 4.3: Training strategies of personalized models. The boxes represent the models investigated, while the arrows indicate training or fine-tuning. The population baseline models (BM) can be fine-tuned by patient-specific (PS) training either with the planning (Plan MRI) or the planning and the first N fraction (F) images yielding PS_{BM} and PS^{FN}_{BM} models, respectively. Repeating the process without the BMs results in PS_{noBM} and PS^{FN}_{noBM} models, respectively.

Personalized models for fractionated treatment at MR-LINACs could be trained including a patient's planning MRI. There are different ways of conducting such training. One option is transfer learning, where a previously trained population model is fine-tuned with the patient's segmented planning image (PS_{BM} in Figure 4.3). Another alternative is training a PS model from scratch without using other models for initialization (PS_{noBM} in Figure 4.3). Both methods have their benefits. Transfer learning exposes PS models to various patient anatomies, which enhances their robustness. On the other hand, training from scratch requires only one segmented training example, so it overcomes a common hurdle in deep learning: the lack of annotated data.
Continuous fine-tuning with newly acquired fraction MRI is another variant of the PS method. This approach can be implemented for PS models trained from scratch and those generated via transfer learning (PS_{BM}^{FN} and PS_{noBM}^{FN} in Figure 4.3). This fine-tuning updates models with the latest patient anatomy and makes it even more up-to-date for the subsequent fractions.
Two frequently raised concerns about PS training are training time and overfitting. PS training must be short since models for several patients might be required daily. Regarding overfitting, PS models are trained or fine-tuned with only one segmented MRI without validation data. It will be shown throughout the thesis that overfitting can be avoided, and careful selection of hyper-parameters can lead to successful personalized models.
At the time of writing this dissertation, several contemporary publications from other research groups describing PS models exist. All of these publications (including the one included in this dissertation, which will be presented in Chapter 5) were developed independently and published within a span of several months. One of the first studies describes intentional deep overfit learning for adaptive radiation therapy [115]. This paper presented a general idea of personalized models for adaptive radiotherapy and suggested several possible use cases including super-resolution MRI reconstruction, pseudo CT generation, and auto-segmentation. A study by Chen et al. [116] implemented PS training for prostate cancer patients receiving fractionated MRgRT at the Elekta Unity MR-LINAC. The study demonstrated that PS fine-tuning effectively catches characteristics of a given patient to improve segmentation with respect to population models. A study by Fransson et al. [117] investigated personalized 2D auto-segmentation models for prostate cancer patients irradiated at the Elekta UnityMR-LINAC. Their models were trained from scratch with one segmented MRI of a patient. The study found PS training successful for many patients. However, the models tended to fail in cases of bigger differences between the planning and fraction MRIs. Another study conducted at Elekta Unity considered patients with tumors in the abdomen [118]. It explored the potential benefits of daily updating 2D PS models trained from scratch with new segmented fraction MRIs. The study concluded that progressive PS training significantly improves the performance of models trained from scratch. Notably, no study was found that trained 3D PS models from scratch. A possible reason for that is that 2D networks have fewer trainable parameters and are easier to train with less data. Moreover, each image slice is treated as a separate training example, which artificially increases the database size.

4.4.2 Registration for deep learning auto-segmentation

Image registration is the core of an alternative approach for DL auto-segmentation with prior knowledge that was investigated in this work. This approach can be employed if the image to be segmented can be registered with another image where the structures of interest are already delineated. A vector field resulting from this registration is used to warp contours from one image to the other. The principle is analogous to the atlas approach detailed in Section 4.3.1. The only difference is that the deformation vector field is predicted by a NN rather than obtained from conventional registration algorithms as described in Section 2.3.
The possible methods for DL image registration are very versatile. They include CNNs, generative adversarial networks (GANs), and reinforcement learning [119–122]. DL registration models can be trained in a supervised, weakly supervised, and unsupervised manner. The possible outputs are a dense displacement field (DDF), sparse transformation on a grid of control points, and a parametric deformation (when referring to the network output in this dissertation, terms such as deformation, transformation, or vector field will be used interchangeably). The following paragraph is limited to the methods used throughout this work. Namely, it outlines CNNs predicting DDFs between a pair of images. Particular training elements are presented in Figure 4.4.

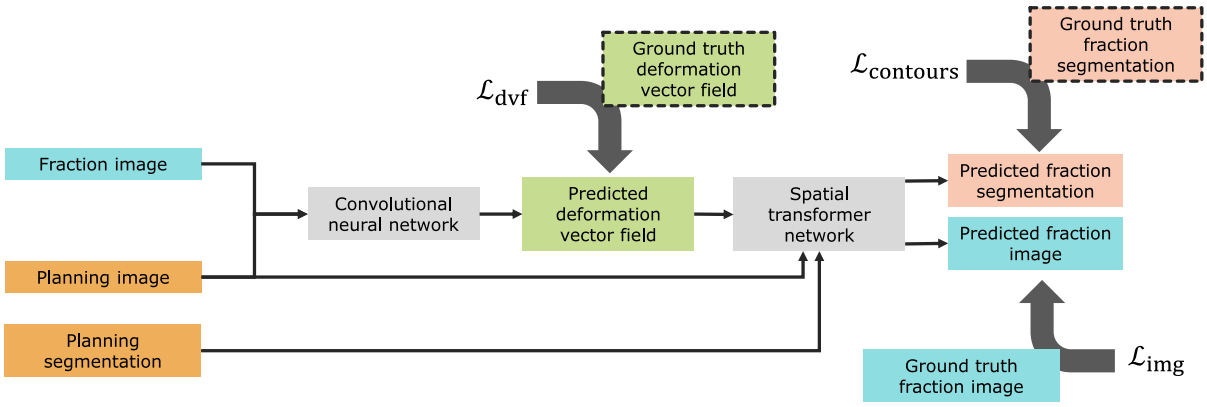

Figure 4.4: Training of combined image registration and contour propagation CNNs using a spatial transformer network. Boxes with dashed borders represent elements that might not exist in all training variants. Different loss function components include a term for vector field similarity and regularization ($\mathcal{L}_{\text{dvf}}$), a term for contour similarity ($\mathcal{L}_{\text{contours}}$), and a term for image similarity ($\mathcal{L}_{\text{img}}$).

In supervised DL image registration the inputs are an image and its copy augmented with a known ground truth vector field. The task of the network is to find this field. The ground truth can represent a rigid, affine, deformable, or random transformation. The field can be modeled (e.g., via B-Splines [123], thin spline models, or specific organ deformation models [43]), obtained from conventional image registration (see Section 2.3), or generated randomly. The mean squared error calculated between the predicted and ground truth fields is a commonly used loss function for DL registration with a ground truth vector field. In addition, a regularization term can be included in the loss function to prevent tissue folding and other unphysical deformations. The other role of the regularization is to incorporate existing knowledge of

the problem into the NN training, e.g., the predicted vector field should be smooth or relatively small in magnitude [43].
Weak supervision is a training variant in which the loss is calculated on a surrogate quantity instead of directly on the network output. For example, the similarity between fixed image contours and moving image contours warped by the predicted vector field can serve as a surrogate quantity. Another possible loss function measures the similarity between the fixed and warped images using, e.g., correlation coefficient or mutual information [43]. Such loss functions are commonly used in NN training as they require no data annotation.
Spatial transformer network (STN) is an architecture that brought significant improvements to the research on DL image registration [124]. STN is a differentiable module that can be inserted into an existing network architecture to allow it to spatially transform feature maps. An STN module comprises a localization net, a grid generator, and a sampler. The localization net takes the feature map as input and outputs transformation parameters to be applied to it. For example, the localization net modeling rigid transformation predicts rotation angles and translation vectors. The input feature map is warped by the subsequent grid generator and sampler using parameters predicted by the localization net. Importantly, the sampling layers are differentiable, which allows loss gradients to flow back towards the localization net to update transformation parameters. STN does not require additional supervision as its parameters can be adjusted during training for the task in question.
VoxelMorph is a well-known architecture using STN that was developed for brain

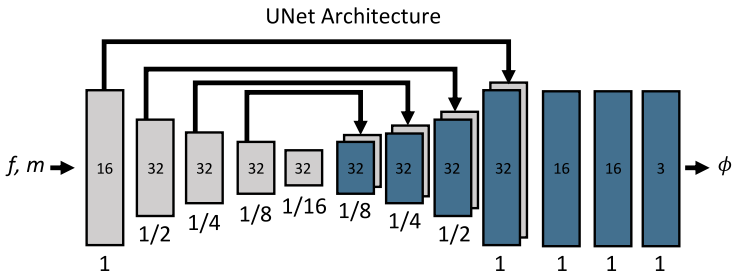

Figure 4.5: VoxelMorph architecture combines a fully convolutional U-Net with three fully connected layers. The input consists of two images to register (f and m), and the output is the predicted DDF ϕ. Figure source: [125].

MRI registration [125]. As shown in Figure 4.5, VoxelMorph comprises a U-Net and three fully connected layers. It predicts DDF between a pair of images and uses the STN's sampler to warp the moving image. It can be trained with an image similarity loss (e.g., cross-correlation) or with an auxiliary contour similarity loss (e.g., DSC) if the images to be registered are segmented.
Another study that aimed to train NNs for combined image registration and segmentation was carried out at the Elekta Unity MR-LINAC [126]. The goal of the study was to train a network to register the planning and fraction MRIs and deform the planning CTV contour to the daily anatomy. The network architecture was a typical U-Net but trained progressively, mimicking the conventional B-Spline registration. In the first epochs, the input image pair was downsampled four times so the network could learn

a rough alignment of the images. As the training progressed, the network was fed three and two times downsampled images, allowing it to focus on smaller details. The last epochs were carried out with image pairs with the original resolution.
Many research groups are currently investigating medical image registration using NNs. Some of these efforts are part of segmentation challenges such as Learn2Reg [127]. Work on DL image registration and segmentation published in the framework of this doctoral study is not included in this paragraph but listed in Chapter 5.

Chapter 5

Publications

This chapter presents publications included in this cumulative dissertation. All four papers have been published in peer-reviewed journals.

5.1 Paper I

The study described in the first paper served as an introductory project to the fields of DL auto-segmentation and radiation therapy. The main research question of this work was: what is the impact of AI-generated OAR and target volume contours on dose optimization during treatment planning. Or, in a more mathematical formulation, is there a correlation between the contour quality (measured by DSC and HD) and the quality of the treatment plan optimized using AI contours (measured by gamma passing rate, conformity index, and DVH parameters). This study was motivated by the fact that the majority of published research on auto-segmentation for radiotherapy reported only geometric metrics for the contours. Although geometric quantities are relevant and enable quick contour comparisons, they do not directly describe the quality of dose distributions obtained during treatment planning. This publication is one of the first works on AI auto-segmentation that addressed the dose as one of the most critical endpoints in radiotherapy. The study involved the most frequently employed geometric metrics in auto-segmentation and clinically relevant parameters used daily for dose distribution assessment.
The data used in this publication included segmented CT images originating from prostate cancer patients treated at the LMU University Hospital. Contours considered as ground truth came either from the clinical workflows or were re-segmented and approved by a radiation oncologist for the purpose of this study. V-Net [94] models were trained for OAR and target volume segmentation while the RayStation was employed as a TPS. Two dose distributions were optimized for each patient using the same optimization settings: one with the ground truth clinical contours and one with the AI-predicted contours. Dose distributions calculated using ground truth contours were considered a reference. DSC and HDs were used as geometric metrics. Conformity index, gamma pass rate, and DVH parameters were used as dosimetric metrics. Finally, the Pearson correlation coefficient was calculated between the geometric and dosimetric metrics.

This study brought the following results. First, the training of DL auto-segmentation networks led to models achieving state-of-the-art geometric performance. Second, treatment plans optimized using DL contours did not result in overdosing the neighboring OARs. Additionally, the differences in target volume coverage were relatively small except for one case. Third, the only statistically significant, moderately positive correlation was found between the prostate DSC and the gamma index. In particular, some patients showed similar contour accuracy yet substantially different gamma pass rates. This led to the conclusion that the high geometric accuracy of predicted contours might not be directly linked to high-fidelity dose distributions.

Radiation Oncology

RESEARCH

Open Access

Dosimetric impact of deep learning-based CT auto-segmentation on radiation therapy treatment planning for prostate cancer

Maria Kawula[1], Dinu Purice[1,2], Minglun Li[1], Gerome Vivar[3], Seyed-Ahmad Ahmadi[3], Katia Parodi[2], Claus Belka[1,4], Guillaume Landry[1,2] and Christopher Kurz[1,2*]

Abstract

Background: The evaluation of automatic segmentation algorithms is commonly performed using geometric metrics. An analysis based on dosimetric parameters might be more relevant in clinical practice but is often lacking in the literature. The aim of this study was to investigate the impact of state-of-the-art 3D U-Net-generated organ delineations on dose optimization in radiation therapy (RT) for prostate cancer patients.

Methods: A database of 69 computed tomography images with prostate, bladder, and rectum delineations was used for single-label 3D U-Net training with dice similarity coefficient (DSC)-based loss. Volumetric modulated arc therapy (VMAT) plans have been generated for both manual and automatic segmentations with the same optimization settings. These were chosen to give consistent plans when applying perturbations to the manual segmentations. Contours were evaluated in terms of DSC, average and 95% Hausdorff distance (HD). Dose distributions were evaluated with the manual segmentation as reference using dose volume histogram (DVH) parameters and a 3%/3 mm gamma-criterion with 10% dose cut-off. A Pearson correlation coefficient between DSC and dosimetric metrics, i.e. gamma index and DVH parameters, has been calculated.

Results: 3D U-Net-based segmentation achieved a DSC of 0.87 (0.03) for prostate, 0.97 (0.01) for bladder and 0.89 (0.04) for rectum. The mean and 95% HD were below 1.6 (0.4) and below 5 (4) mm, respectively. The DVH parameters, $V_{60/65/70\,Gy}$ for the bladder and $V_{50/65/70\,Gy}$ for the rectum, showed agreement between dose distributions within ± 5% and ± 2%, respectively. The $D_{98/2\%}$ and $V_{95\%}$, for prostate and its 3 mm expansion (surrogate clinical target volume) showed agreement with the reference dose distribution within 2% and 3 Gy with the exception of one case. The average gamma pass-rate was 85%. The comparison between geometric and dosimetric metrics showed no strong statistically significant correlation.

Conclusions: The 3D U-Net developed for this work achieved state-of-the-art geometrical performance. Analysis based on clinically relevant DVH parameters of VMAT plans demonstrated neither excessive dose increase to OARs nor substantial under/over-dosage of the target in all but one case. Yet the gamma analysis indicated several cases with low pass rates. The study highlighted the importance of adding dosimetric analysis to the standard geometric evaluation.

Keywords: 3D U-Net, Automatic segmentation, Radiation therapy, Prostate cancer, Neural networks, Deep learning

*Correspondence: Christopher.Kurz@med.uni-muenchen.de
[1] Department of Radiation Oncology, University Hospital, LMU Munich, Munich, Germany
Full list of author information is available at the end of the article

Background

The anatomical structure of the male pelvic region with the prostate surrounded by seminal vesicles, bladder, and

rectum, makes modern intensity modulated radiation therapy (RT) a favorable technique for the treatment of localized prostate cancer [1–3]. However, due to variable bladder and rectal filling, random shifts, and deformations of neighboring organs, online adaptation of the treatment plan would be necessary in order to take full advantage of modern radiotherapy techniques [4, 5].

Recontouring of the target volume (TV) and organs at risk (OARs) is an important step in treatment plan adaptation. Previous studies have shown that manual delineation is not only time-consuming (in the order of several minutes) but also prone to inter- and intra-physician variability [6–8].

To address these problems, considerable scientific efforts have been made to develop efficient automatic segmentation tools. Previously, auto-segmentation methods such as (multi)atlas based and hybrid techniques have been considered state-of-the-art [9]. Over time, methods based on convolutional neural networks (CNN) [10] gained more attention [11, 12]. Milletari et al. [13] proposed a 3D fully convolutional neural network architecture trained end-to-end on magnetic resonance (MR) prostate images, referred to as V-Net, and introduced a novel objective function based on the Dice similarity coefficient (DSC). Balagopal et al. [14] presented a hybrid network, having an additional 2D localization network prior to the 3D segmentation network to delineate prostate, bladder, rectum, and femoral heads on pelvic computed tomography (CT) images. In order to overcome the challenges of low soft tissue contrast in CT images as well as blurry boundaries, Wang et al. [15] and Tong et al. [16] focused additionally on edge enhancement techniques. Sultana et al. [17] proposed a two-stage network combining U-Net and generative adversarial network (GAN) architectures [18] for structure localization followed by precise prediction of organ delineation.

Evaluation metrics that are commonly used to measure segmentation performance focus purely on geometric accuracy. The most frequently used are the DSC, the mean, 95%, or maximal Hausdorff distance (HD), the positive prediction value (PPV) or the sensitivity [19]. The two main ideas behind them are: (1) a pixel-wise comparison of ground-truth and predicted segmentation and (2) measuring the distance between the ground-truth and the predicted contours. What carries a higher relevance in clinical practice, however, is the dosimetric accuracy and the quality of the treatment plans that can be achieved on the basis of the predicted segmentations [12, 20]. At the time of writing, no studies exist that have investigated and quantified the dosimetric impact of CT organ delineations for prostate cancer patients obtained from deep CNNs.

In this work a state-of-the-art 3D U-Net architecture for automatic organ segmentation in CT images of low-grade prostate cancer patients was trained. The training was carried out separately for the bladder, prostate, and rectum which are the most important structures for prostate cancer treatment. Since in patients with low-grade prostate cancer, tumorous tissue is located only in the prostate, seminal vesicles were not considered for segmentation. Clinically acceptable VMAT plans were created for all test cases using manual segmentations and the automatic segmentations obtained from the 3D U-Net. This allowed to infer the dosimetric impact of deep learning delineations, which is still rarely present in the literature. The quality of the treatment plans optimized on the automatically generated contours was compared with the reference plans in terms of dose volume-histogram (DVH) parameters, conformity index (CI) and gamma pass rate. In addition, a standard contour-based analysis based on DSC as well as on average and 95th percentile HD calculation was performed. Both, geometric and dosimetric evaluation metrics, were compared in terms of Pearson correlation coefficient to investigate a possible correlation between them.

Methods

Database

The dataset used in this study consisted of 69 CT images, along with delineated structures associated with the low-grade prostate cancer treatment performed at the Klinikum Großhadern of the Ludwig Maximilian University (LMU) of Munich. Patients with substantial CT artifacts due to the presence of metal hip implants (1 patient) and fiducial markers (9 patients), causing artifacts throughout the image and especially in the prostate area, were not included in this study. The use of an ultrasound probe for prostate monitoring during irradiation in several cases, did not interfere with CT imaging of the pelvic region, therefore such cases were also included. Similarly, the presence of prostate calcification did not rule out the inclusion of images in the study. CT data have been acquired with a Toshiba Acquilion LB CT scanner (Canon Medical Systems, Japan) using 512×512 pixels in the axial plane and a variable number of slices. Voxel size was $1.074 \times 1.074 \times 3$ mm^3. OARs, in particular bladder and rectum, were delineated by a trained radiation oncologist and stored as point clouds (DICOM RT-structs). The prostate contours were redrawn under the supervision of a trained physician according to guidelines for low grade (stage I and II) prostate tumor patients. Using plastimatch [21] images and segmentations were converted from the DICOM RT-struct format, which is required by treatment planning systems and contouring software, into binary masks that are used during the

neural network training. Images and binary masks were resampled with the help of nearest neighbor interpolation for masks and linear interpolation for images, to a $1 \times 1 \times 1$ mm^3 spaced grid, which was advantageous for the subsequent data augmentation at training stage. While aiming to minimize the influence of contour conversion between the DICOM RT-struct format, defined on a $1.074 \times 1.074 \times 3$ mm^3 grid, and binary masks, defined on a $1 \times 1 \times 1$ mm^3 grid, we found that employing resampling with nearest neighbor interpolation introduced negligible alterations to the structures. Finally, the dataset has been split into a training, validation, and test sets of 47, 11, and 11 images, respectively. This partitioning was a trade-off between providing enough statistic for testing and validation as well as introducing sufficient variability into the training set.

3D U-Net

The 3D U-Net presented here is based on the V-Net architecture [13], developed initially for prostate delineation on MR images. The encoding arm of the network is composed of five levels (including the lowest one) each comprising one (1st level), two (2nd level) or three (3rd–5th levels) convolutional layers and having 16, 32, 64, 128, 256 channels, respectively. The kernel size has been set to $5 \times 5 \times 5$, stride to $1 \times 1 \times 1$ and group normalization has been applied after each convolution. The output of a given level is used in the subsequent one as input for the first convolution and is added to the output of the last convolution, thus creating a residual connection. For downsampling between the network levels convolution with a kernel of size $2 \times 2 \times 2$ and stride 2 was used. Throughout the network the PReLU activation was applied. The decoding arm of the 3D U-Net is built in an analogous way, with up-convolution to increase the image size instead. The output of each level of the encoding arm (before the dowsampling) is concatenated with the corresponding input of the decoding arm. The last layer of the network uses the soft-max activation and thresholding of 0.5 to produce two binary masks representing segmentation of the structures and the background. For this project only the segmentation of the structures is relevant.

Data augmentation

The data augmentation, applied with probability p_{aug} to each input pair, i.e. image and its segmentation, included 3D rotations around the image center (always aligned with the prostate center of mass), translations, B-Spline-based deformations, and zooming. Translations can be described by three parameters $[x_{\text{trans}}, y_{\text{trans}}, z_{\text{trans}}]$ denoting the maximal translation distances along each axis. Similarly, Euler rotations can be denoted by the maximal rotation angles $[\alpha, \beta, \gamma]$ around the superior-inferior, anterior-posterior and medial-lateral axis, respectively. Zooming re-sizes each axis by a factor randomly drawn from $[l_{\text{min}}, l_{\text{max}}]$. The pixel intensities have been truncated to fit the soft tissue window $[I_{\text{min}}, I_{\text{max}}]$ and subsequently rescaled to $[-1, 1]$. The deformation field is defined on a grid of $n \times n \times n$ control points with random shifts drawn from a Gaussian distribution $[\mu, \sigma]$. In the last step of the augmentation pipeline, a central part of each image has been cropped to 128×128×128 due to memory limitations on the GPU. Nevertheless, the clinically relevant high dose regions close to the prostate were not affected by the cropping. While setting the initial values for the data augmentation parameters, special care was taken not to introduce strong artifacts or create unrealistic deformations.

Training

Training on single-label data has been performed separately for three regions of interest: prostate, rectum, and bladder. Each model has been trained on an NVIDIA Quadro P6000 GPU with the Keras implementation of the Adam optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1e-07$) and the Dice loss function applied to both, segmentations and the background. The set of hyper-parameters to be optimized can be divided into two sub-groups: data augmentation related parameters such as maximal translation shifts, rotation angles, zooming and soft-tissue window limits, B-Spline deformation parameters, augmentation probability and training related parameters such as the learning rate and number of epochs. The optimization of the hyper-parameters was performed via a random search. Training with a certain set of hyperparameters was performed until the loss function evaluated on the validation data did not decrease further for several dozen epochs.

Treatment planning

For all test cases, single arc photon VMAT treatment plans were generated using a research version of the commercial treatment planning system (TPS) RayStation (version 8.99, RaySearch, Sweden). All plans aimed at a total dose of 74 Gy in 37 fractions. The generic beam model of an Elekta Synergy Linac (Elekta, Sweden) with Agility multi-leaf-collimator was used. For each test case, two treatment plans were optimized on the same planning CT image, one based on the expert segmentation and one based on the 3D U-Net segmentation of rectum, bladder, and prostate. In both scenarios, in accordance with our facility's clinical guidelines, a PTV margin of 6 mm (posterior 5 mm) was applied around the prostate. The same optimization settings, i.e., the same objectives and weights for planning target volume

(PTV), bladder, and rectum, for both manual and automatic segmentation were used. Settings were chosen using the expert segmentation such that a PTV coverage of at least $V_{95\%} = 100\%$ was achieved (no normalization was applied after optimization), while dose to OARs was below the recommendations of the QUANTEC report [22]. Since the dose optimization problem does not have a unique solution, calculation outcomes might be different, despite using highly similar sets of contours. In order to perform a dosimetric evaluation that captures differences in dose distributions caused primarily by variations in the delineated structures and not by the solution ambiguity of the optimization problem, care was additionally taken to choose optimization settings that produce consistent planning results by applying small perturbations to the manual segmentation. For this, the original RT-structs were converted to binary masks and back to DICOM RT-structs. Then a new plan was generated with the same optimization settings and dosimetrically compared to the initial plan using the original RT-structs. With the final parameters (see weights in Table 1) dose distributions for all test cases were achieved that deviated less than $\pm 2\%$ in the considered OAR and target DVH parameters (see following section) but were not statistically significant. For all test patients and all calculated dose distributions, the ICRU Report 83 guidelines concerning the PTV [23], i.e. $D_{98\%} \geq 95\%$ of the prescribed dose and $D_{2\%} \leq 107\%$ of the prescribed dose, were met as well. These settings were then used to optimize treatment plans using the 3D U-Net segmentations without further user interaction. Table 1 summarizes the goals of the treatment planning along with the importance of each factor.

Table 1 Clinical goals used in the TPS RayStation for VMAT plan generation

Function	ROI	Description	Weight
Max dose	Rectum	74 Gy	0.03
Max EUD, A = 12	Rectum	64 Gy	0.11
Max EUD, A = 8	Bladder	63 Gy	0.03
Min dose	PTV	74 Gy	0.42
Uniform dose	PTV	74 Gy	0.07
Max dose	PTV	77.7 Gy	0.21
Dose fall-off	External	[H]74 Gy, [L] 10 Gy, Low dose distance 1 cm	0.13

For each region of interest (ROI) a given objective function was assigned. Weights were normalized to 1 and indicate the importance of each parameter during plan optimization

Data evaluation

In order to evaluate the network-generated contours, DSC, average HD and 95% HD (defined as 95th percentile of the distances between boundary points), have been calculated for all test cases with expert delineations as the reference ground truth. Since there is no clear boundary between the rectum and colon, evaluation of the network predictions was limited to the slices containing the ground truth segmentation, i.e. no additional penalty was applied for colon misclassification. Apart from that, geometric data evaluation (DSC, HD_{avg}, and $HD_{95\%}$) has been restricted to the $128 \times 128 \times 128$ volume.

The dose distributions for predicted and ground truth contours were analyzed using a 3D global gamma-criterion with a pass-rate of (3%, 3 mm), where only voxels with at least 10% of the prescribed dose were considered. Additionally, CI defined by Paddick [24] was calculated. This index has an ideal value of one and plan quality decreases with decreasing index value. Both dose distributions were also compared in terms of clinically relevant target and OAR DVH parameters. For prostate and its 3 mm expansion (surrogate CTV), values of $D_{98\%}$, $D_{2\%}$ and $V_{95\%}$ were determined. Similarly, for the rectum $V_{50/65/70\,Gy}$ and for the bladder $V_{60/65/70\,Gy}$ were calculated. All DVH parameters were determined using the ground truth segmentations and the dose distributions optimized either on the predicted or on the ground truth contours. To assess the statistical differences between DVH parameters for plans optimized on the manually and the U-Net generated contours, a Wilcoxon signed-rank test with a statistical significance threshold of $p = 0.05$ was used.

To investigate the correlation between the dosimetric and geometric metrics, the Pearson correlation coefficient [25] between (1) DSC of prostate and gamma index, (2) average DSC and gamma index, and (3) DSC and DVH parameters were calculated.

Results

Hyperparameter optimization

The following values of hyperparameters have lead to satisfactory results: $p_{aug} = 0.93$, rotation angles $\alpha = 20°$, $\beta = \gamma = 10°$, translation shifts $x_{trans} = y_{trans} = z_{trans} = 10\,mm$, $l_{min} = 0.9$, $l_{max} = 1.1$, $I_{min} = -150\,HU$, $I_{max} = 150\,HU$, grid control points $n \times n \times n = 15 \times 15 \times 15$, $\mu = 0$, $\sigma = 30$. After 20k epochs with a batch size of two, we found all the loss functions to converge with no signs of overfitting. The learning rate of 10^{-3} has been shown to perform best.

Contour-based analysis

Figure 1 illustrates ground truth and automatically-generated delineations of prostate, rectum, and bladder for

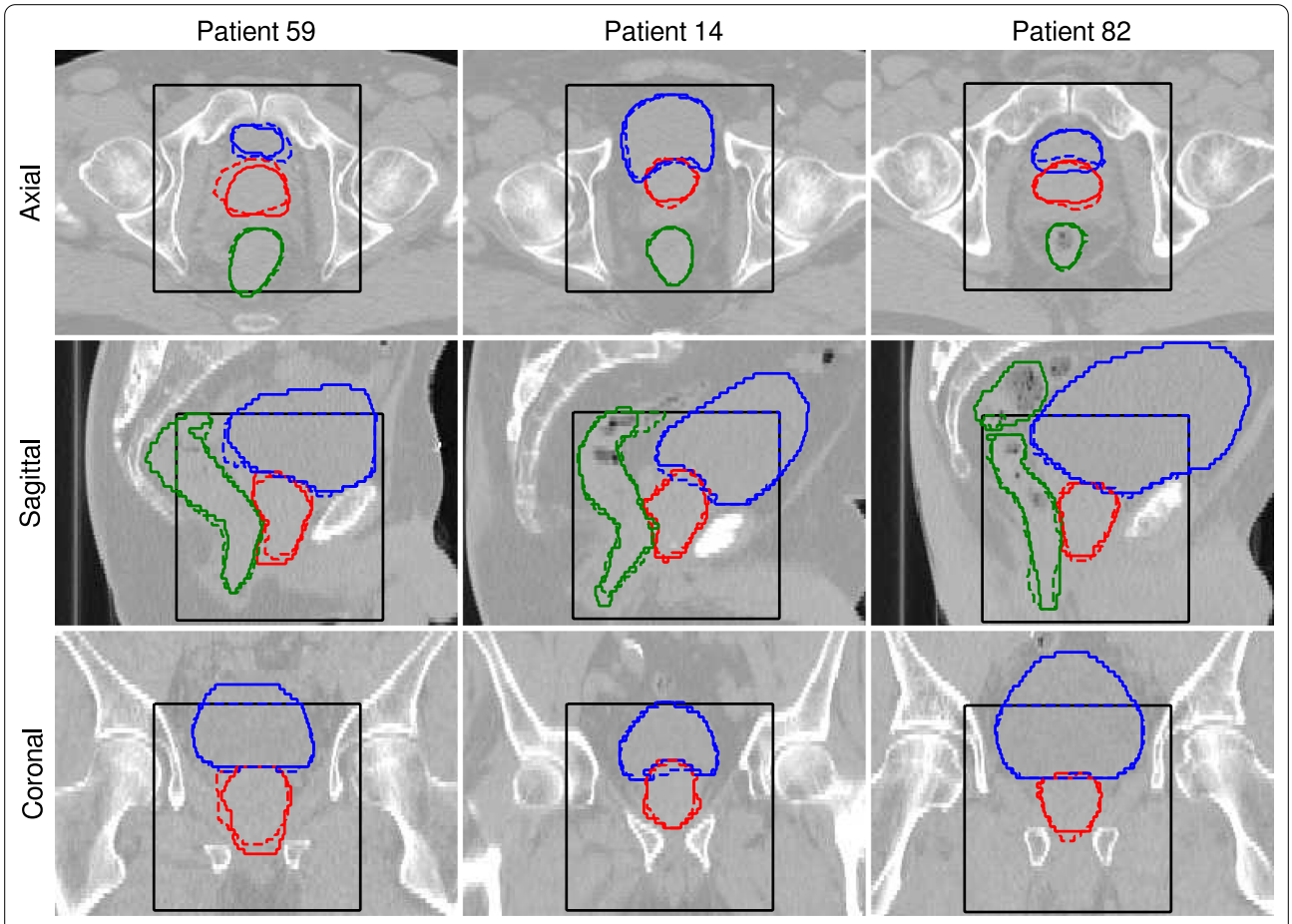

Fig. 1 Axial, sagittal, and coronal slices showing (solid lines) the ground truth contours and (dashed lines) predictions generated by the 3D U-Net. (Red) prostate, (green) rectum, and (blue) bladder delineations are presented for three test patients showing (left) the worst, (middle) closest to the average, and (right) the best agreement with the ground truth by means of DSC for prostate. The black box indicates the region where the contours were predicted by the U-net

Table 2 Contour based metrics: DSC, average Hausdorff distance (HD_{avg}) and 95% Hausdorff distance ($HD_{95\%}$) of all test patients

	DSC			(HD_{avg}/$_{95\%}$) (mm)		
	Prostate	Bladder	Rectum	Prostate	Bladder	Rectum
Pat. 11	0.90	0.96	0.90	1.4/4.5	1.0/2.3	1.2/4.0
Pat. 14	0.88	0.96	0.88	1.5/3.6	1.0/3.6	1.2/3.5
Pat. 27	0.86	0.97	0.91	1.5/3.7	0.9/2.2	1.1/3.0
Pat. 32	0.87	0.96	0.78	2.2/5.1	1.1/3.2	3.4/14.9
Pat. 43	0.85	0.94	0.90	1.7/4.2	1.5/3.2	1.2/3.3
Pat. 44	0.88	0.96	0.92	1.3/3.6	0.8/2.1	0.8/2.2
Pat. 52	0.83	0.97	0.92	2.0/5.5	0.9/2.3	1.4/3.5
Pat. 59	0.82	0.97	0.90	2.3/6.2	0.9/2.6	1.3/4.9
Pat. 81	0.91	0.97	0.87	1.2/3.4	0.9/2.1	1.8/8.3
Pat. 82	0.92	0.97	0.88	1.0/2.3	0.8/2.1	1.2/3.5
Pat. 90	0.85	0.97	0.91	1.6/4.3	0.8/2.1	1.0/2.9
Mean (STD)	0.87 (0.03)	0.97 (0.01)	0.89 (0.04)	1.6 (0.4)/4 (1)	0.95 (0.2)/2.5 (0.5)	1.4 (0.7)/5 (4)

The last row presents the mean and standard deviation (STD) over all test cases

three test patients. Images with the best, closest to the average, and the worst values of DSC for prostate are displayed.

Table 2 collects the results of the geometric analysis for all test patients. Mean DSCs (standard deviation) of 0.87 (0.03), 0.97 (0.01), 0.89 (0.04) were achieved for the prostate, bladder, and rectum, respectively. The highest average DSC value was observed for the bladder, which can be attributed to its relatively large size. A slightly worse performance has been observed for rectum and subsequently prostate. The values of the average HD were 1.6 (0.4) mm, 0.95 (0.2) mm, 1.4 (0.7) mm for prostate, bladder, and rectum, respectively. The values of the 95% HD show the same trend 4 (1) mm, 2.5 (0.5) mm, 5 (4) mm for prostate, bladder, and rectum, respectively.

Dosimetric analysis

Figures 2, 3 and 4 illustrate dose distributions of three exemplary patients with the highest, the average, and the lowest gamma pass-rate in axial, sagittal and coronal views. The reference dose distribution optimized using the ground truth contours, the 3D U-Net dose distribution optimized using the predicted delineations, and their difference are shown. Deviations from the reference plan were found to be in the range of $\pm 10\%$ and were located primarily outside of the prostate. The largest differences were found close to the borders of the PTV region, where dose gradients are steep (6 mm away from the prostate boundary).

The quantitative results of the dosimetric comparison are summarized in Table 3. The value of the CI for the reference plans is in the range of 0.81 and 0.89 with an average (standard deviation) of 0.85 (0.03). For the plans calculated on 3D U-Net generated contours the CI is in the range of 0.69 and 0.88 with an average of 0.78 (0.06). The gamma-pass rates (3 mm, 3%) were between 71 and 94%, with an average value of 85%.

Figure 5 illustrates differences between clinically relevant DVH parameters of the two optimized dose distributions, evaluated on the reference, i.e. manually delineated, contour set. Again, the reference dose distribution was optimized using the ground truth delineations and compared the the dose distribution optimized on the 3D U-Net predicted contours. For rectum and bladder, all the differences are below 5% and 2%, respectively.

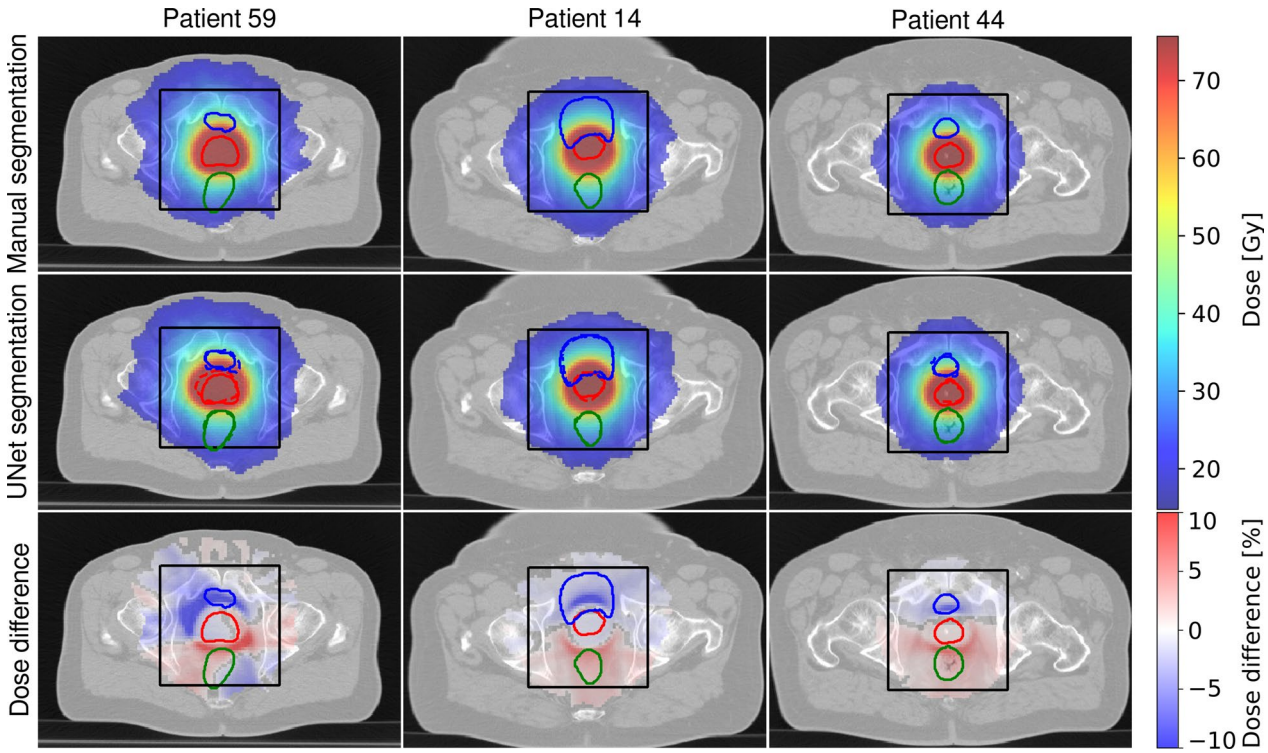

Fig. 2 Dose distributions on axial slices of three test patients showing (left) the worst, (middle) the average and (right) the best agreement quantified by the gamma-index of the treatment plan optimized on (top) the manual segmentation and (middle) the 3D U-Net segmented images. Additionally, relative dose differences are presented. For improved visibility, dose below 25% of the dose prescribed to PTV and deviations below 0.4% on the difference plot are not displayed. Ground truth contours of (green) rectum, (blue) bladder, and (red) prostate, are also shown. The black box indicates the region where the contours were predicted by the U-Net

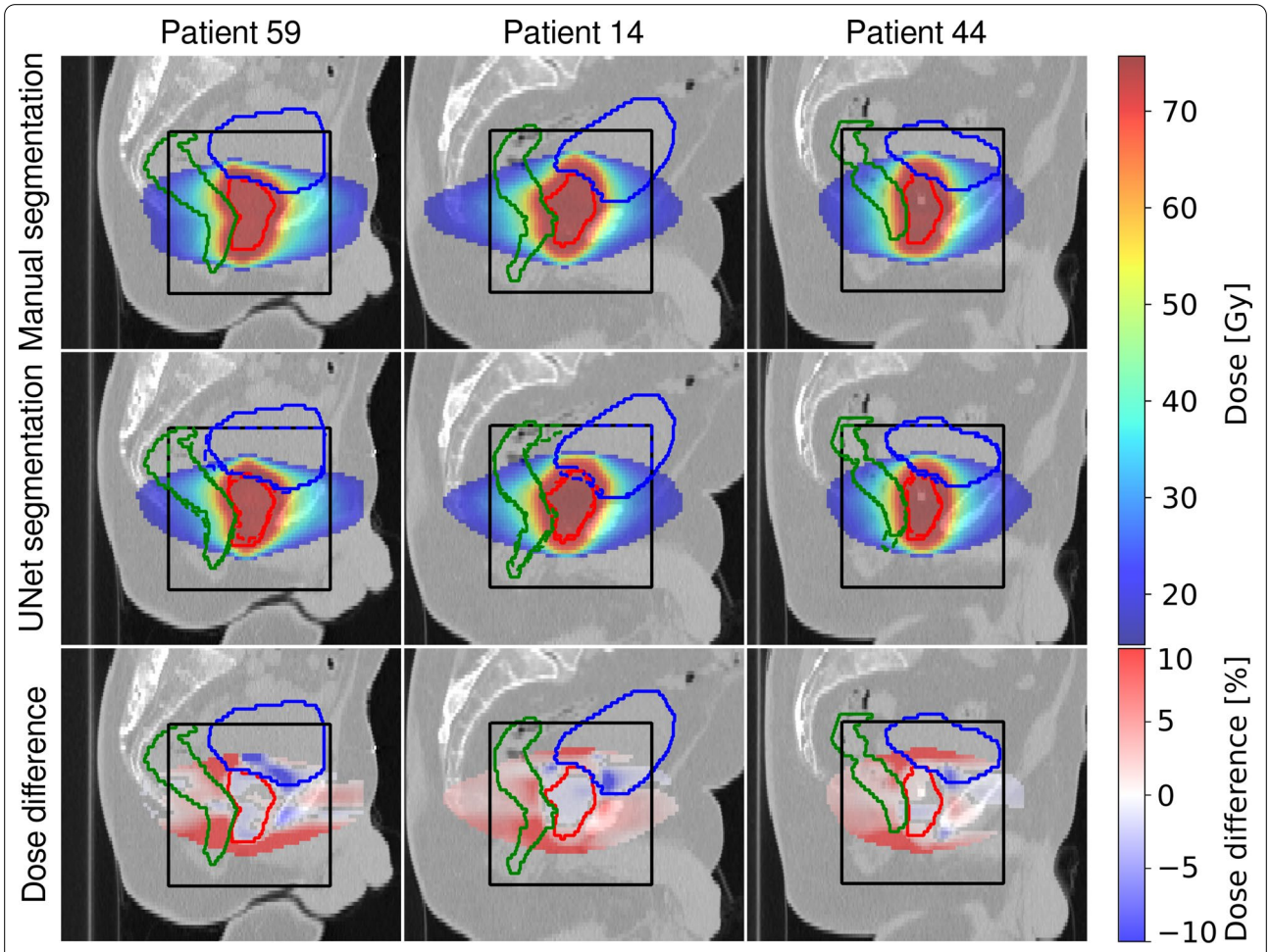

Fig. 3 Dose distributions on sagittal slices of three test patients showing (left) the worst, (middle) the average and (right) the best agreement quantified by the gamma-index of the treatment plan optimized on (top) the manual segmentation and (middle) the 3D U-Net segmented images. Additionally, relative dose differences are presented. For improved visibility, dose below 25% of the dose prescribed to PTV and deviations below 0.4% on the difference plot are not displayed. Ground truth contours of (green) rectum, (blue) bladder and (red) prostate are also shown. The black box indicates the region where the contours were predicted by the U-Net

None of them has been found to be statistically significant ($p \geq 0.05$). No clear trend of increased or decreased bladder and rectum dose for the 3D U-Net segmentation-based plans was found. Similarly, differences for the target volume are mostly below 3 Gy/2% for D_{98}, D_2 and V_{95}, apart for one outlier (patient 59, 10% of the test set) where the network struggled to delineate the prostate, which is also reflected in the relatively low DSC of 0.82 and gamma index of 71%. The only statistically significant differences have been found for the surrogate CTV for D_{98} and V_{95}. No tendencies for the D_2 parameter have been observed, but the 3D U-Net based plans tend to have reduced values of D_{98} and V_{95} for both, prostate and its 3 mm expansion, indicating a slight reduction of target coverage which is in line with the reduced CI values.

Pearson correlation coefficient

The Pearson correlation coefficient with the p value for the DSC of prostate and gamma index was 0.67 ($p = 0.023$), which shows a moderate positive correlation. No statistically significant results were obtained for the other parameters.

Discussion

In this work a 3D U-Net has been successfully trained and applied for CT-based organ segmentation in the male pelvic area. The evaluation of the network's performance was based not only on the commonly used geometric metrics, but also on clinically relevant dosimetric parameters.

Satisfactory performance was observed with regard to the geometric accuracy of the contour delineation, indicating a high degree of similarity between

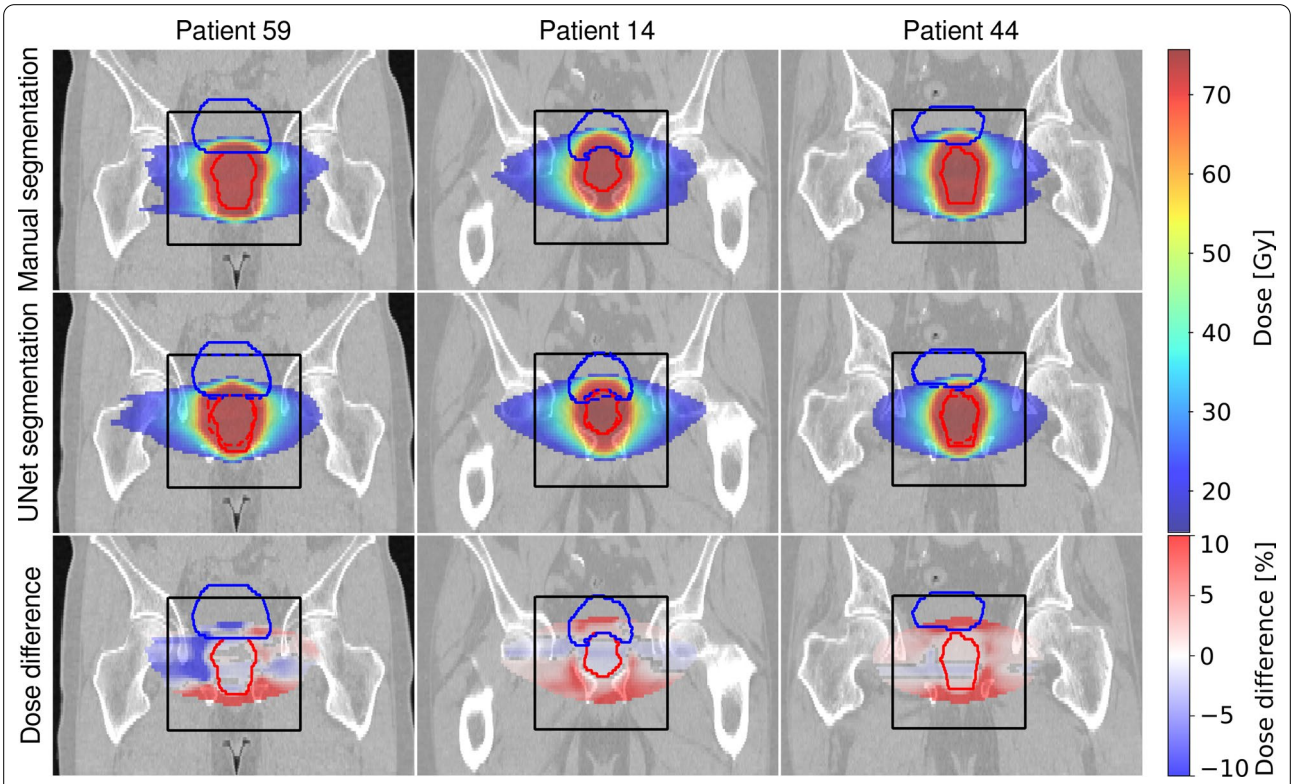

Fig. 4 Dose distributions on coronal slices of three test patients showing (left) the worst, (middle) the average and (right) the best agreement quantified by the gamma-index of the treatment plan optimized on (top) the manual segmentation and (middle) the 3D U-Net segmented images. Additionally, relative dose differences are presented. For improved visibility, dose below 25% of the dose prescribed to PTV and deviations below 0.4% on the difference plot are not displayed. Ground truth contours of (green) rectum, (blue) bladder and (red) prostate are also shown. The black box indicates the region where the contours were predicted by the U-Net

Table 3 Gamma pass rate (3 mm, 3%) and conformity index calculated for plans optimized on manual (CI_{man}) and 3D U-Net ($CI_{3DU-Net}$) generated segmentations of all test patients

	CI_{man}	$CI_{3DU-Net}$	(3 mm, 3%) (%)
Pat. 11	0.87	0.78	91
Pat. 14	0.83	0.83	89
Pat. 27	0.83	0.75	88
Pat. 32	0.89	0.77	79
Pat. 43	0.83	0.78	93
Pat. 44	0.82	0.88	94
Pat. 52	0.84	0.69	77
Pat. 59	0.87	0.70	71
Pat. 81	0.88	0.82	87
Pat. 82	0.85	0.85	92
Pat. 90	0.81	0.72	74
Mean	0.85 (0.03)	0.78 (0.06)	85 (8)

automated and manual segmentations. The best results were observed for bladder segmentation, followed by the rectum, and prostate. The best values of DSC and HD for the bladder can be explained firstly, by its simple geometry and secondly, by its relatively large size, which makes an incorrect prediction of a group of edge pixels less relevant with regard to the correctly classified central part of this organ. The low contrast of the prostate on the CT images makes its segmentation most challenging, which was reflected in a DSC of 0.87. With the exception of one case (Pat. 32) in which a substantial portion of the colon was misclassified as part of the rectal contour, the rectum segmentation showed a relatively high dice equal to 0.87. Since the rectum-colon boundary is visually difficult to identify and is not located in the high dose region, we decided to reduce the penalty for this type of misclassification during the final evaluation (testing) by truncating the volume of interest to the axial slices that contained the ground truth segmentation.

Quantitative test outcomes showed state-of-the-art network performance in terms of DSC, mean and 95% HD. The 2D–3D hybrid network for localization and subsequent organ segmentation proposed by Balagopal et al. [14] achieved a DSC of 0.9 for prostate, 0.95 for

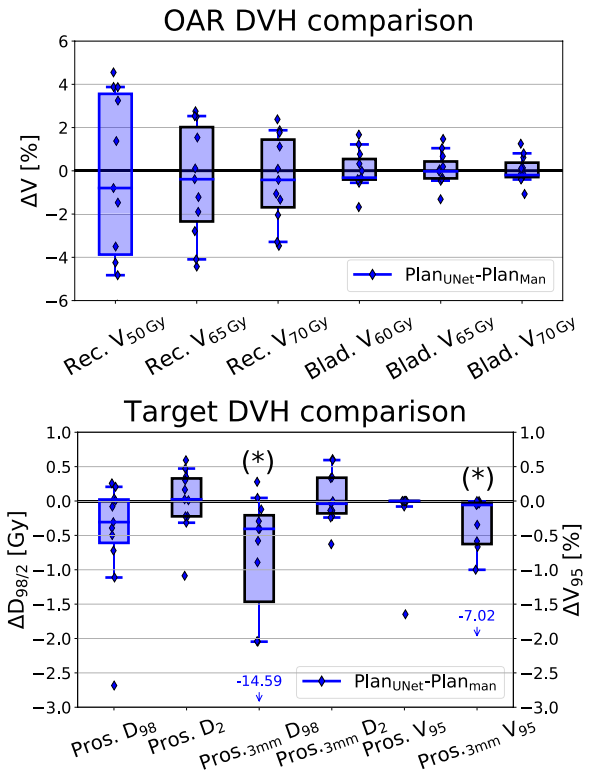

Fig. 5 Differences between the two treatment plans optimized using U-Net- and manually-generated contours. DVH parameters for (top) OARs (bladder and rectum) and (bottom) TV (prostate, prostate + 3 mm margin). Each patient is represented by a single data point, while whiskers indicate the 5th–95th percentile. Dose values correspond to the total dose of 74 Gy delivered during fractionated treatment. Asterisks indicate the statistically significant differences

bladder and 0.84 for rectum. The edge-calibrated multi-task network by Tong et al. [16] showed an overall bladder, rectum, and prostate segmentation performance of DSC = 0.89. The UNet-GAN hybrid architecture by Sultana et al. [17] achieved DSC = 0.90 for prostate. A more detailed comparison is shown in Table 4. In all studies, bladder achieved the highest segmentation accuracy, followed by prostate and rectum.

In the current work, 1 patient with a metal hip implant and 9 patients with fiducial markers were excluded from the study due to artifacts. Applying the trained network to these cases resulted in a DSC of 0.60 (7) for prostate and average Hausdorff distance of 32.5 (8) mm, demonstrating that the trained network cannot be used for images with such artifacts. The available 10 cases are neither sufficient to train a separate model nor to expect a visible effect on the training in combination with the other training data-sets (several images would also have to be set aside for validation and testing, further reducing the training dataset). A potential solution to this issue could be collecting a larger database of images with artifacts and carrying out an independent training.

The ground truth bladder and rectum segmentations were assembled over a course of 2.5 years at the LMU Klinikum and originated from several physicians. In contrary, prostate segmentation has been re-drawn for the purpose of this study. Multi-observer contours in the training set might be seen as an advantage, as the network learns how to generalize and does not adjust to the contouring style of one physician only. On the other hand this might lead to lower testing outcomes, since the network predictions compared against contours drawn by different physicians will be ranked differently. This also

Table 4 Quantitative comparison of geometric metrics with state-of-the-art segmentation algorithms

	Present work	Balagopal et al. [14]	Sultana et al. [17]	Tong et al. [16]
Prostate				
DSC	0.87 ± 0.03	0.90 ± 0.02	0.90 ± 0.05	0.86 ± 0.06
HD_{avg}	1.6 ± 0.4	–	1.56 ± 0.37	1.01 ± 0.65
$HD_{95\%}$	4 ± 1	–	5.21 ± 1.2	3.51 ± 1.66
Bladder				
DSC	0.96 ± 0.01	0.95 ± 0.02	0.95 ± 0.02	0.96 ± 0.02
HD_{avg}	0.95 ± 0.2	–	0.95 ± 0.15	0.97 ± 0.53
$HD_{95\%}$	2.5 ± 0.5	–	4.37 ± 0.56	3.17 ± 3.61
Rectum				
DSC	0.89 ± 0.04	0.84 ± 0.04	0.84 ± 0.04	0.86 ± 0.07
HD_{avg}	1.4 ± 0.7	–	1.78 ± 1.3	1.22 ± 1.05
$HD_{95\%}$	5 ± 4	–	6.11 ± 1.5	4.34 ± 5.30

sets an upper limit on the network performance measured by means of geometric metrics which is in the order of the expectable inter-observer differences [26].

Due to GPU memory limitations, images were cropped around the prostate center of mass, causing truncation of bladder and rectum parts in some cases. On the one hand, this could have made it easier to predict the outer walls, on the other hand, this reduced the organ volume. Since these factors have the opposite effect on DSC and are small in themselves, the effect on DSC is deemed negligible, while the value of HD might have been slightly underestimated. The truncated sections were always located in the low dose region and therefore dosimetric analysis and plan optimization were not affected.

In the scope of the additional dosimetric analysis, target volume D_{98}, D_2 and V_{95} of the plans optimized using 3D U-Net contours were found to differ only slightly from the reference plans based on expert delineations, however a trend of lower D_{98} and V_{95} was observed as shown in Fig. 5. In only one case (patient 59), major deviations, i.e. $D_{98} = -14.59$ Gy and $V_{95} = -7.02\%$ for surrogate CTV, were observed. This can be attributed to an incorrect prostate contouring that is shifted towards the bladder, as can be seen in Fig. 1.

The average value of the CI was 0.78 (0.06) for the plans optimized on 3D U-Net generated contours and 0.85 (0.03) for reference plans. The lower value of the average CI confirms slightly worse target coverage. The treatment plans derived from automatic contours yielded lower CI since the evaluation was performed using the ground truth contours. In contrary, the reference plans have been optimized and evaluated on the same set of contours, and are thus biased towards higher values by design.

Due to the lack of an absolute reliability of the automatic segmentation, human review is still unavoidable. Nonetheless, introducing a method that has a potential to accelerate the contouring process in the majority of cases, as it was show in [27] or in a similar study considering lung cancer patients [28], would be an improvement with respect to current clinical practice.

Analysis of DVH parameters for rectum showed that treatment plans optimized on 3D U-Net-generated contours did not result in statistically significant differences measured by $V_{50/65/70\ \mathrm{Gy}}$. No statistically significant differences were found for the bladder as well. Results indicate that plans optimized on automatically generated contours do not overdose the neighboring OARs, i.e. bladder and rectum.

The gamma index analysis resulted in pass rates of 71–94% with a mean value of 85%. The most prominent differences between dose distributions have been detected close to the PTV border. The degree of the discrepancies correlates closely with the discrepancies between PTV borders (ground truth and predicted) as steep dose gradients are desirable during dose optimization. Thus, the main organs affected by these differences were the bladder and the rectum, for which the most relevant DVH indices have been carefully analyzed in this study. Inside the PTV we did not observe any 'hot-spots' exceeding 107% of the prescribed dose. We also did not notice any consistent dose clustering outside of the PTV. The maximum dose delivered to femoral heads was always below 35 Gy, which is significantly lower than the recommended threshold of 50 Gy.

The only statistically significant correlation was found between the DSC of the prostate and the gamma index. The Pearson coefficient showed a moderately positive correlation only. No statistically significant correlation was found between the gamma pass-rate and the DSC values of OARs and between the DVH parameters and the DSC. On the contrary, we have observed that it is not uncommon for patients to show a very similar DSC for the prostate, which is the most important segmentation in relation to the treatment planning of prostate cancer, while showing a very different gamma pass-rate e.g. $\mathrm{DSC}_{\mathrm{Pat.43}} = \mathrm{DSC}_{\mathrm{Pat.90}} = 0.85$ while $\gamma_{\mathrm{Pat.43}} = 93$ and $\gamma_{\mathrm{Pat.90}} = 74$ or $\mathrm{DSC}_{\mathrm{Pat.44}} = 0.88$, $\mathrm{DSC}_{\mathrm{Pat.81}} = 0.91$ while $\gamma_{\mathrm{Pat.44}} = 94$ and $\gamma_{\mathrm{Pat.81}} = 87$. This leads to the conclusion, that a high geometric similarity between contours, commonly evaluated by the means of DSC, does not necessarily result in a high fidelity dose distribution optimized using these contours. Since eventually, the dosimetric analysis is clinically more relevant the results of this study highlight that the latter should always be carried out in addition to the geometric analysis.

Another important factor to consider is the contour conversion between two formats: the point cloud format (DICOM RT-Struct) required by the contouring software as well as the TPS, and the binary masks required for CNN training. The use of nearest neighbors interpolation in the conversion pipeline did not introduce any noticeable differences during structure conversion.

One possible improvement to this study could be to prepare separate training images for the bladder and rectum by cropping images around their mass centers and adjusting the soft tissue window to match closer their HU range. This could help create more precise contours, but should not significantly affect the dosimetric analysis as the parts of the OAR structures relevant for treatment planning are located in close vicinity of the prostate, which was used as center for cropping in this study. Furthermore, prostate patients with

tumor stages III and IV could be included in future studies by including seminal vesicles in the prostate contour or training a separate network. However, this is a challenging task since in clinical practice the CTV/PTV might contain different proportions of seminal vesicles depending on the exact tumor stage. Therefore, the CTV/PTVs including the seminal vesicles might have more pronounced variations between patients and thus more training data would be required.

Conclusions

A 3D U-Net was successfully trained for organ segmentation on CT images of the male pelvic region. The geometric accuracy measured with DSC, mean and 95% HD showed state-of-the-art performance of our algorithm. Analysis based on clinically relevant DVH parameters of VMAT plans did not show excessive dose enhancement to OARs and proved sufficient for treatment target volume coverage in nine out of ten cases. Nevertheless, the gamma pass rate was not always acceptable, indicating that human review is crucial. No strong statistically relevant correlation between geometric and dosimetric metrics was observed, suggesting that both types of analysis should be included in the evaluation of automatic organ segmentation in the scope of radiotherapy.

Abbreviations

3D U-Net: 3 dimensional U-Net architecture; CI: Conformity index; CNN: Convolutional neural network; CT: Computed tomography; CTV: Clinical target volume; DICOM: Digital imaging and communications in medicine; DSC: Dice similarity coefficient; DVH: Dose-volume histogram; GAN: Generative adversarial network; GPU: Graphics processing unit; HD: Hausdorff distance; HD_{avg}: average Hausdorff distance; 95% HD or $HD_{95\%}$: 95th percentile Hausdorff distance; HU: Hounsfield unit; RT: Radiation therapy; MR: Magnetic resonance; OAR(s): Organ(s) at risk; PPV: Positive prediction value; PReLU: Parametric rectified linear unit; PTV: Planning target volume; TPS: Treatment planning system; TV: Target volume; VMAT: Volumetric Modulated Arc Therapy.

Acknowledgements

The first author wishes to thank Martin Rädler for help in creating figures for this manuscript.

Authors' contributions

MK trained the final networks, performed data analysis, comparison of geometric and dosimetric metrics and was a major contributor in writing the manuscript. DP adapted the 3D U-Net code to the CT data, performed data preprocessing, data augmentation and hyperparameter optimization. ML supervised prostate recontouring for the whole dataset. GV and AA provided the core part of the 3D U-Net implementation. KP and CB reviewed the manuscript and helped to finalize it. GL and CK designed the study, participated in all stages of this work from data preparation, network training, data analysis and writing the manuscript. All authors read and approved the final manuscript.

Funding

This work was supported by the Wilhelm Sander-Stiftung (2019.162.1) and the German Research Foundation (DFG) within the Research Training Group GRK 2274.

Availability of data and materials

The patient data will not be available due to missing ethics approval for public sharing. V-Net code: https://github.com/faustomilletari/VNet.

Declarations

Ethics approval and consent to participate

This retrospective study was exempt from requiring ethics approval. Bavarian state law (Bayrisches Krankenhausgesetz/Bavarian Hospital Law §27 Absatz 4 Datenschutz (Dataprotection)) allows the use of patient data for research, provided that any person's related data are kept anonymous. German radiation protection laws request a regular analysis of outcomes in the sense of quality control and assurance, thus in the case of purely retrospective studies no additional ethical approval is needed under German law.

Consent for publication

Not applicable.

Competing interests

The Department of Radiation Oncology of the University Hospital of the LMU Munich has ongoing researchagreements with Elekta Inc., Brainlab GmbH and ViewRay Inc.

Author details

[1]Department of Radiation Oncology, University Hospital, LMU Munich, Munich, Germany. [2]Department of Medical Physics, Faculty of Physics, Ludwig-Maximilians-Universität München, Garching, Germany. [3]German Center for Vertigo and Balance Disorders, Ludwig-Maximilians-Universität München, Planegg, Germany. [4]German Cancer Consortium (DKTK), Munich, Germany.

Received: 26 July 2021 Accepted: 10 January 2022
Published online: 31 January 2022

References

1. Hummel S, Simpson E, Hemingway P, Stevenson M, Rees A. Intensity-modulated radiotherapy for the treatment of prostate cancer: a systematic review and economic evaluation. Health Technol Assess. 2010;14(47):1–108.
2. Guckenberger M, Flentje M. Intensity-modulated radiotherapy (IMRT) of localized prostate cancer. Strahlenther Onkol. 2007;183(2):57–62.
3. Chen MJ, Weltman E, Hanriot RM, Luz FP, Cecílio PJ, Da Cruz JC, et al. Intensity modulated radiotherapy for localized prostate cancer: Rigid compliance to dose-volume constraints as a warranty of acceptable toxicity? Radiat Oncol. 2007;2(1):1–7.
4. Wu QJ, Thongphiew D, Wang Z, Mathayomchan B, Chankong V, Yoo S, et al. On-line re-optimization of prostate IMRT plans for adaptive radiation therapy. Phys Med Biol. 2008;53(3):673.
5. McVicar N, Popescu IA, Heath E. Techniques for adaptive prostate radiotherapy. Physica Med. 2016;32(3):492–8.
6. Choi H, Kim Y, Lee S, Lee Y, Park G, Jung J, et al. Inter- and intra-observer variability in contouring of the prostate gland on planning computed tomography and cone beam computed tomography. Acta Oncol (Stockh Swed). 2011;03(50):539–46.
7. Nyholm T, Jonsson J, Söderström K, Bergström P, Carlberg A, Frykholm G, et al. Variability in prostate and seminal vesicle delineations defined on magnetic resonance images, a multi-observer, -center and-sequence study. Radiat Oncol. 2013;8(1):1–12.
8. Wong J, Fong A, McVicar N, Smith S, Giambattista J, Wells D, et al. Comparing deep learning-based auto-segmentation of organs at risk and clinical target volumes to expert inter-observer variability in radiotherapy planning. Radiother Oncol. 2020;144:152–8.
9. Cardenas CE, Yang J, Anderson BM, Court LE, Brock KB. Advances in auto-segmentation. In: Seminars in radiation oncology. Elsevier; 2019. p. 185–97.
10. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553):436–44.
11. Savenije MH, Maspero M, Sikkes GG, van der Voort JR, van Zyp TJ, Kotte AN, Bol GH, et al. Clinical implementation of MRI-based organs-at-risk

auto-segmentation with convolutional networks for prostate radiotherapy. Radiat Oncol. 2020;15:1–12.
12. Chung SY, Chang JS, Choi MS, Chang Y, Choi BS, Chun J, et al. Clinical feasibility of deep learning-based auto-segmentation of target volumes and organs-at-risk in breast cancer patients after breast-conserving surgery. Radiat Oncol. 2021;16(1):1–10.
13. Milletari F, Navab N, Ahmadi SA. V-net: fully convolutional neural networks for volumetric medical image segmentation. In: 2016 Fourth international conference on 3D vision (3DV). IEEE; 2016. pp. 565–71.
14. 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.
15. Wang S, He K, Nie D, Zhou S, Gao Y, Shen D. CT male pelvic organ segmentation using fully convolutional networks with boundary sensitive representation. Med Image Anal. 2019;54:168–78.
16. Tong N, Gou S, Chen S, Yao Y, Yang S, Cao M, et al. Multi-task edge-recalibrated network for male pelvic multi-organ segmentation on CT images. Phys Med Biol. 2021;66(3):035001.
17. Sultana S, Robinson A, Song DY, Lee J. Automatic multi-organ segmentation in computed tomography images using hierarchical convolutional neural network. J Med Imaging. 2020;7(5):055001.
18. Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial networks. arXiv preprint arXiv:1406.2661. 2014;.
19. Taha AA, Hanbury A. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. BMC Med Imaging. 2015;15(1):1–28.
20. Guo H, Wang J, Xia X, Zhong Y, Peng J, Zhang Z, et al. The dosimetric impact of deep learning-based auto-segmentation of organs at risk on nasopharyngeal and rectal cancer. Radiat Oncol. 2021;16(1):1–14.
21. Sharp GC, Li R, Wolfgang J, Chen G, Peroni M, Spadea MF, et al. Plastimatch: an open source software suite for radiotherapy image processing. In: Proceedings of the XVI'th international conference on the use of computers in radiotherapy (ICCR), Amsterdam, Netherlands. 2010.
22. Marks LB, Yorke ED, Jackson A, Ten Haken RK, Constine LS, Eisbruch A, et al. Use of normal tissue complication probability models in the clinic. Int J Rad Oncol Biol Phys. 2010;76(3):S10–9.
23. Prescribing I. recording, and reporting intensity-modulated photon-beam therapy (IMRT) (ICRU Report 83). J ICRU. 2010;10(1):555–9.
24. Paddick IA. simple scoring ratio to index the conformity of radiosurgical treatment plans. J Neurosurg. 2000;93(supplement–3):219–22.
25. Benesty J, Chen J, Huang Y, Cohen I. Pearson correlation coefficient. In: Noise reduction in speech processing. Springer; 2009. p. 1–4.
26. Sanders J, Mok H, Tang C, Hanania A, Venkatesan A, Bruno T, et al. Benchmarking automatic segmentation algorithms against human interobserver variability of prostate and organs at risk delineation on prostate MRI. Int J Radiat Oncol Biol Phys. 2021;111(3):e291–2.
27. Zabel WJ, Conway JL, Gladwish A, Skliarenko J, Didiodato G, Goorts-Matthews L, et al. Clinical evaluation of deep learning and atlas-based auto-contouring of bladder and rectum for prostate radiation therapy. Pract Radiat Oncol. 2021;11(1):e80–9.
28. 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.

Publisher's Note

5.2 Paper II

The second paper focused on OAR and CTV auto-segmentation for prostate cancer patients undergoing MRgRT at the MRIdian MR-LINAC. The data utilized in this study comprised manually segmented planning and fraction MRIs from two facilities: the LMU University Hospital in Munich and the Gemelli University Hospital in Rome. Several aspects were investigated. First, how good are the DL-generated contours, and do they require manual corrections before clinical implementation? Are DL contours better for fraction images than the currently employed planning contours registered to the daily MRI in the TPS? Second, can networks trained on images from one facility be used directly at a different one? Or is additional fine-tuning with the data from the new target facility necessary? Third, can personalized (at a single patient level) auto-segmentation networks surpass conventional population models? The third question was inspired by the observation that the prostate CTV is defined differently depending on the patient. Differences in contouring are also present at the rectum's superior end. However, it was crucial to observe that the planning images are segmented in a certain way, and this "style" is generally kept throughout the fractions. To recapitulate, the third question was aimed at exploring if the segmented planning MRI can be used to teach the network the expected shape of a given structure.
This is the first work that considered patient-specific models for prostate cancer patients treated at the MRIdian MR-LINAC. Moreover, this study is the first to compare DL-generated contours to TPS suggestions, evaluating the potential benefits of introducing them for online treatment adaptation.
The architecture of choice was a state-of-the-art residual U-Net. It was trained for single-class segmentation of the bladder, rectum (OARs), and CTV. The baseline population models were trained with data from one facility only. Subsequently, facility-specific models were trained by fine-tuning the population models with a subset of images from the target facility. Finally, personalized models were obtained by fine-tuning the population models with the segmented planning image of each patient. All methods were assessed using geometric metrics. Additionally, an experienced radiation oncologist reviewed the suitability of population model OAR contours for treatment adaptation and compared them to the currently used ones that are provided by the TPS.
The study yielded the following conclusions corresponding to the three primary study aspects. First, the trained population models achieved satisfactory geometric performance for all investigated structures. The DL predictions of OARs were graded better and were shown to require fewer manual corrections than the currently used TPS contours. Second, the facility-specific fine-tuning did not substantially improve OAR segmentation compared to population models. Thus, it is possible to use OAR models trained at one institute directly at the other. However, this conclusion did not extend to the CTV as CTV population models benefited noticeably from the facility-specific fine-tuning. The third part of the results corresponded to the personalized training. It was shown that patient-specific models were especially beneficial for CTV segmentation and in cases of unusual OAR shapes. However, they were more sensitive to inaccuracies and errors in the planning segmentation.

Received: 3 May 2022 | Revised: 23 September 2022 | Accepted: 25 September 2022

DOI: 10.1002/mp.16056

RESEARCH ARTICLE

MEDICAL PHYSICS

Patient-specific transfer learning for auto-segmentation in adaptive 0.35 T MRgRT of prostate cancer: a bi-centric evaluation

Maria Kawula[1] | **Indrawati Hadi**[1] | **Lukas Nierer**[1] | **Marica Vagni**[2] | **Davide Cusumano**[2] | **Luca Boldrini**[2] | **Lorenzo Placidi**[2] | **Stefanie Corradini**[1] | **Claus Belka**[1,3] | **Guillaume Landry**[1] | **Christopher Kurz**[1]

[1]Department of Radiation Oncology, University Hospital, LMU Munich, Munich, Germany

[2]Fondazione Policlinico Universitario "Agostino Gemelli" IRCCS, Rome, Italy

[3]German Cancer Consortium (DKTK), Munich, Germany

Correspondence
Christopher Kurz, Department of Radiation Oncology, University Hospital, LMU Munich, Munich, Germany.
Email: Christopher.Kurz@med.uni-muenchen.de

Funding information
Wilhelm Sander-Stiftung, Grant/Award Number: 2019.162.1

Abstract

Background: Online adaptive radiation therapy (RT) using hybrid magnetic resonance linear accelerators (MR-Linacs) can administer a tailored radiation dose at each treatment fraction. Daily MR imaging followed by organ and target segmentation adjustments allow to capture anatomical changes, improve target volume coverage, and reduce the risk of side effects. The introduction of automatic segmentation techniques could help to further improve the online adaptive workflow by shortening the re-contouring time and reducing intra- and inter-observer variability. In fractionated RT, prior knowledge, such as planning images and manual expert contours, is usually available before irradiation, but not used by current artificial intelligence-based autocontouring approaches.

Purpose: The goal of this study was to train convolutional neural networks (CNNs) for automatic segmentation of bladder, rectum (organs at risk, OARs), and clinical target volume (CTV) for prostate cancer patients treated at 0.35 T MR-Linacs. Furthermore, we tested the CNNs generalization on data from independent facilities and compared them with the MR-Linac treatment planning system (TPS) propagated structures currently used in clinics. Finally, expert planning delineations were utilized for patient- (PS) and facility-specific (FS) transfer learning to improve auto-segmentation of CTV and OARs on fraction images.

Methods: In this study, data from fractionated treatments at 0.35 T MR-Linacs were leveraged to develop a 3D U-Net-based automatic segmentation. Cohort C1 had 73 planning images and cohort C2 had 19 planning and 240 fraction images. The baseline models (BMs) were trained solely on C1 planning data using 53 MRIs for training and 10 for validation. To assess their accuracy, the models were tested on three data subsets: (i) 10 C1 planning images not used for training, (ii) 19 C2 planning, and (iii) 240 C2 fraction images. BMs also served as a starting point for FS and PS transfer learning, where the planning images from C2 were used for network parameter fine tuning. The segmentation output of the different trained models was compared against expert ground truth by means of geometric metrics. Moreover, a trained physician graded the network segmentations as well as the segmentations propagated by the clinical TPS.

Results: The BMs showed dice similarity coefficients (DSC) of 0.88(4) and 0.93(3) for the rectum and the bladder, respectively, independent of the facility. CTV segmentation with the BM was the best for intermediate- and high-risk cancer patients from C1 with DSC=0.84(5) and worst for C2 with DSC=0.74(7). The PS transfer learning brought a significant improvement in the CTV segmentation, yielding DSC=0.72(4) for post-prostatectomy and low-risk patients and DSC=0.88(5) for intermediate- and high-risk patients. The FS training did not improve the segmentation accuracy considerably. The physician's assessment of the TPS-propagated versus network-generated structures showed a clear advantage of the latter.
Conclusions: The obtained results showed that the presented segmentation technique has potential to improve automatic segmentation for MR-guided RT.

KEYWORDS
0.35 T MR-Linac, adaptive radiotherapy, automatic segmentation, deep learning, patient-specific transfer learning, prostate cancer

1 | INTRODUCTION

The introduction of magnetic resonance (MR) linear accelerators (Linacs) into clinical practice has facilitated online adaptive radiotherapy.[1–4] Fully integrated daily MR imaging enables fast dose re-optimization based on the anatomy of the day, which has the potential to improve tumor coverage and reduce gastrointestinal and genitourinary toxicity in abdominal and pelvic targets.[5,6] With the current state-of-the-art, these benefits come at the cost of longer workflows, notably due to the need for online re-contouring.[7] The median fraction time excluding the irradiation itself can be as long as 30 min, as presented by Sahin et al.[8] for 500 fractions delivered to 72 patients. Other studies reported 54 min for adapted abdominal and pelvic stereotactic body radiotherapy (SBRT) fractions,[9] 50 min for liver tumors,[10] and up to 71 min in MR-guided SBRT boosts for gynecological cancer patients.[11] During the adaptation process at 0.35 T MR-Linacs (MRIdian, ViewRay Inc, Cleveland, OH),[12] the planning MRI is matched to the daily MRI using deformable image registration (DIR) and subsequently the planning contours are propagated to the anatomy of the day using either the same deformation field or rigid registration for the CTV. The propagated structures are corrected manually by radiation oncologists and only then can be used for dose evaluation and optimization. An automatic or semi-automatic segmentation, which requires no or fewer corrections, has the potential to shorten the treatment time and thus increase patient throughput at MR-Linacs.[13–15] It could also help to avoid the inter- and intra-physician variability caused by work under time pressure, fatigue and the level of individual experience.[16]

Several studies have been conducted to address the problem of auto-contouring in cancer patients by means of state-of-the-art machine learning techniques in the scope of MR-guided radiation therapy (MRgRT). Liang et al.[17] described an approach regarding abdominal multi-organ auto-contouring integrating information from the manually segmented simulation 0.35 T MR images with predictions generated by a support vector machine (SVM). Fu et al.[18] presented an architecture comprising a segmentation convolutional neural network (CNN) followed by two correction CNNs that was trained for liver, kidney, stomach, bowel, and duodenum automatic delineation for MRgRT. Eppenhof et al.[19] proposed a CNN for contour propagation based on DIR during fractionated prostate cancer treatment at a 1.5 T MR-Linac system. The architecture implemented by Eppenhof et al. is a UNet which is frequently used for organ segmentation and broadly discussed in the literature.[20] Friedrich et al.[21] investigated the stability of conventional and machine learning-based 2D tumor auto-segmentation techniques for 2D tumor tracking at a 0.35 T MR-Linac.

However, until now there are very few studies that leverage the scheme of fractionated MRgRT at MR-Linacs, and the available prior knowledge such as initial treatment planing segmentation. For online plan adaptation, prior knowledge could be beneficial for organ segmentation in patients with unusual anatomies or for clinical target volume (CTV) delineation, since the latter does not necessarily follow visible organ boundaries and requires additional clinical information.

The aim of this work was to use a 3D U-Net architecture[22] with customized data augmentation to generate organs-at-risk (OARs), that is bladder and rectum, and CTV segmentation for prostate cancer patients treated at a 0.35 T MR-Linac. In order to investigate the transferability of trained models, the network performance was additionally tested with data from an independent facility which operates the same MR-Linac. Furthermore, the network-generated contours were compared with the structures automatically propagated by the treatment planning system (TPS) during the online adaptive MRgRT workflow and graded with regard to their clinical usability for treatment adaptation. Facility-specific (FS) transfer learning has been performed to test if the trained baseline neural network can

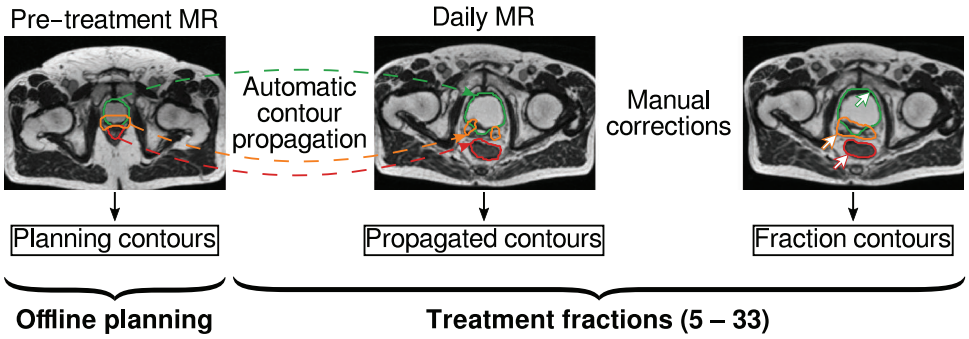

FIGURE 1 Illustration of the adaptive radiotherapy workflow at the MRIdian presenting the different types of contours incorporated in the study.

improve its performance on data from an independent facility by adapting to the specific segmentation style as suggested by Balagopal et al.[23] Finally and most importantly, patient-specific (PS) transfer learning was carried out in order to investigate whether incorporating prior knowledge, as typically available in fractionated adaptive MRgRT, further improves segmentation performance for fraction images.[24]

2 | MATERIALS AND METHODS

2.1 | Database

A total of 92 prostate cancer patients treated between January 2018 and June 2021 with online adaptive MRgRT at the Department of Radiation Oncology of the University Hospital of the LMU Munich (19 patients) and the Gemelli University Hospital in Rome (73 patients) were included in this study. At both facilities, MR imaging was performed at the ViewRay 0.35 T MRIdian MR-Linac system. The images were acquired using the clinical balanced steady-state free-precession (bSSFP) sequence resulting in a T2*/T1 image contrast, and had a resolution of 1.5 mm × 1.5 mm × 1.5 mm or 1.5 mm × 1.5 mm × 3 mm.[12] The latter were resampled to 1.5 mm × 1.5 mm × 1.5 mm in the scope of this study, using the plastimatch *convert*[25] function with nearest neighbor interpolation.

All patients were treated following a similar workflow (Figure 1), which consisted of an initial offline planning phase and irradiation in 5–33 fractions. After the acquisition of a planning MR image, OARs, including the bladder and the rectum, as well as the CTV were manually delineated by trained consulting physicians (*planning contours*). The CTV was defined as a volume of tissue that contains a demonstrable gross target volume and/or sub-clinical malignant disease at a certain probability considered relevant for therapy. Depending on the tumor development, different regions of the seminal vesicles were included in the CTV: none for low-, proximal for intermediate- (int) and entire for high-risk prostate cancer. There were no other additional differences in contouring between the risk groups. A separate subgroup comprises post-prostatectomy (pp) patients. For them, the CTV includes only the remaining parts of the prostate and seminal vesicles after surgery, which makes them visibly different from the rest of the patients. Then, the planning target volume (PTV) was generated as a CTV expansion by 4 mm/posterior 3 mm at the LMU Hospital and isotropically by 5 mm at Gemelli Hospital (which due to the TPS rounding to a full pixel size of 1.5 mm^3 results in 4.5 mm/3 mm at LMU and 4.5 mm for Gemelli) and clinical treatment plans were created. At each fraction, a daily MRI was acquired with the same imaging sequence as the one used for the offline planning and rigidly aligned with the pre-treatment image. The planning MRI was then matched to the fraction image with DIR and the planning structures were propagated by the ViewRay TPS using the same DIR for all OARs, while the CTVs were propagated using rigid registration, according to the clinical guidelines followed in our institutes. The resulting contours will be referred to as *propagated contours*. Subsequently, they were inspected by a physician and, if necessary, corrected, which led to the final *fraction contours*. These were used for adaptation of the daily treatment plan, if deemed necessary. After dose re-optimization, a new plan was delivered.

All contours were initially stored in the DICOM RT-struct format, which represents structures as point clouds. The segmentations were converted into binary masks using plastimatch[25] with nearest neighbors interpolation, in order to be suitable for the subsequent neural network training. The image-binary mask pairs were cropped/padded around the PTV center to a size of 220 × 220 × 220 pixels, which in all but one case, covered all structures of interest with a substantial margin. The exception case had a part of the bladder cropped.

Throughout this work, the planning and fraction contours, as generated and approved by the radiation oncologists, were considered as ground truth, while the propagated structures were used only for comparison in the evaluation phase. The Gemelli dataset, cohort 1 (C1), consisted exclusively of planning MRs and corresponding manual expert delineations, while the LMU

TABLE 1 Datasets used in the study.

	Cohort	Type	Stage	Number
OARs	C1	Planning	–	73
	C2	Planning	–	19
		Fraction	–	240
		Propagated	–	24 (5 patients)
CTV	C1	Planning	pp & low	10
		Planning	int & high	57
	C2	Planning	pp & low	8
		Fraction	pp & low	91
		Planning	int & high	11
		Fraction	int & high	144

Note: For each subgroup, the origin of the data (C1 or C2), the type of contours (planning, fraction, or propagated, see Figure 1), and the number of images available are given. For the CTV, it was differentiated between intermediate- and high-risk patients (int & high) and the remaining cases, that is, post-prostatectomy (pp) and low-risk (low) patients.

dataset, cohort 2 (C2), included planning as well as fraction images along with their contours. Propagated OAR contours were available for a subset of C2 patients, in addition to expert delineations on each image. Table 1 summarizes the characteristics of the dataset.

2.2 | 3D U-net

In this work, the MONAI[26] implementation of the residual U-Net developed by Kerfoot et al.[27] was used. The network follows the well-known architecture with encoding and decoding arms linked at each level via skip connections. The network consists of five levels. Each of them contains two convolutions with $3 \times 3 \times 3$ kernels, followed by instance normalization[28] and PReLU[29] activation with the initial slope for negative arguments of 0.2. In the encoding arm, the second convolution has a stride of 2 serving also for down-sampling, while in the decoding arm a transpose convolution is used for up-sampling. The output layer of the network has soft-max activation[30] and thresholding at 0.5, which generates a binary image corresponding to the predicted structure. A loss function based on the dice similarity coefficient (DSC)[31] and the Adam[32] optimizer were employed throughout the training.

2.3 | Data augmentation and preprocessing

The data augmentation applied during training included random spatial transformations such as rotations, translations, scaling, B-Spline deformation, along with MR-specific random transformations mimicking the occurrence of bias fields, motion artifacts, and noise. To harmonize the data fed into the network an intensity normalization based on image mean and standard deviation, followed by scaling to the (0, 1) range was applied to all images (training, validation, testing). Finally, the image and binary mask pairs were centrally cropped to the size of $192 \times 192 \times 192$ pixels, while the pixel spacing of 1.5 mm × 1.5 mm × 1.5 mm was preserved. In all but one patient (with bladder extending exceptionally high in the superior direction), the cropping resulted in images with substantial margins around the structures of interest. Further details on the data augmentation and hyperparameter tuning are given in the Supporting information.

2.4 | Baseline training

A single optimal combination of hyperparameters was sought while training three independent models for the segmentation of bladder, rectum, and CTV. Since there was a non-zero overlap between some structures, for example, bladder and CTV or rectum and CTV, and based on previous experience, no multi-organ segmentation was performed. At this point, only C1 patients were included in order to provide an independent test cohort (C2) in the later evaluation phase and PS training was not considered. For OARs, the C1 data split was 53/10/10 for training, validation, and testing. However, six cases had to be excluded from the validation and test sets in the case of CTV segmentation, as the tumor was located outside the prostate gland (e.g., lymphatic pathways), which led to a division of 53/7/7. Approximately 90% of the cases were intermediate- and high-risk patients, meaning that the CTV contained at least parts of the seminal vesicles in most cases. Therefore, the baseline CTV model is considered suitable for the intermediate- and high-risk cases, and its performance for low-risk and pp patients will be tested only to allow comparison at later stages during PS training. The relatively small number of low-risk and pp patients in the training set did not affect the network performance on the remaining cases, therefore they were not excluded.

2.5 | Baseline models evaluation

The performance of the baseline models (BMs) was tested separately on three data subsets: 10 planning C1 images that were not used for training, 19 C2 planning images, and 240 C2 fraction images. Again, for the BM evaluation we did not consider PS training.

2.6 | Network-predicted versus treatment planning system-propagated contours

During treatment adaptation, propagated contours are available to physicians and form the basis for their

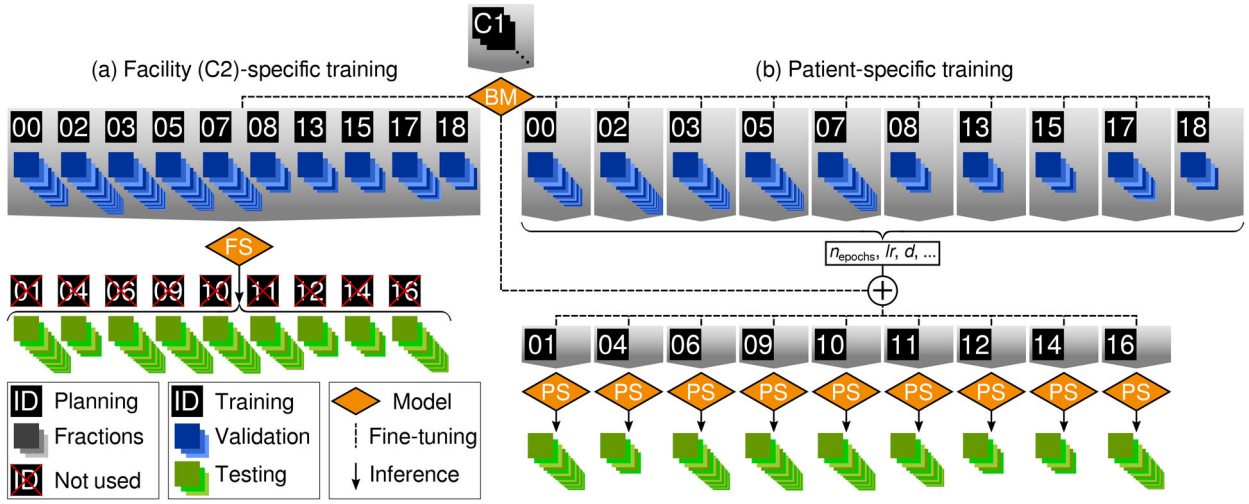

FIGURE 2 Representation of the training scheme as well as the patients (ID) split for (a) the facility-specific (FS) and (b) the patient-specific (PS) training. The gray background indicates images considered together for model training and validation. Both variants share the same test set. The depicted frames and the patient IDs show the actual data base and the training/validation/testing split.

corrections. Due to their potentially insufficient quality, the contours have to be checked and adjusted manually most of the time, which prolongs the treatment. The aim of this section was to compare the quality of the propagated contours with the network predictions and to determine which would potentially require less corrections.

The ground truth fraction delineations were generated from the propagated contours by applying manual corrections. Under time pressure physicians mostly correct pronounced errors of the propagated structures, which means that they may artificially be closer to the propagated contours, introducing a considerable bias in favor of propagated contours evaluated by means of DSC or HD. Therefore, an additional qualitative analysis investigating contour usability during plan adaptation has been carried out. Please note, that prior to contour propagation, the planning and fraction images are rigidly aligned and it is ensured, that the MR scanner/Linac isocenter is roughly at the center of the PTV.

The propagated contours were retrieved for 24 fractions from 5 patients of C2. A radiation oncologist working at the LMU MR-Linac was presented two sets of contours in random order: the predicted and the propagated, for each fraction. First, the physician was asked to choose the contour considered more useful during plan adaptation, and secondly, to rate each delineation on a four-point scale: 1-ready to use, 2-small corrections required, 3-major corrections required, and 4-not useful.[33] In order to eliminate personal bias, the physician was neither informed about the study goal nor the origin of the examined delineations. Since CTV segmentation requires additional knowledge, such as the patient's medical record and cancer risk category, this analysis was restricted to the OARs.

2.7 | Facility- and patient-specific transfer learning

The study also aimed at investigating whether transfer learning can improve segmentation accuracy in fraction images. Two approaches have been taken: FS and PS transfer learning. In both training types, network weights and biases were initialized with parameters of the BM and further trained with a planning image (or images) of interest, adjusting all network parameters. The hyperparameter search was carried out analogously to the BM optimization. In FS transfer learning, the BMs were fine-tuned with a set of planning images from C2, while in PS transfer learning a single C2 planning image for a particular patient was used for fine-tuning. The goal of this approach is to slightly adjust the BM using information from the planning image. The approach is similar to Chun and Park et al.[24] To prevent overfitting to the anatomy seen on the planning image, data augmentation was applied to mimic possible anatomical changes occurring over the following fractions. Figure 2 shows the design of both transfer learning approaches with the data subdivision and patient split.

The FS training was carried out with ten randomly selected patients from C2. Planning images were used to optimize data augmentation, hyperparameters, and fine-tune the network parameters, while the corresponding fraction data were employed for validation. The trained model was tested on the fraction data of the nine remaining C2 patients.

In the PS training, no validation data are available to select the stopping epoch when applying the procedure to test data. Thus, ten separate models were fine-tuned simultaneously for each of the ten preselected training patients (see Figure 2). Again, the planning images

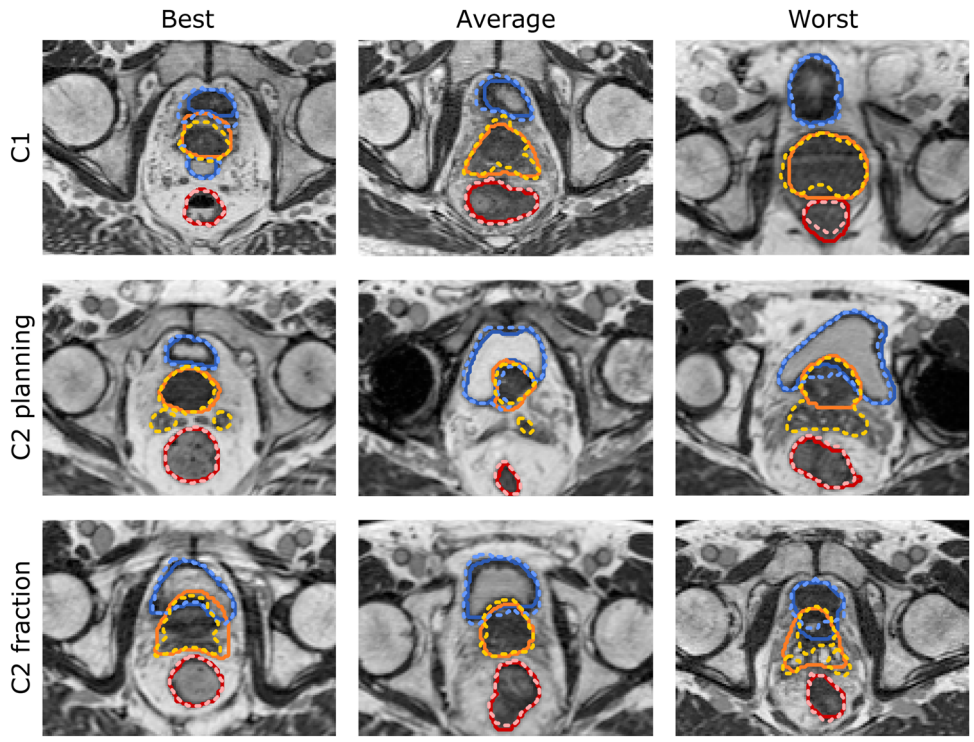

FIGURE 3 Image slices showing (left) one of the best, (middle) average, and (right) worst, baseline model performance. Image slices from (top) C1, (middle) C2 planning, and (bottom) C2 fraction MRs are shown. The (solid line, saturated colors) ground truth and (dashed, faded counterparts) network predictions for the investigated organs (blue) bladder, (orange) prostate, and (red) rectum are presented.

were used for model fine-tuning and the fraction images for validation. Collecting validation results from all 10 patients allowed to adjust the data augmentation, learning rate, and number of training epochs the same for all patients. Finally, models were fine-tuned for the nine test patients using their planning images and fixed hyperparameters. Both FS and PS training shared the same test set of 115 fraction images.

2.8 | Data evaluation

The network predictions were compared to the ground truth via DSC, the 95th percentile and the average Hausdorff distance, HD_{95} and HD_{avg}, respectively. The evaluation of the rectum segmentation considered slices including the PTV and 10 additional slices reaching 1.5 cm above and below the upper and lower PTV ends. We performed the analysis separately for planning and fraction images. The CTV contours for the intermediate- and high-risk cases were considered separately from the post-prostatectomy and low-risk patients, due to the considerable differences in the inclusion of seminal vesicles. To determine whether the differences between different methods or datasets are statistically significant, the Wilcoxon-signed rank test was performed with the p-value < 0.05 being considered statistically significant.

2.9 | Technical details

The network architecture and the training loop were implemented using MONAI,[26] PyTorch,[34] and TorchIO[35] libraries. The computations were carried out in a Docker container built from the projectmonai/monai image version 0.6.0 on Nvidia Quadro RTX 8000 and/or Nvidia RTX A6000 GPUs.

3 | RESULTS

3.1 | Baseline training

The BMs were trained over 300 epochs with a batch size of 2, which required approximately 4 min/epoch and resulted in a training duration of 20 h. The same set of hyperparameters was used for the final training of models for all three organs. The final values and details on the hyper-parameter optimization are given in the Supporting information.

3.2 | Baseline model evaluation

Figure 3 collects exemplary slices showing cases with one of the best, average, and poor network segmentations for the C1 test patients, the C2 planning, and

TABLE 2 Numerical outcomes of the baseline models performance for the OARs and the CTV.

Dataset	N	Bladder DSC HD_{95} (mm) HD_{avg} (mm)	Rectum DSC HD_{95} (mm) HD_{avg} (mm)	N	CTV int&high DSC HD_{95} (mm) HD_{avg} (mm)	N	CTV low&pp DSC HD_{95} (mm) HD_{avg} (mm)
C1	10	0.93(0.03)	0.88(0.03)	5	0.84(0.05)	2	0.82(0.09)
planning		3.7(1.8)	3.6(1.4)		5.2(2.4)		9.2(4.2)
		1.3(0.4)	1.2(0.3)		1.8(0.5)		3.0(1.7)
C2	19	0.93(0.03)	0.88(0.04)	11	0.76(0.06)	8	0.35(0.19)
planning		3.6(3.5)$^{(ss)}$	3.7(1.6)		8.8(3.0)		15(8)
		1.3(0.7)$^{(ss)}$	1.2(0.3)		3.1(0.8)		6.9(5.1)
C2	240	0.90(0.07)	0.87(0.08)	144	0.75(0.06)	91	0.39(0.17)
fraction		6.2(5.6)$^{(ss)}$	4.9(3.3)		8.6(2.8)		14(5)
		1.8(1.1)$^{(ss)}$	1.5(1.0)		3.1(0.8)		6.0(2.8)

Note: Dice similarity coefficient (DSC), average and 95th percentile Hausdorff distance (HD_{avg}, HD_{95}), with (standard deviation of the mean) are presented for a given number *N* of C1 test patients, C2 planning, and C2 fraction images. Low-risk and post-prostatectomy (low & pp) patients were considered separately from the intermediate and high-risk (int & high) cases. The statistically significant pairs are marked with $^{(ss)}$.

the C2 fraction images. The average DSC, HD_{95}, and HD_{avg} comparing the network-generated segmentation and the ground truth delineation are given in Table 2. Apart from the HDs between the planning and fraction bladder contours of C2, there were no statistically significant differences between the three test sets examined. For the rectum, mean DSC was 0.87–0.88 and for the bladder it was 0.90–0.93. For both OARs, the HDs increased for fraction contours compared to the planning images from approximately 3.6–3.7 to 4.9–6.2 mm for the HD_{95} and from 1.2–1.3 to 1.5–1.8 mm for the HD_{avg}. Analysis of the CTV predictions showed the best outcomes for intermediate- and high-risk C1 test patients, that is, DSC=0.84(0.05), HD_{95}=5.2(2.4) mm, and HD_{avg}=1.8(0.5) mm, thus having the same risk category as the majority of patients in the training set. The delineations for the remaining C1 test patients (low-risk and post-prostatectomy) showed a comparable DSC value of 0.82(0.09), yet worse HD_{95} of 9.2(4.2) mm and HD_{avg} of 3.0(1.7) mm. However, these results should be treated with caution, as only two low-risk patients were available for testing and therefore, the results are not statistically significant. Applying the same network to intermediate- and high-risk C2 patients yielded worse results of DSC=0.75(0.06), HD_{95} = 8.8(3.0) mm, and HD_{avg} = 3.1(0.8) mm, regardless of the contour type (fraction or planning). The network performance on the remaining C2 cases, both planning and fraction, yielded worse outcomes of DSC<0.4, HD_{95}=15(8) mm, and HD_{avg}=6.9(5.1) mm. Here as well, no considerable differences between planning and fraction contours were observed.

Figure 4 illustrates the DSC for the C2 cohort, separately for each patient. For the bladder, 10 of 19 test patients consistently showed a DSC above 0.9

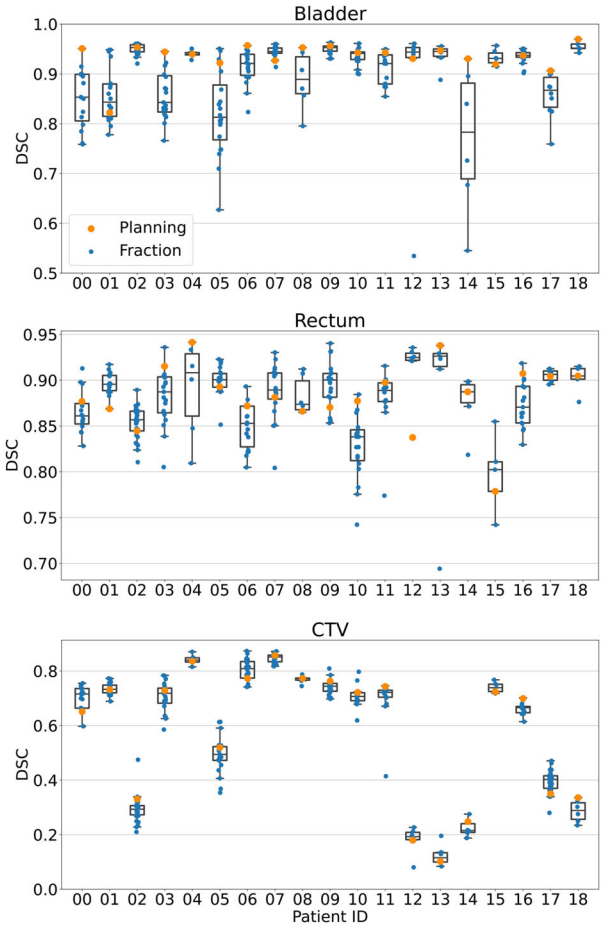

FIGURE 4 The baseline model outcomes. Dice similarity coefficient (DSC) for the bladder, rectum, and clinical target volume (CTV) segmentation for all 19 C2 patients separately. For each patient (horizontal black line) the median value, (orange) performance on the planning data, and (blue) performance on fraction data are marked.

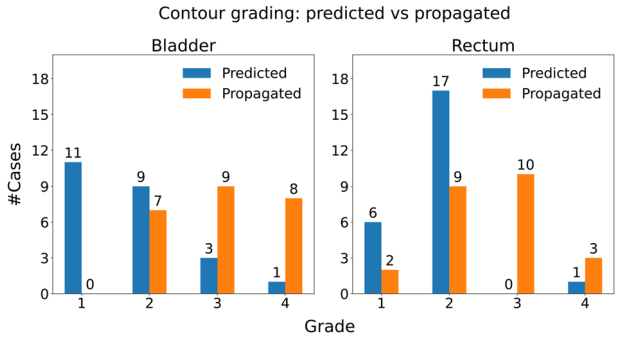

FIGURE 5 Bar plots showing physician's grading of the network predictions (baseline models) and the treatment planning system (TPS)-propagated delineations. The grading is defined as follows: 1—ready-to-use, 2—small corrections, 3—major corrections required, and 4—not useful.

TABLE 3 Quantitative outcomes evaluating the BM-predicted and TPS-propagated OAR contours.

Method	*N*	Bladder DSC HD_{95} (mm) HD_{avg} (mm)	Rectum DSC HD_{95} (mm) HD_{avg} (mm)
Network	24	0.91(0.09)	0.81(0.02)$^{(ss)}$
predicted		1.5(0.9)	2.2(2.9)$^{(ss)}$
		4.1(2.6)	5.8(6.6)$^{(ss)}$
TPS-	24	0.91(0.1)	0.88(0.16)$^{(ss)}$
propagated		1.5(1.3)	1.2(1.7)$^{(ss)}$
		5.2(4.9)	3.4(4.1)$^{(ss)}$

Note: DSC, HD_{avg}, and HD_{95} with (standard deviation of the mean) are given. The statistically significant pairs are marked with $^{(ss)}$. BM, baseline model; DSC, Dice similarity coefficient; OARs, organs at risk, TPS, treatment planning system.

for all planning and fraction images. A slight tendency towards more accurate network contouring on planning compared to fraction images was observed. The considerable DSC variations in several patients, for example, 5 and 14, were caused by the acquisition of some fraction images with an empty bladder, in contrast to the planning stage, when all patients followed closely the clinical recommendations of a filled bladder.

For the rectum, the DSC for most patients was above 0.80 for both planning and fraction data. There was no clear tendency towards better DSC in the planning data.

The CTV segmentation showed the largest variation in the DSC among the three structures examined. All patients with an average DSC < 0.6 were low-risk and post-prostatectomy patients, while those with DSC > 0.6 were intermediate- and high-risk cases. No consistent performance differences were observed between the planning and the fraction MRIs.

3.3 | Network-predicted versus treatment planning system-propagated contours

In the physician examination, the OAR contours generated by the network were preferred over the TPS-propagated contours for the bladder and the rectum in 22 and 23 out of 24 cases, respectively. Figure 5 presents the outcomes of the additional assessment, which graded the contour quality. In almost half of the cases (11 out of 24) the network delineations of the bladder were ready to use directly and further 38% (9 out of 24) required only minor corrections. For the remaining four instances (constituting 17% of the test set), the physician declared the need for major changes or rejection of the predicted contours. On the contrary, none of the propagated contours was considered as ready-to-use and in as many as 17 cases (68%) major corrections would be necessary or the contours were declared not useful.

Similarly, an advantage of the predicted contours over the propagated ones was visible for the rectum. In all cases but one, which was labeled not useful, the predicted rectum contours were either ready-to-use or required only minor corrections. Among the propagated contours, 11 (45% of the cases) needed no or minor corrections, and the remaining 13 (55%) were labeled as requiring major corrections or not useful.

Table 3 presents the quantitative evaluation of the contours. Only the differences for the rectum were statistically significant. It can be observed that the TPS-propagated contours score equally good or even higher in terms of quantitative analysis (see Table 3) and clearly worse in the qualitative assessment (see Figure 5). This can be explained by the potential bias in favor of TPS contours measured by DSC and HD as already described in Section 2.6. Due to this bias, the quantitative results should be interpreted with caution.

3.4 | Facility- and patient-specific transfer learning

Fine-tuning over 500 epochs was found sufficient during training and validation in all cases for both FS and PS transfer learning. The learning rate lr and the maximum displacement d for the B-spline deformation field were decreased in both training variants to $\mathrm{lr} = 10^{-4}$ and $d = 30$ mm compared to the baseline training (see Supporting information). The total training time was 9.5 h and 2 h for the FS and PS models, respectively. Figure 6 and Table 4 collect evaluation outcomes for the nine test patients. No signs of overfitting to the planning image anatomy were observed in any of the ten patients, and training was performed until performance stopped improving on the validation data, that is, the corresponding fraction images.

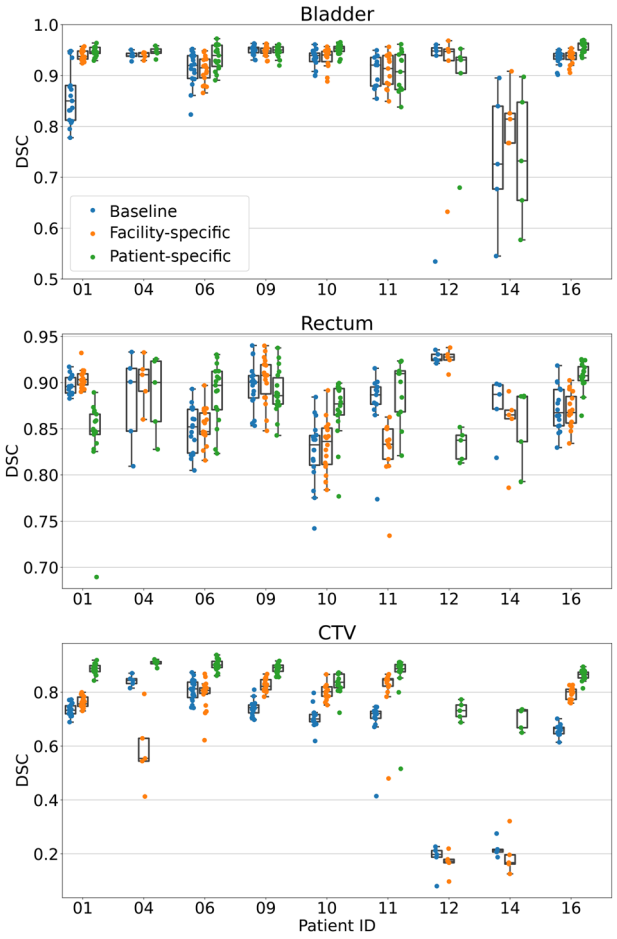

FIGURE 6 Box plots comparing the outcomes of the (blue) baseline, the (orange) facility-, and the (green) patient-specific training for the nine test patients. A single point on the plot represents dice similarity coefficient (DSC) of a predicted fraction contour.

Both types of transfer learning resulted in minor enhancements in the bladder segmentation accuracy. The only exception was patient 01, in which the incorrect inclusion of a substantial part of the surrounding tissue has been corrected for. In the remaining eight instances, patients with a wider range of the DSC values on the BM showed also a similar spread in both transfer learning variants.

The PS training was helpful to adjust the top and the bottom of the rectum according to the planning contours. This resulted in DSC improvements in patients 06, 10, and 16. However, by design, the training was prone to major differences between planning and fraction anatomy, for example, due to different rectum filling, which was the case for patients 01 and 12.

A clear benefit was observed in case of the CTV for PS training, which can be seen in Figure 7 and is summarized in Table 4. The average DSC improved by 0.52 for low-risk and post-prostatectomy cases (patients 12 and 14) and by 0.14 for intermediate- and high-risk (remaining patients), respectively. Also, the HD_{95}/HD_{avg} decreased by 14/5.9 mm for the first ones and by 5.3/1.7 mm for the latter. The predictions generated by the PS model overlap well with the ground truth contours. In particular, the correct parts of seminal vesicles and normal tissue surrounding the prostate gland were included in the predicted CTVs. The PS-generated contours do not follow the visible organ boundaries but adjust to the planning delineations.

4 | DISCUSSION

In this work, we investigated the feasibility of deep learning for the automatic segmentation of the CTV, bladder, and rectum in prostate cancer patients treated at a

TABLE 4 Outcomes of the FS and PS training compared to the baseline models (BMs).

Model	*N*	Bladder DSC HD_{95} (mm) HD_{avg} (mm)	Rectum DSC HD_{95} (mm) HD_{avg} (mm)	*N*	CTV int&high DSC HD_{95} (mm) HD_{avg} (mm)	*N*	CTV low&pp DSC HD_{95} (mm) HD_{avg} (mm)
BM	114	0.91(0.07) 6.0(5.1) 1.8(1.1)	0.87(0.04)$^{(ns)}$ 5.2(2.8)$^{(ns)}$ 1.5(0.5)$^{(ns)}$	105	0.73(0.07) 9.6(2.8) 3.3(0.8)	10	0.2(0.05)$^{(ns)}$ 17(4) 7.2(1.8)$^{(ns)}$
FS		0.92(0.04) 3.8(1.8) 1.4(0.4)	0.87(0.04)$^{(ns)}$ 5.0(2.7)$^{(ns)}$ 1.4(0.5)$^{(ns)}$		0.78(0.07) 8.6(3.2) 2.9(1.1)		0.18(0.06)$^{(ns)}$ 12(3) 6.6(1.6)$^{(ns)}$
PS		**0.93(0.06)** **3.5(2.6)** **1.2(0.7)**	**0.90(0.03)** **3.7(2.1)** **1.1(0.4)**		**0.88(0.05)** **4.3(1.5)** **1.7(0.6)**		**0.72(0.04)** **3.2(0.4)** **1.3(0.1)**

Note: DSC, average and 95th percentile Hausdorff distance (HD_{avg}, HD_{95}), with (standard deviation of the mean). The evaluation has been restricted to fraction images of the nine test patients. Results of the best performing models in bold. The non-statistically significant differences are marked with $^{(ns)}$. CTV, clinical target volume; DSC, Dice similarity coefficient; FS, facility-specific; PS, patient-specific.

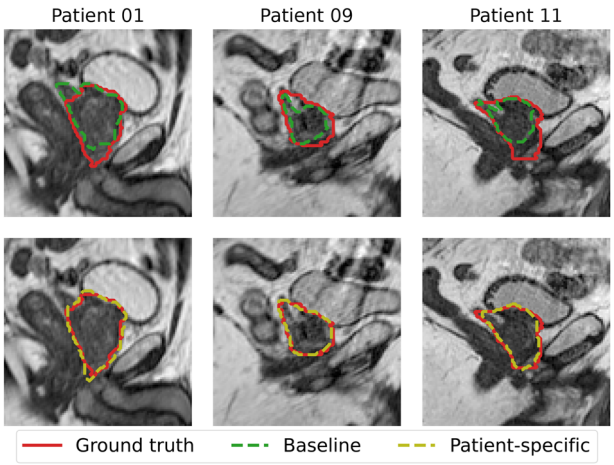

FIGURE 7 Image slices showing the comparison between clinical target volume (CTV) segmentation performed by the (top) baseline and (bottom) patient-specific models.

TABLE 5 Overview of the performance of automatic OAR delineation techniques on MR images.

Study	Method	Bladder DSC HD_{95} (mm)	Rectum DSC HD_{95} (mm)
Elguindi et al.[36]	DeepLabV3+	0.93(0.04)	0.82(0.05)
		–	–
Savenije et al.[37]	DeepMedic	0.96(0.02)	0.88(0.05)
		2.5(1.1)	7.4(4.4)
Sanders et al.[38]	DenseNet	0.96(0.03)	0.91(0.05)
		3.49(6.9)	9.16(6.9)
Huang et al.[39]	U-Net variation	0.90(0.09)	0.78(0.07)
		8.7(9.4)	11.8(8)
This study	3D U-Net	0.93(0.03)	0.88(0.03)
	(Baseline)	3.6(3.0)	3.6(1.5)

Note: A brief description of the method is reported together with DSC and HD_{95} metrics. DSC, Dice similarity coefficient; OAR, organs at risk; MR, magnetic resonance.

0.35 T MR-Linac. Data from two independent facilities were used to test for generalizability of trained models. In addition, contours propagated by the TPS were compared to the network predictions and evaluated regarding their clinical usability during treatment adaptation. Furthermore, the data of the fractionated adaptive treatment course were leveraged, first, to examine differences between planning and fraction contour prediction accuracy and, second, to generate facility- and patient-specific models for the automatic delineation of fraction images by fine-tuning the network parameters on the planning data.

The analysis of the BM yielded no considerable differences between OAR segmentation on planning images from two independent facilities. The mean DSC values for the bladder and the rectum were around 0.93 and 0.88, respectively, while the HD_{95} and HD_{avg} were below 3.7 and 1.3 mm, regardless of the OAR. This suggests that models trained in one of the institutes can be directly used in the other without the necessity of additional model fine-tuning.

This, however, does not apply to the CTV. All three employed metrics indicate more severe errors, that is, drop in DSC by 0.08 and an increase of HD_{95}/HD_{avg} by 3.6 mm/1.3 mm for intermediate- and high-risk cases and more pronounced miss-classifications for low-risk and post-prostatectomy patients when applying the BM to C2 planning images. This potentially rules out model generalizability for CTV delineation and is potentially related to more pronounced differences in contouring style between different facilities for the CTV.

Table 5 presents the outcomes of several recent studies on neural networks for pelvic region auto-segmentation in MRI. The performance of the BM is comparable to those presented in the recent literature. One should bear in mind, however, that the data collected in Table 5 are given as reported by the authors using different training and testing sets. Therefore, they should be interpreted as an estimate of what can be achieved for OAR segmentation on MR images and not as a direct comparison.

The analysis of the BM predictions on planning and fraction OAR contours showed differences in the average DSC between the subsets below 0.03, yet both HD_{95} and HD_{avg} were higher for fraction contours by up to 2.6 and 0.5 mm. The difference could be caused by the limited time that can be dedicated to correct the propagated structures and the fact that mainly the region close to the PTV, that is, the high dose region, is subjected to additional contour adjustments.

According to our institutional protocol, patients were instructed to show up consistently with at least half-full bladder. All patients followed the recommendations for the planning image acquisition, but not always for fractions. This was frequently observed in patients 05 and 14 and resulted in a considerable DSC spread of approximately 0.35. The same was observed in several fractions of patients 01, 08, 12, and 17, represented by the lowest points on the plot (see Figure 4). The bladder volume of patient 03 was about three times larger than average. Both, empty and exceptionally big bladders, were underrepresented in the training set.

Larger variations in rectum DSC, as visible in Figure 4, were caused mostly by the challenges in capturing the sigmoid-rectum transition. The network has tended to segment several additional slices of the large intestine compared to the ground truth segmentation. This issue has been improved upon after PS training, when the precise rectum end for a given patient has been adapted from the planning contours. The source of the problem lies in the hardly visible colon–rectum boundary and the fact that this is a low-dose region,

meaning, that the physician's attention is shifted rather to areas of greater importance, which potentially leads to discrepancies in ground truth contours. Training a network with inconsistent segmentation might lead to an average segmentation style, which will naturally lower performance on the test set.

The physician evaluation clearly showed the advantage of the network predicted structures over the TPS-propagated ones. In contrast to the propagated contours, the vast majority of the predicted structures, 83% of the bladder and 96% of the rectum contours, could be used either directly or after small corrections, thus potentially shortening the time required for re-contouring in the adaptive MRgRT workflow. In order to minimize the impact of personal bias on the results, the physician who performed this analysis was not informed about the details of the study. In the quantitative assessment, it could have been expected, that the TPS-propagated contours would show equally good or even higher DSC and HD due to the way they were generated. Under time pressure, when the patient is lying on the couch, physicians mostly correct pronounced errors of the propagated structures with the main focus on the high-dose region. Slices that are not ideally contoured, but are of quality sufficient for plan adaptation or located in a low-dose region, might be left unchanged. This gives the propagated structures a considerable advantage over the ground truth segmentation in terms of geometric metrics.

The biggest challenge of the CTV segmentation was classifying the correct amount of seminal vesicles and normal tissue surrounding the prostate gland. The network was trained on data, where 90% of the cases constituted intermediate- and high-risk cases and therefore assumed the CTV to include parts of the seminal vesicles. An alternative training that excluded the low-risk cases did not improve segmentation results, therefore all cases including all risk categories, were kept in the baseline training set. Yet, the low-risk and post-prostatectomy cases were taken into account separately while testing. It can be also noticed on the upper part of Figure 7 that the BM assumed no additional margin around the prostate, which might, however, sometimes be required in CTV definition.

For the OARs, the FS and PS training improved the average DSC accuracy only slightly, yet brought a decrease in HD_{95} and HD_{avg}. The PS training was beneficial mostly for determining the correct colon-rectum boundary (patients 06, 10, 16) and correcting for misclassification of larger areas of normal tissue (bladder, patient 01). However, if the rectum filling was remarkably different on the planning day than on the day of irradiation (e.g., patients 01 and 12), the PS training reduced accuracy. This behavior can be observed in Figure 6. Both types of transfer learning are intrinsically sensitive to the quality of the planning segmentation and might be affected by large changes in organ shape with respect to the planning image. Although advantageous for patients with unusual anatomies, it could propagate errors in initial contouring and over-favor the planning shape. Therefore, we believe that for the OARs a BM trained on more examples of unusual anatomies, for example, various bladder fillings, would be the better choice than the PS training.

A clear benefit was observed for the CTV undergoing a PS training. The models learned the geometry of the planning CTV and successfully applied it to delineate fraction contours. Especially, they learned to include the correct amount of seminal vesicles and normal tissue as can be seen in Figure 7. For the nine test patients, the DSC improved from 0.68(0.16) to 0.86(0.06), the HD_{95} from 10(4) mm to 4.2(1.5) mm and the HD_{avg} from 3.7(1.4) mm to 1.6(0.6) mm, which corresponds to approximately one and three pixels, respectively. It should be noted that an average CTV volume is much smaller than the size of a (half) full bladder and therefore, a high score on DSC is harder to achieve here. In the context of MRgRT, where expert delineations can be expected on a planning image, PS transfer learning may lead to time gains during online adaptive fractions.

In order to achieve the desired accuracy, the PS networks were fine-tuned over 500 epochs, which took about 2 h. If needed, this could be shortened to 300 epochs with only a small loss in performance, reducing the training time by roughly 50 min. Since the first fraction takes place several days after the planning MR acquisition, the proposed PS training is feasible in a typical clinical workflow. The time required to predict a single contour with a trained model, approximately 1 s, is negligible compared to the duration of the treatment adaptation procedure.

The study presented here has its limitations. Due to the lack of a complete model reliability, physician review remains unavoidable. However, as suggested in[14,15] the time required to correct network-generated structures might be significantly shorter than contouring from scratch. One can also speculate that in our case the correction of network predictions is shorter than adjusting the TPS-propagated contours, given the better grading observed in our study (see Figure 5). The quality of bladder autosegmentation could be improved by including cases with variable bladder filling in the training set, since not all patients follow the clinical protocol that recommends filling the bladder before each fraction. For the low-risk CTV, one could consider collecting a larger database and training a dedicated BM as the basis for PS transfer learning.

Another study limitation concerns the manual localization of the PTV. The augmentation pipeline takes input data of size $220 \times 220 \times 220$ pixels and crops it further to $192 \times 192 \times 192$. Despite the final size of 192^3 voxels, which corresponds a relatively large volume of 28.8^3 cm^3, an approximate isocenter position

might be determined by an additional network for full automization.

This study focused on the CTV, bladder, and rectum segmentation as crucial structures with regard to prostate cancer RT. Delineations of more OARs might be required in the future, especially in other anatomical sites, where a significant segmentation burden is expected (e.g., abdomen). However, there are no conceptual limitations to expand the network toward the prediction of further structures.

Currently, the biggest limitation is the quality of ground truth segmentation. The contours were created by several physicians with the assumption to be sufficiently accurate for treatment planning. However, while small inconsistencies, especially outside of the high-dose region, do not affect the dose calculation, they can decrease DSC considerably. As previously mentioned, the random nature of these inconsistencies did not have a strong impact on network learning, as the differences naturally average out, and the trained models approach the visible boundaries of the organs. However, this negatively impacts validation and testing. Using consistently segmented datasets would help to solve this problem.

5 | CONCLUSIONS

In this work, 3D U-Nets for CTV, rectum, and bladder segmentation were successfully trained for prostate cancer patients treated at two 0.35 T MR-Linacs at two independent facilities. The quality of the predicted contours was confirmed by the high DSC and low HD scores. In addition, the investigated network delineations of OARs were preferred over the currently used structures that are suggested by the clinical system. It was shown that the accuracy of the OAR segmentation was transferable to a second cohort from an independent institute. Moreover, for the first time the usefulness of PS training to improve CTV autosegmentation was demonstrated, which could be an effective method for exploiting the prior knowledge available due to the fractionated type of data seen in adaptive MRgRT.

ACKNOWLEDGMENTS

The authors wish to thank Vanessa Filipa Da Silva Mendes for her support with the ViewRay treatment planning system and the data export. We would like to acknowledge the support from Prof. Sibylle Ziegler, Claudio Votta, Gabriele Turco as well as Dr. Seyed-Ahmad Ahmadi for his introduction to the MONAI framework. Special thanks to Martin Rädler for comments and suggestions throughout the study, help in designing figures and support in writing the paper. This work was funded by the Wilhelm Sander-Stiftung (2019.162.1).

CONFLICTS OF INTEREST

The Department of Radiation Oncology of the University Hospital of LMU Munich has a research agreement with ViewRay. ViewRay did not fund this study and was not involved and had no influence on the study design, the collection or analysis of data, or on the writing of the manuscript.

DATA AVAILABILITY STATEMENT

No data to share.

REFERENCES

1. Winkel D, Bol GH, Kroon PS, et al. Adaptive radiotherapy: the Elekta unity MR-linac concept. *Clin Transl Radiat Oncol*. 2019;18:54-59.
2. Henke L, Contreras J, Green O, et al. Magnetic resonance image-guided radiotherapy (MRIgRT): a 4.5-year clinical experience. *Clin Oncol*. 2018;30:720-727.
3. Da Silva Mendes V, Nierer L, Li M, et al. Dosimetric comparison of MR-linac-based IMRT and conventional VMAT treatment plans for prostate cancer. *Radiat Oncol*. 2021;16:1-12.
4. Corradini S, Alongi F, Andratschke N, et al. Mr-guidance in clinical reality: current treatment challenges and future perspectives. *Radiat Oncol*. 2019;14:1-12.
5. Finazzi T, Palacios MA, Spoelstra FO, et al. Role of on-table plan adaptation in MR-guided ablative radiation therapy for central lung tumors. *Int J Radiat Oncol Biol Phys*. 2019;104:933-941.
6. Bruynzeel AM, Tetar SU, Oei SS, et al. A prospective single-arm phase 2 study of stereotactic magnetic resonance guided adaptive radiation therapy for prostate cancer: early toxicity results. *Int J Radiat Oncol Biol Phys*. 2019;105:1086-1094.
7. Güngör G, Serbez I, Temur B, et al. Time analysis of online adaptive magnetic resonance-guided radiation therapy workflow according to anatomical sites. *Pract Radiat Oncol*. 2021;11:e11-e21.
8. Sahin B, Mustafayev TZ, Gungor G, et al. First 500 fractions delivered with a magnetic resonance-guided radiotherapy system: initial experience. *Cureus*. 2019;11(12):e6457.
9. Lamb J, Cao M, Kishan A, et al. Online adaptive radiation therapy: implementation of a new process of care. *Cureus*. 2017;9(8):e1618.
10. Rogowski P, von Bestenbostel R, Walter F, et al. Feasibility and early clinical experience of online adaptive MR-guided radiotherapy of liver tumors. *Cancers*. 2021;13:1523.
11. Hadi I, Eze C, Schönecker S, et al. MR-guided SBRT boost for patients with locally advanced or recurrent gynecological cancers ineligible for brachytherapy: feasibility and early clinical experience. *Radiat Oncol*. 2022;17:1-9.
12. Klüter S. Technical design and concept of a 0.35 T MR-Linac. *Clin Transl Radiat Oncol*. 2019;18:98-101.
13. Cusumano D, Boldrini L, Dhont J, et al. Artificial intelligence in magnetic resonance guided radiotherapy: medical and physical considerations on state of art and future perspectives. *Phys Med*. 2021;85:175-191.
14. Cha E, Elguindi S, Onochie I, et al. Clinical implementation of deep learning contour autosegmentation for prostate radiotherapy. *Radiother Oncol*. 2021;159:1-7.
15. Zabel WJ, Conway JL, Gladwish A, et al. Clinical evaluation of deep learning and atlas-based auto-contouring of bladder and rectum for prostate radiation therapy. *Pract Radiat Oncol*. 2021;11:e80-e89.
16. Fiorino C, Reni M, Bolognesi A, Cattaneo GM, Calandrino R. Intra- and inter-observer variability in contouring prostate and seminal vesicles: implications for conformal treatment planning. *Radiother Oncol*. 1998;47:285-292.

17. Liang F, Qian P, Su KH, 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.
18. Fu Y, Mazur TR, Wu X, et al. A novel MRI segmentation method using CNN-based correction network for MRI-guided adaptive radiotherapy. *Med Phys*. 2018;45:5129-5137.
19. Eppenhof KA, Maspero M, Savenije M, et al. Fast contour propagation for MR-guided prostate radiotherapy using convolutional neural networks. *Med Phys*. 2020;47:1238-1248.
20. Cai L, Gao J, Zhao D. A review of the application of deep learning in medical image classification and segmentation. *Ann Transl Med*. 2020;8(11):713.
21. Friedrich F, Hörner-Rieber J, Renkamp CK, et al. Stability of conventional and machine learning-based tumor auto-segmentation techniques using undersampled dynamic radial bSSFP acquisitions on a 0.35 T hybrid MR-linac system. *Med Phys*. 2021;48:587-596.
22. Çiçek Ö, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O. 3D U-Net: learning dense volumetric segmentation from sparse annotation. *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer; 2016:424-432.
23. Balagopal A, Morgan H, Dohopolski M, et al. PSA-Net: deep learning-based physician style-aware segmentation network for postoperative prostate cancer clinical target volumes. *Artif Intell Med*. 2021;121:102-195.
24. Chun J, Park JC, Olberg S, et al. Intentional deep overfit learning (IDOL): a novel deep learning strategy for adaptive radiation therapy. *Med Phys*. 2022;49:488-496.
25. Sharp GC, Li R, Wolfgang J, et al. Plastimatch: an open source software suite for radiotherapy image processing. *Proceedings of the XVI'th International Conference on the use of Computers in Radiotherapy (ICCR)*, Amsterdam, Netherlands. 2010.
26. Ma N, Li W, Brown R, et al. Project MONAI. *Zenodo CERN*. 2021.
27. Kerfoot E, Clough J, Oksuz I, Lee J, King AP, Schnabel JA. Left-ventricle quantification using residual U-Net. *International Workshop on Statistical Atlases and Computational Models of the Heart*. Springer; 2018:371-380.
28. Ulyanov D, Vedaldi A, Lempitsky V. Instance normalization: the missing ingredient for fast stylization. 2016. arXiv:1607.08022.
29. Ding B, Qian H, Zhou J. Activation functions and their characteristics in deep neural networks. Chinese control and decision conference (CCDC). IEEE; 2018:1836-1841.
30. Bridle JS. Training stochastic model recognition algorithms as networks can lead to maximum mutual information estimation of parameters. *Adv Neural Inf Process Syst*. 1990;2:211-217.
31. Milletari F, Navab N, Ahmadi SA. V-net: Fully convolutional neural networks for volumetric medical image segmentation. Fourth international conference on 3D vision (3DV). IEEE; 2016: 565-571.
32. Kingma DP, Ba J. Adam: a method for stochastic optimization. 2014. *arXiv:1412.6980*.
33. Ruskó L, Capala ME, Czipczer V, et al.Deep-learning-based segmentation of organs-at-risk in the head for MR-assisted radiation therapy planning. *BIOIMAGING*. 2021;2:31-43.
34. Paszke A, Gross S, Massa F, et al. Pytorch: an imperative style, high-performance deep learning library. *Adv Neural Inf Process Syst*. 2019;32:8026-8037.
35. Pérez-García F, Sparks R, Ourselin S. TorchIO: a Python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning. *Comput Methods Programs Biomed*. 2021;208:106236.
36. Elguindi S, Zelefsky MJ, Jiang J, 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;12:80-86.
37. Savenije MH, Maspero M, Sikkes GG, et al. Clinical implementation of MRI-based organs-at-risk auto-segmentation with convolutional networks for prostate radiotherapy. *Radiat Oncol*. 2020;15:1-12.
38. Sanders JW, Lewis GD, Thames HD, et al. Machine segmentation of pelvic anatomy in MRI-assisted radiosurgery (MARS) for prostate cancer brachytherapy. *Int J Radiat Oncol Biol Phys*. 2020;108:1292-1303.
39. Huang S, Cheng Z, Lai L, et al. Integrating multiple MRI sequences for pelvic organs segmentation via the attention mechanism. *Med Phys*. 2021;48:7930-7945.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Kawula M, Hadi I, Nierer L, et al. Patient-specific transfer learning for auto-segmentation in adaptive 0.35 T MRgRT of prostate cancer: a bi-centric evaluation. *Med Phys*. 2023;50:1573–1585.
https://doi.org/10.1002/mp.16056

A Supplementary information: Data augmentation and hyper-parameter search

Table A.1 collects functions and hyperparameters used for data augmentation of the baseline models, the limits in which they were tested as well as the final values.

Table A.1: Functions used for data augmentation and hyper-parameters to be optimized. The investigated ranges and the values selected for final model training are given.

	Function	Parameter	Tested range	Final value
	Probability	p_{agm}	0.7-1	0.9
	Learning rate	lr	$10^{-4} - 10^{-2}$	10^{-3}
Spatial	Rotation	α_{max}	5°- 20°	20°
	Translation	Δ_{max}	-	15 mm
	Deformation	n_{cp}	10-20	10
		d	22.5- 45 mm	37.5 mm
	Zooming	z_{min}, z_{max}	-	0.9, 1.1
MR	Motion	m_α	5°- 15°	10°
		m_Δ	20- 50 mm	45 mm
	Bias field	deg	1-3	1
		c_{mag}	0-1	0.4
	Noise	σ	0-1	0.25
		μ	-	0

The first subgroup of hyperpatameters contained two values to be determined: the augmentation probability p_{agm} and the learning rate lr. The functions and the starting parameter values for the spatial transformations have been taken from our previous work[40]. These include parameters for rotation (the maximum rotation angle α_{max}), translation (the maximum displacement Δ_{max}), zooming (the zoom and cropping ratio z_{min}, z_{max}) and the B-Spline deformation (number of control points n_{cp} and the maximum displacement d). To simulate MR-specific artifacts, random motion, bias field, and random noise have been introduced. The random motion artifact aims to mimic the influence of patient movement during image acquisition (defined by the maximum possible translation m_Δ and the rotation angle m_α). The bias field caused by signal attenuation within patients body is modeled via polynomials of a given degree (`deg`) and with specified magnitude (c_{mag}) of its coefficients. Detailed information on the nature as well as mathematical description of these phenomena can be found in Surde *et al.*[41] and Shaw *et al.*[42]. The random noise was sampled from a normal distribution with mean μ and standard deviation σ.
The parameter fine-tuning has been performed as follows. The initial parameter values were either chosen to introduce only subtle alterations to the image-binary mask pairs or, in the case of spatial transformations, motivated by our previous work[40]. All values were then gradually changed until the network performance either stopped improving on the validation data or the loss computed on the training set settled on a value that was noticeably higher than the loss on the not-altered validation cases, indicating unrealistic augmentation.
The MR-specific augmentations and the B-spline elastic deformation were imported from the TorchIO package[35], the remaining functions from the MONAI[26] library.

5.3 Paper III

The study described in Paper 3 investigated using segmented planning images to improve auto-segmentation during fractionated MRgRT. The target group was prostate cancer patients irradiated at the MRIdian MR-LINAC. The aim was to develop neural networks for combined image registration and contour propagation. Such networks could be used to register the planning and fraction MRIs and deform planning contours to the anatomy of the day. An advantage of this registration approach is that the planning delineations serve as a starting point for the fraction contours. Hence, the decisions of contouring physicians are taken into account, for example, how much normal tissue and seminal vesicles should be encompassed by the CTV contour or where to define the superior end of the rectum. In this study, the registration method was compared to the personalized models developed in Study 2, i.e., patient-specific networks generated by fine-tuning population models with the planning image of a given patient. This is the first study to investigate neural networks for combined image registration and contour propagation for MRgRT at the MRIdian MR-Linac.
The chosen architecture was a U-Net trained for single-class segmentation of the bladder, rectum, and prostate CTV. The U-Net was designed to predict a dense displacement field between a pair of input images and propagate contours between them. An important part of this study was the loss function design, which involved up to three terms. The first term quantified the similarity between the predicted and simulated ground truth displacement fields. This term could also include regularization to restrict the predictions to anatomically plausible deformations. The second loss term measured the similarity between the target and the deformed image. The third loss term quantified the similarity between the target and the deformed contours. These three terms were investigated in different combinations to find an optimal loss function. Another investigated aspect was network training on gradually increasing resolutions to mimic a classical multi-stage deformable registration. Finally, patient-specific models were trained as described in Paper 2.
The study showed that registration networks used in the work can model only small deformations regardless of the choice of the loss terms. Training on different resolutions also did not bring the expected improvements. Therefore, registration networks were found to be unsuitable for organs that can undergo substantial changes, such as the bladder and rectum. Patient-specific models were found to be a superior method for these two OARs. However, registration models were found useful in segmenting prostate CTVs, as delineation of the latter should not change substantially throughout the treatment to ensure proper dose coverage of the tumor. The presumed reason for predicting only limited deformations was the large dimensionality of the expected network output in relation to the training data set size. In fact, the degrees of freedom for a 3D deformation field scale cubically with the image size.

Physics and Imaging in Radiation Oncology 28 (2023) 100498

Contents lists available at ScienceDirect

Physics and Imaging in Radiation Oncology

journal homepage: www.sciencedirect.com/journal/physics-and-imaging-in-radiation-oncology

Original Research Article

Prior knowledge based deep learning auto-segmentation in magnetic resonance imaging-guided radiotherapy of prostate cancer

Maria Kawula [a], Marica Vagni [b], Davide Cusumano [b,c], Luca Boldrini [b], Lorenzo Placidi [b], Stefanie Corradini [a], Claus Belka [a,d,e], Guillaume Landry [a], Christopher Kurz [a,*]

[a] Department of Radiation Oncology, LMU University Hospital, LMU Munich, Munich, Germany
[b] Fondazione Policlinico Universitario "Agostino Gemelli" IRCCS, Rome, Italy
[c] Mater Olbia Hospital, Olbia (SS), Italy
[d] German Cancer Consortium (DKTK), Partner Site Munich, A Partnership Between DKFZ and LMU University Hospital Munich, Germany
[e] Bavarian Cancer Research Center (BZKF), Munich, Germany

ARTICLE INFO

Keywords:
Auto-segmentation
Patient-specific models
Spatial transformer layer
Deep learning
MRgRT
MR-Linac
Prostate cancer

ABSTRACT

Background and purpose: Automation is desirable for organ segmentation in radiotherapy. This study compared deep learning methods for auto-segmentation of organs-at-risk (OARs) and clinical target volume (CTV) in prostate cancer patients undergoing fractionated magnetic resonance (MR)-guided adaptive radiation therapy. Models predicting dense displacement fields (DDFMs) between planning and fraction images were compared to patient-specific (PSM) and baseline (BM) segmentation models.

Materials and methods: A dataset of 92 patients with planning and fraction MR images (MRIs) from two institutions were used. DDFMs were trained to predict dense displacement fields (DDFs) between the planning and fraction images, which were subsequently used to propagate the planning contours of the bladder, rectum, and CTV to the daily MRI. The training was performed either with true planning-fraction image pairs or with planning images and their counterparts deformed by known DDFs. The BMs were trained on 53 planning images, while to generate PSMs, the BMs were fine-tuned using the planning image of a given single patient. The evaluation included Dice similarity coefficient (DSC), the average (HD_{avg}) and the 95th percentile (HD_{95}) Hausdorff distance (HD).

Results: The DDFMs with DSCs for bladder/rectum of 0.76/0.76 performed worse than PSMs (0.91/0.90) and BMs (0.89/0.88). The same trend was observed for HDs. For CTV, DDFM and PSM performed similarly yielding DSCs of 0.87 and 0.84, respectively.

Conclusions: DDFMs were found suitable for CTV delineation after rigid alignment. However, for OARs they were outperformed by PSMs, as they predicted only limited deformations even in the presence of substantial anatomical changes.

1. Introduction

Integrated magnetic resonance (MR) linear accelerators (MR-Linacs) facilitate daily MR image (MRI) acquisition and dose adaptation [1–3]. This enables the reduction of safety margins and hypofractionation [4–7]. However, online adaptive MR-guided radiation therapy (MRgRT) has longer workflows than conventional linacs. The most time-consuming step besides irradiation is the generation of updated organ-at-risk (OAR) and target volume contours on the fraction MRIs [8–10].

In clinical workflows [1], the planning and fraction MRIs are first rigidly aligned and then registered with deformable image registration (DIR) in the treatment planning system (TPS). The clinical target volume (CTV) is rigidly copied to the daily MRI, while the OARs are propagated with a displacement vector field estimated in the TPS. Since the quality of propagated contours is suboptimal for dose evaluation and optimization, manual corrections are required. Automatic segmentation could shorten the adaptation time as suggested by Cha et al. [11] or Zabel et al. [12], and reduce inter- and intra-physician variability [13,14].

Despite several studies on MRI segmentation [15,16], the utilization of planning contours to enhance auto-segmentation on fraction images is

* Corresponding author.
E-mail address: Christopher.Kurz@med.uni-muenchen.d (C. Kurz).

https://doi.org/10.1016/j.phro.2023.100498
Received 30 May 2023; Received in revised form 3 October 2023; Accepted 4 October 2023
Available online 10 October 2023

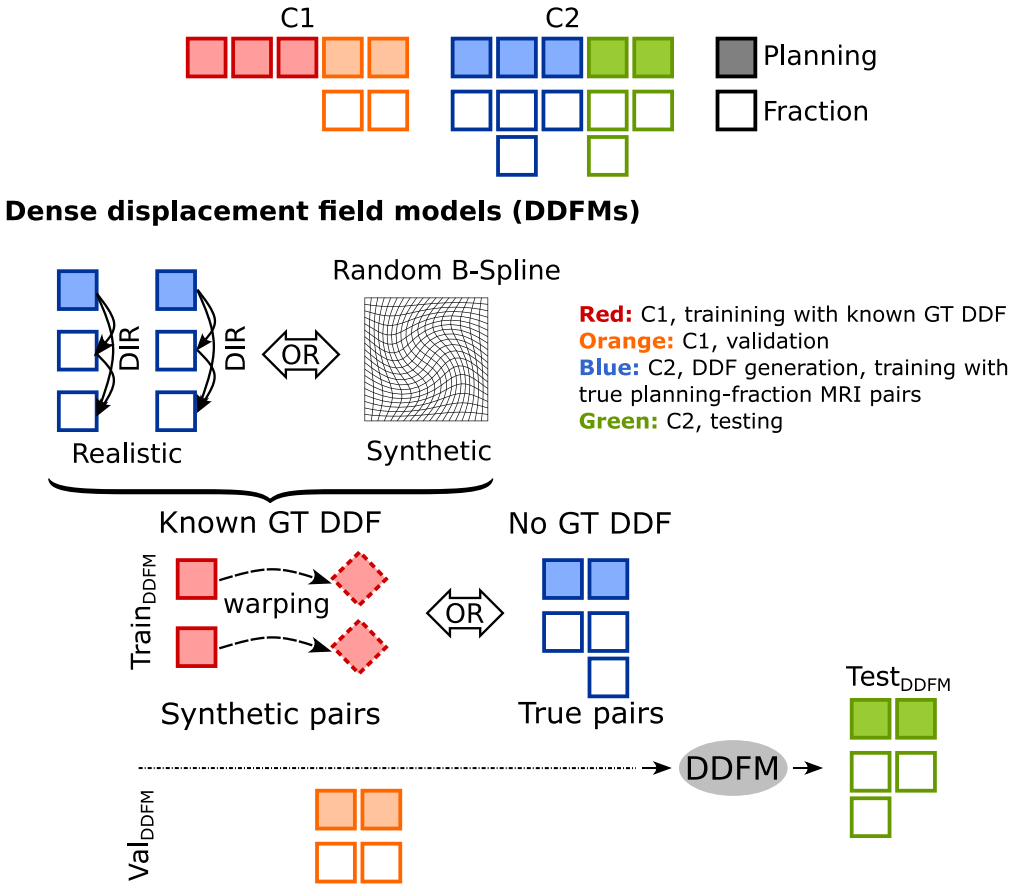

Fig. 1. DDFM training and evaluation scheme. Cohorts 1 (C1) and 2 (C2) were utilized in the study. In the first approach of the DDFM training with known ground truth (GT) DDFs, planning-fraction image pairs were generated by warping the C1 MRIs using either the "realistic" or the "synthetic" dense displacement fields (DDFs). The former were derived from deformable image registration (DIR) of the C2 images. The second approach had no ground truth (GT) DDF but was trained on the true planning-fraction image pairs from C2. Both variants were validated and tested on the remaining patients from C1 and C2, respectively.

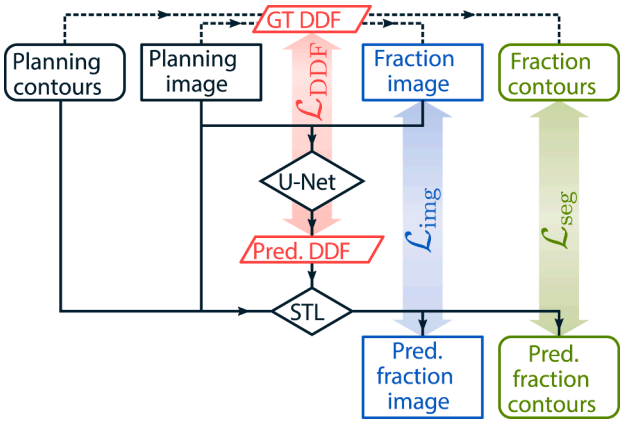

Fig. 2. Training scheme and possible loss terms $\mathscr{L}$. The dense displacement field (DDF) predicted by the U-Net is passed to the spatial transformer layer (STL) along with the planning data to generate predictions for the fraction image and contours. The usage of the data variant with the synthetic planning-fraction pair and the known ground truth (GT) DDF has been indicated with dashed lines.

still limited. Fransson et al. [17] showed single-patient 2D models trained with the first fraction image to predict contours on consecutive treatment days. Li et al. [18] used a similar 2D training strategy, but suggested refining the model after each irradiation. Kawula et al. [19] proposed using the planning image to fine-tune generic models, which adjusted the model to the patient of interest without losing generality. Eppenhof et al. [20] presented another strategy for CTV segmentation based on the prediction of a dense displacement field (DDF) between the planning and fraction images. Our work is the first study investigating this method for the segmentation of OARs in prostate cancer patients.

This study thus aimed to train models predicting DDFs (DDFMs) between the planning and fraction images, enabling propagation of planning contours to the daily anatomy. Moreover, patient-specific models (PSMs) [19] were trained by fine-tuning generic models (baseline models, BMs) using segmented planning MRIs. All methods were compared to DIR with contour propagation and rigid contour copying.

2. Materials and Methods

2.1. Dataset

Datasets from two facilities were used: 73 patients from the Gemelli University Hospital in Rome formed the first cohort (C1). For ten patients the planning and one fraction MRI were available, while for the remaining 63, only the planning image was included. The second cohort (C2) had 19 patients from the LMU University Hospital. Patients in C2 had one planning and 5–33 fraction images (240 fraction MRIs in total). Informed written consent was obtained from all patients and the study was carried out carried out in accordance with relevant ethics guidelines and regulations (LMU: ethics project number 20–291, Gemelli: EC authorization number 3460).

All delineations used as ground truth were created during the clinical workflow. Rectum, bladder, and CTV contours on planning MRIs

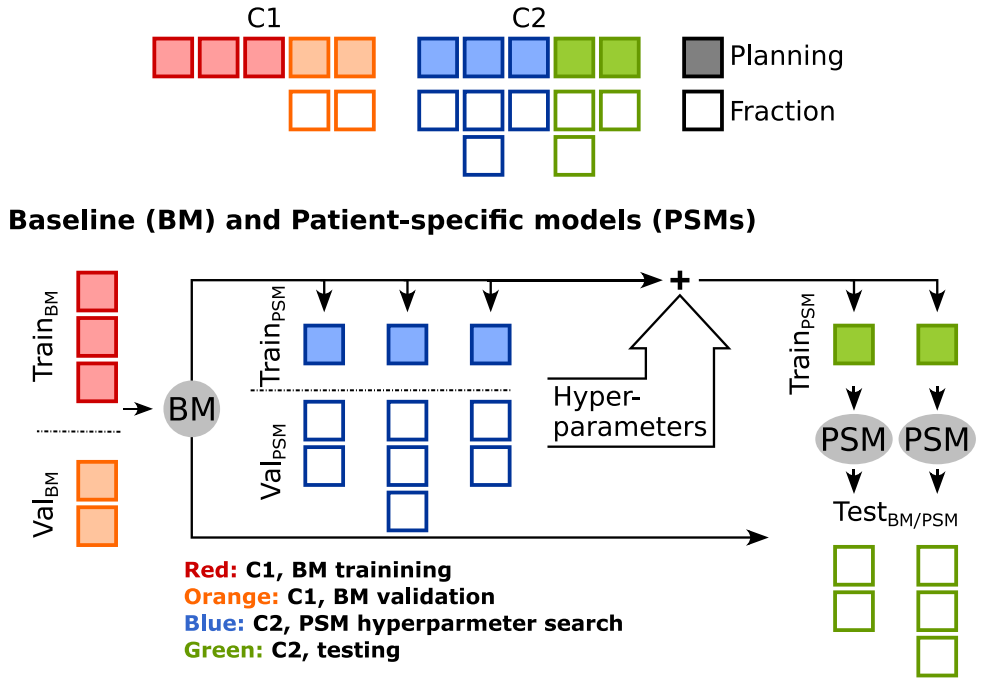

Fig. 3. Baseline (BM) and patient-specific model (PSM) training and evaluation scheme. Cohorts 1 (C1) and 2 (C2) utilized in the study. The BM was trained and validated on MRIs from C1. For the PSMs, a hyperparameter search was conducted using MRIs from ten C2 patients. The final PSMs were generated through fine-tuning the BM with planning MRIs of the 9 C2 test patients with fixed hyperparameters. Both methods shared the same test set of fraction data of these 9 C2 patients.

(planning contours) were manually drawn by physicians. At every fraction a new MRI was acquired for potential plan adaptation. A radiation therapy technologist (RTT) performed manual rigid registration of the images, prioritizing the alignment of the prostate and seminal vesicles, followed by the adjustment of bones and outline. Subsequently the planning and fraction images were registered with DIR in the TPS. The resulting DDF was used to propagate the planning OAR contours to the daily MRI. Usually, a physician manually corrected them due to their insufficient quality, focusing mostly on the high-dose region, i.e. the proximity of the planning target volume (PTV). The resulting delineations will be referred to as fraction contours. In our previous work [19], no differences in OAR annotation styles were found between the two cohorts. The C2 CTV contours included slightly more normal tissue than the C1 CTV structures, however, differences were subtle. The percentages of low, intermediate and high-risk cases, which determine the CTV contour size, were alike in both cohorts.

All MRIs were acquired at 0.35 T MR-Linacs (MRIdian, ViewRay Inc, Cleveland, Ohio) with a balanced steady-state free precession (bSSFP) sequence resulting in T2*/T1 contrast [1]. The in-plane spacing was 1.5 mm × 1.5 mm and the axial slice thickness was 1.5 mm (~90% of the MRIs in both cohorts) or 3 mm (remaining ~10%). The latter were resampled to 1.5 mm slice thickness using the open-source tool Plastimatch [21] with either linear (for images) or nearest neighbor (for contours) interpolation. To mimic the manual rigid alignment done clinically by RTTs, all images were cropped to 192 × 192 × 192 voxels around the CTV centroid.

2.2. Architectures

For each anatomical structure, a separate network was trained. The MONAI [22] implementation of a 3D U-Net architecture [23] was used for segmentation with all investigated models. It comprised five resolution levels, i.e., there were four down- and four up-sampling operations. Each level had two convolutions with 3 × 3 × 3 kernels, followed by instance normalization and PReLU activation. The down- and up-sampling employed double-stride and up-convolution, respectively.

For BMs and PSMs, the U-Net was predicting which voxels of the input image belong to the foreground. For DDFMs it was predicting the DDF of size 192 × 192 × 192 × 3 between the planning and the fraction image. The output of the DDFM U-Net was given to a spatial transformer layer (STL) [24] that served as a grid interpolator and had no learnable parameters.

2.3. Training of DDFMs

Two types of input data were considered. The first type were the C1 planning images and contours warped by known training DDFs generating synthetic fraction images and contours. The second type were the true planning-fraction image pairs of C2 patients (125 pairs for training and 115 pairs for testing), without a ground truth DDF. In both scenarios ten C1 patients, for whom fraction data were available, were used as an independent validation set. Figure 1 illustrates the data split and the DDFM training scheme.

Training DDFs were generated in two ways. For the *synthetic* DDFs, TorchIO's [25] RandomElasticDeformation function was employed on-the-fly during training, separately for each mini-batch. Its parameters were set to yield clearly visible transformations (number_of_control_points = 10, max_displacement = 35 mm) due to the expected substantial volume changes in the OARs. The so-called *realistic* DDFs were extracted from the DIR between all image pairs within a patient's dataset (planning and fraction MRIs). This was done for ten C2 patients using Plastimatch with multistage B-spline registration [26] and resulted in a total of 860 different realistic DDFs. Information on the B-spline deformation are in the supplementary material (section A).

DDFM training was carried out in two ways. The first followed the progressive training scheme described by Eppenhof et al. [27]. The network was trained at different resolutions, starting with a coarse alignment of $n-1 = 4$ times sub-sampled MRIs (corresponding to $n = 5$ U-Net levels) and gradually adding higher-resolution data as the training progressed. In the second variant, the network was trained at once.

During training, either true or the synthetic input data pairs were used without data augmentation. For the approach with the synthetic image pairs, either the synthetic or the realistic DDFs were used for model training. The training variants are shown in Figure 1.

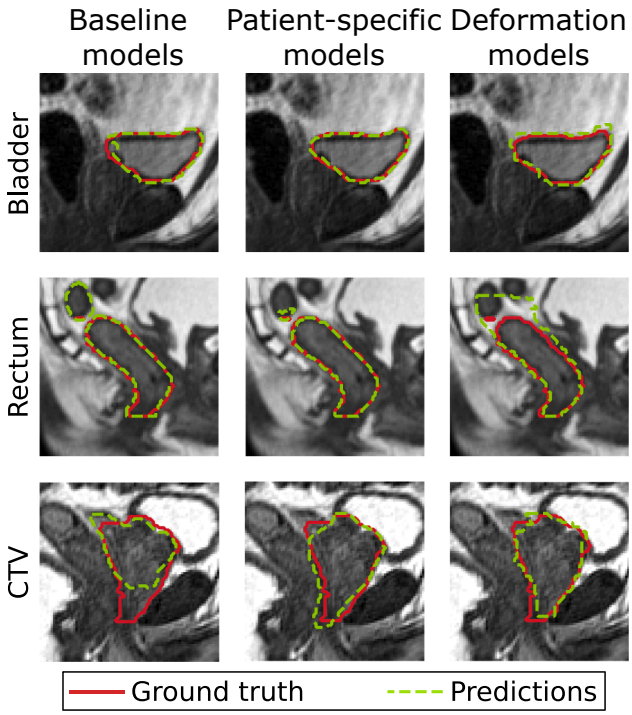

Fig. 4. Sagittal view of exemplary test patients showing the comparison between segmentations performed by (left) the baseline models, (middle) patient-specific models, and (right) dense displacement field (deformation) models for (top) the bladder, (middle) rectum, and (bottom) clinical target volume (CTV). Rectum contours are shown only within the evaluation volume, i.e. slices including the planning target volume (PTV) and 1.5 cm above and below its upper and lower ends.

The total loss function $\mathscr{L}_{\text{tot}}$ was defined as follows:

$$\mathscr{L}_{\text{tot}} = \lambda_{\text{DDF}}\mathscr{L}_{\text{DDF}} + \lambda_{\text{seg}}\mathscr{L}_{\text{seg}} + \lambda_{\text{img}}\mathscr{L}_{\text{img}}. \quad (1)$$

The DDF term $\mathscr{L}_{\text{DDF}} = \lambda_{\text{DDF,L2}}L_2 + \lambda_{\text{DDF,reg}}L_{\text{reg}}$ measured the similarity between the ground truth and the predicted DDF using the L_2 norm and regularized the predicted DDF with the bending energy L_{reg} [28]. For the true planning-fraction image pairs, there was no ground truth DDF, and therefore no L_2 term. The segmentation loss $\mathscr{L}_{\text{seg}}$ quantified the similarity between the contours with the (multi-scale) Dice Similarity Coefficient (DSC) [29]. The DSC-based loss function was chosen due to the class imbalance with relatively few voxels belonging to the foreground (OARs and CTV), and the majority belonging to the background. The image loss $\mathscr{L}_{\text{img}}$ aimed to optimize the similarity between the registered images using either cross-correlation or L_2 norm. The λ coefficients determined the contribution of each term in the total loss. All training variations are depicted in Figures 1 and 2. Among the investigated DDFMs the best DSC for the ten C1 validation images was obtained while training the entire network at once (no progressive learning) with true planning-fraction image pairs, and the loss function including terms with bending energy loss, multi-scale DSC, and L_2 norm for image similarity. The optimal weighting parameters were found to be $\lambda_{\text{DDF}} = 1$, $\lambda_{\text{DDF,L2}} = 0, \lambda_{\text{DDF,reg}} = 10, \lambda_{\text{seg}} = 100, \lambda_{\text{img}} = 1$, and the number of epochs to be 200 epochs. The training was carried out on Nvidia Quadro RTX 8000 or Nvidia RTX A6000 GPUs (used also later for the BM and PSM training) and took two days. No signs of overfitting were present at that stage, and the DDFM predictions did not improve significantly beyond that point. All results presented later were obtained using the settings listed above.

2.4. Training of baseline and patient-specific models

For the single-label BMs, three 3D U-Net models have been trained on 53 planning images from 53 C1 patients for bladder, rectum, and CTV segmentation. BM networks were trained over 300 epochs with a batch size of two for 20 h. Patient-specific transfer learning applied to the BMs. For each patient, the BM was fine-tuned with its on-the-fly augmented planning image and contours. Hyperparameters were adjusted using ten C2 patients. Figure 3 shows the training scheme. PSMs did not benefit from fine-tuning beyond 500 epochs which took 3 h. There were small improvements in DSC after 300 epochs, where training could have stopped. For the remaining nine C2 cases (the same as for the DDFMs) PSMs were fine-tuned over 500 epochs with fixed hyperparameters for the final testing. The process of hyperparameter optimization was described in detail by Kawula et al. [19].

2.5. Benchmarking and data evaluation

For benchmarking, a classical multistage B-spline registration with Plastimatch and rigid copying of the planning contours to the daily anatomy were performed (see supplementary material section A).

All investigated methods shared the same test set of nine C2 patients having 115 fractions in total. Network predictions were compared to the ground truth via DSC, the 95$^{\text{th}}$ percentile (HD_{95}) and the average (HD_{avg}) Hausdorff distance (HD), calculated using Plastimatch. To prevent bias towards patients with more fractions, we initially calculated the average DSC and HDs for each patient. The final model performances were then reported as the averages over all test patients. Since the calculated metrics (average values of DSC and HDs for each patient) followed a normal distribution, as determined by the Kolmogorov–Smirnov test, we computed the mean values along with their corresponding standard deviations. Evaluation of rectum segmentations considered slices including the PTV and ten additional slices reaching 1.5 cm above and below the upper and lower PTV ends. To assess statistical significance of differences among the methods the paired t-test was conducted for all three metrics, employing a significance level of 0.05.

3. Results

Fig. 4 shows sagittal slices from exemplary test patients with the predictions of BMs, PSMs, and DDFMs versus ground truth

Table 1
Mean and (standard deviation) of Dice similarty coefficient (DSC), 95th percentile (HD_{95}), and average (HD_{avg}) Hausdorff distance for the test set patients. The dense displacement field (DDFM), and patient-specific models (PSM) [19] are compared to the baseline models (BMs), conventional deformable image registration with Plastimatch, and rigid copying of the planning contours to the fraction anatomy. The analysis was performed for the nine C2 test patients (in total 115 fractions). The results of the best-performing models are in bold.

Method	Bladder DSC HD_{95} [mm] HD_{avg} [mm]	Rectum DSC HD_{95} [mm] HD_{avg} [mm]	CTV DSC HD_{95} [mm] HD_{avg} [mm]
Plastimatch	0.79(0.14) 9.6(4.7) 3.6(2.0)	0.78(0.08) 7.4(4.1) 2.3(1.0)	0.83(0.11) 4.5(2.3) 2.3(2.2)
Copying	0.72(0.11) 12(4) 4.8(1.6)	0.70(0.03) 8.8(3.2) 3.1(0.8)	**0.89(0.02)** 2.9(0.9) 1.1(0.4)
BM	0.89(0.07) 5.9(4.2) 1.8(0.9)	0.88(0.03) 4.8(1.7) 1.4(0.4)	0.62(0.24) 11(4) 4.1(1.9)
DDFM	0.76(0.09) 11(3) 4.1(1.3)	0.76(0.03) 5.0(1.7) 1.7(0.3)	0.87(0.06) **2.3(0.7)** **1.0(0.5)**
PSM	**0.91(0.07)** **4.0(2.6)** **1.4(0.7)**	**0.90(0.02)** **3.6(0.8)** **1.1(0.2)**	0.84(0.07) 4.0(0.8) 1.6(0.3)

segmentation. All methods segmented the bladder similarly well. For the rectum the DDFM did not capture its volume increase with respect to the planning day, while the BM did not determine the cranial end correctly. PSM and DDFM delineated the CTV well, but the BM had larger deviations from the ground truth.

The average DSC, HD_{95}, and HD_{avg} from comparing the segmentations generated by the investigated methods and the ground truth delineations are given in Table 1. For the OARs, the PSMs gave the highest mean DSC of 0.91/0.90 for the bladder/rectum. BMs yielded slightly lower mean DSC values of 0.89/0.88 for bladder/rectum, however the difference for the rectum was not statistically significant. Among the three examined deep learning approaches, the DDFMs delivered the lowest (in all cases statistically significant) DSC with a mean of 0.76 for both OARs. HD_{avg} and HD_{95} showed a similar trend as the DSC assessment, however, not all differences were statistically significant. Conventional DIR with Plastimatch slightly outperformed the DDFM for the OARs, however the differences were not statistically significant. The results of the statistical analysis are provided in the supplementary material (section B).

Apart from the BM, all methods gave similar results for the CTV. The highest DSC was observed for the rigidly copied contours, while the best HDs were achieved for DDFMs. Nevertheless, for the DSC and HD_{avg} the differences between Plastimatch, copying and DDFMs lacked statistical significance.

Fig. 5 illustrates the DSC and HDs separately for each test patient. In most cases, the PSMs worked best and improved the BMs predictions considerably. The exceptions were patients 12 and 14, showing considerable differences in bladder filling between the planning and fraction days. For the bladder, the DDFMs performed notably worst, but in terms of HDs for the rectum they outperformed BMs in four out of nine cases.

4. Discussion

For OARs, the PSMs gave the best outcomes in terms of DSC and HDs, comparable to the state-of-the art in automatic pelvic segmentation [17,30–32]. For the bladder, PSMs mostly corrected larger volume misclassifications of the BM as observed in patient 01. For the rectum, the PSMs accurately identified the superior and inferior ends on fraction MRIs, in agreement with the planning contours. However, this intrinsic feature might hinder PSM performance for organs that are likely to change shape in the course of treatment or in the case of imprecise planning contours. This is evident in Fig. 5, patient 12, who did not follow the drinking protocol consistently. For OARs, the DDFMs did not perform satisfactorily. Regardless of the presence or absence of the regularization term during training the predicted fields were limited to

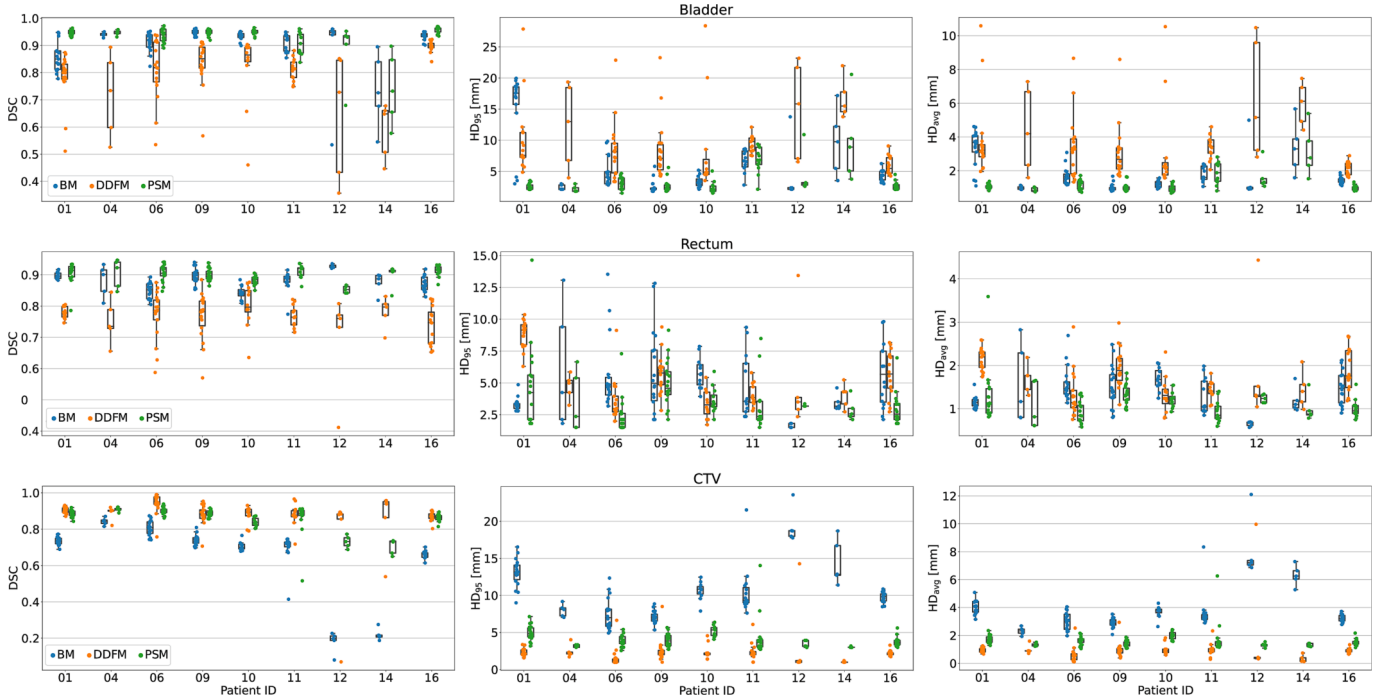

Fig. 5. Results for (top) bladder, (middle) rectum, and (bottom) clinical target volume (CTV) comparing the predictions of (blue) the baseline, (orange) deformation vector field, and (green) the patient-specific training for the nine test patients. A single point on the plot represents (left) the Dice similarity coefficient (DSC), (middle) 95th percentile Hausdorff distance (HD_{95}), and (right) average Hausdorff distance (HD_{avg}) for a predicted fraction contour.

small deformations. Moderate volume differences in OAR fillings, which could easily be captured by the PSMs, were not predicted by the DDFMs. The presence of substantial anatomical changes requires the DDFMs to model long-distance relationships between voxels, while in segmentation tasks (BMs, PSMs), local information is more critical. This is in agreement with the work of Luo et al. [33] which suggests that due to the limited effective receptive field of convolution operations their ability to model long-range spatial relations is limited. Moreover, it has been demonstrated in Li et al. [34] that as the convolutional layers deepen, the impact of far-away voxels lowers quickly. This problem could be potentially solved by architectures based on transformers, as the intrinsic self-attention mechanisms have larger effective receptive fields, making them capable of capturing long-range spatial information [35].

Nevertheless, the potential benefit of DDFMs over the BMs lies in the utilization of the planning delineations as the starting point for the predicted contours. Aside from the rectum, this may be beneficial for other tubular organs, e.g. the esophagus or spinal canal, where physicians choose to contour only a section of the organ near the PTV. This benefit can be seen in Fig. 5 where the rectum HDs for several patients is lower for DDFM than for BM.

The advantage of using planning segmentation as a starting point for the fraction contours, applies also to the conventional DIR by Plastimatch. Both methods yielded similar DSC and HDs. However, the DIR took approximately 1 min to register a pair of images and deform one contour set, while the generation of DDFM contours required 1-2 s.

For the CTV segmentation, the methods involving planning contours in some way (all but BM) performed similarly well, with rigid copying having the highest DSC and DDFMs showing the best HDs. However, cropping of all images around the CTV centroid most likely led to a better performance of contour copying than could have been achieved clinically. The high performance of methods utilizing planning contours was to be expected due to the way prostate CTV is delineated in clinical practice. To avoid unexpected changes from the applied deformations, the planning CTV contours are rigidly copied to the fraction anatomy and only slightly adjusted, if necessary. Similar good performance of DDFMs for CTV has been confirmed by Eppenhof et al. [20], showing DSC/HD_{95} of 0.86/5.66 mm. However, we are not aware of any studies showing high performance of DDFMs for OARs.

There are some limitations of this study. Only the two key OARs for prostate cancer patients were considered. Nevertheless, the methods should be applicable to other organs. Moreover, bladder and rectum undergo considerable volumetric changes, and serve as an excellent evaluation scenario for the networks considered in this work. The dataset size is another potential limitation. For most patients only the planning MRI was included due to the tedious process of data export and to the best of our knowledge, the are no publicly available datasets collecting MR-Linac data. Additionally, only geometric assessment with DSC and HDs has been provided, while a qualitative analysis such as physician's grading [19] could better gauge the clinical value of the predicted contours. Finally, some OAR ground truth fraction contours showed sub-optimal quality, as time constraints led to corrections being applied primarily around the PTV. Based on the geometric metrics and our visual inspection, none of the investigated segmentation methods appear fully reliable, necessitating physician's inspection before clinical use. Nevertheless, prior studies suggest, that correcting deep learning contours is faster than manual contouring from scratch [11,12].

To summarize, on average patient-specific U-Net models (PSMs) improved segmentation compared to BMs. DDFMs predicted only limited deformations and achieved good results for the CTV, while being less suitable for organs undergoing substantial volume changes. As a next step, transformer models [36] involving attention mechanisms will be investigated as an alternative to DDFMs. Another method to be explored will be U-Nets taking the planning MRI with manual contours as an additional input [37] to aid segmentation on fraction images.

Ethics statement

All LMU patients provided informed written consent within the scope of an ethics approved study protocol in place at the Department of Radiation Oncology of the LMU Munich University Hospital (ethics project number 20–291).

Images from Fondazione Policlinico Universitario "Agostino Gemelli" were collected under EC authorization number 3460 for image analysis.

CRediT authorship contribution statement

Maria Kawula: Investigation, Software, Formal analysis, Data curation, Writing - original draft, Visualization. **Marica Vagni:** Data curation, Writing - review & editing. **Davide Cusumano:** Data curation, Writing - review & editing. **Luca Boldrini:** Data curation, Writing - review & editing. **Lorenzo Placidi:** Data curation, Writing - review & editing. **Stefanie Corradini:** Supervision, Writing - review & editing. **Claus Belka:** Supervision, Writing - review & editing. **Guillaume Landry:** Conceptualization, Writing - review & editing, Supervision. **Christopher Kurz:** Conceptualization, Writing - review & editing, Supervision, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. The Department of Radiation Oncology of the University Hospital of LMU Munich has a research agreement with ViewRay. ViewRay did not fund this study and was not involved and had no influence on the study design, the collection or analysis of data, or on the writing of the manuscript.

Acknowledgments

The authors wish to thank Vanessa Filipa Da Silva Mendes and Lukas Nierer for their support with the data export as well as the Thesis Advisory Committee members Prof. Sibylle Ziegler and Dr. Seyed-Ahmad Ahmadi. Special thanks to Martin Rädler for discussions throughout the study, help in designing figures and proofreading the manuscript.

This work was funded by the Wilhelm Sander-Stiftung (2019.162.1 and 2019.162.2).

References

[1] Klüter S. Technical design and concept of a 0.35 T MR-Linac. Clin Transl Oncol 2019;18:98–101. doi: 10.1016/j.ctro.2019.04.007.
[2] 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 2018;30:720–7. doi: 10.1016/j.clon.2018.08.010.
[3] 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;14:1–12. https://doi.org/10.1186/s13014-019-1308-y.
[4] Finazzi T, Palacios MA, Spoelstra FOB, Haasbeek CJA, Bruynzeel AME, Slotman BJ, et al. Role of On-Table Plan Adaptation in MR-Guided Ablative Radiation Therapy for Central Lung Tumors. Int J Radiat Oncol Biol Phys 2019;104:933–41. https://doi.org/10.1016/j.ijrobp.2019.03.035.
[5] Bruynzeel AME, Tetar SU, Oei SS, Senan S, Haasbeek CJA, Spoelstra FOB, et al. A Prospective Single-Arm Phase 2 Study of Stereotactic Magnetic Resonance Guided Adaptive Radiation Therapy for Prostate Cancer: Early Toxicity Results. Int J Radiat Oncol Biol Phys 2019;105:1086–94. https://doi.org/10.1016/j.ijrobp.2019.08.007.
[6] Kontaxis C, Bol GH, Lagendijk JJW, Raaymakers BW.A new methodology for inter- and intrafraction plan adaptation for the MR-linac. Phys Med Biol 2015;60:7485. doi: 10.1088/0031-9155/60/19/7485.
[7] 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 Oncol 2019;394:385–95. https://doi.org/10.1016/s0140-6736(19)31131-6.

[8] Güngör G, Serbez Í, Temur B, Gür G, Kayalílar N, Mustafayev TZ, et al. Time Analysis of Online Adaptive Magnetic Resonance-Guided Radiation Therapy Workflow According to Anatomical Sites. Pract Radiat Oncol 2021;11:e11–21. https://doi.org/10.1016/j.prro.2020.07.003.
[9] Rogowski P, von Bestenbostel R, Walter F, Straub K, Nierer L, Kurz C, et al. Feasibility and Early Clinical Experience of Online Adaptive MR-Guided Radiotherapy of Liver Tumors. Cancers 2021;13:1523. https://doi.org/10.3390/cancers13071523.
[10] Hadi I, Eze C, Schönecker S, von Bestenbostel R, Rogowski P, Nierer L, et al. MR-guided SBRT boost for patients with locally advanced or recurrent gynecological cancers ineligible for brachytherapy: feasibility and early clinical experience. Radiat Oncol 2022;17:1–9. https://doi.org/10.1186/s13014-022-01981-z.
[11] Cha E, Elguindi S, Onochie I, Gorovets D, Deasy JO, Zelefsky M, et al. Clinical implementation of deep learning contour autosegmentation for prostate radiotherapy. Radiother Oncol 2021;159:1–7. https://doi.org/10.1016/j.radonc.2021.02.040.
[12] Zabel WJ, Conway JL, Gladwish A, Skliarenko J, Didiodato G, Goorts-Matthews L, et al. Clinical Evaluation of Deep Learning and Atlas-Based Auto-Contouring of Bladder and Rectum for Prostate Radiation Therapy. Pract Radiat Oncol 2021;11:e80–9. https://doi.org/10.1016/j.prro.2020.05.013.
[13] Veiga-Canuto D, Cerdà-Alberich L, Sangüesa Nebot C, Martínez de las Heras B, Pötschger U, Gabelloni M, et al. Comparative Multicentric Evaluation of Inter-Observer Variability in Manual and Automatic Segmentation of Neuroblastic Tumors in Magnetic Resonance Images. Cancers 2022;14:3648. doi: 10.3390/cancers14153648.
[14] Chlebus G, Meine H, Thoduka S, Abolmaali N, Van Ginneken B, Hahn HK, et al. Reducing inter-observer variability and interaction time of MR liver volumetry by combining automatic CNN-based liver segmentation and manual corrections. PLoS One 2019;14:e0217228. https://doi.org/10.1371/journal.pone.0217228.
[15] Liang F, Qian P, Su K-H, 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. https://doi.org/10.1016/j.artmed.2018.07.001.
[16] 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;45:5129–37. https://doi.org/10.1002/mp.13221.
[17] Fransson S, Tilly D, Strand R. Patient specific deep learning based segmentation for magnetic resonance guided prostate radiotherapy. Phys Imaging Radiat Oncol 2022;23:38–42. https://doi.org/10.1016/j.phro.2022.06.001.
[18] Li Z, Zhang W, Li B, Zhu J, Peng Y, Li C, et al. Patient-specific daily updated deep learning auto-segmentation for MRI-guided adaptive radiotherapy. Radiother Oncol 2022;177:222–30. https://doi.org/10.1016/j.radonc.2022.11.004.
[19] Kawula M, Hadi I, Nierer L, Vagni M, Cusumano D, Boldrini L, et al. Patient-specific transfer learning for auto-segmentation in adaptive 0.35 T MRgRT of prostate cancer: a bi-centric evaluation. Med Phys 2023;50:1573–85. doi: 10.1002/mp.16056.
[20] Eppenhof KAJ, Maspero M, Savenije MHF, de Boer JCJ, van der Voort van Zyp JRN, Raaymakers BW, et al. Fast contour propagation for MR-guided prostate radiotherapy using convolutional neural networks. Med Phys 2020;47:1238–48. doi: 10.1002/mp.13994.
[21] Sharp GC, Li R, Wolfgang J, Chen GTY, Peroni M, Spadea MF, et al. Plastimatch: An Open Source Software Suite for Radiotherapy Image Processing. Proceedings of the XVI'th International Conference on the use of Computers in Radiotherapy (ICCR), Amsterdam, Netherlands, 2010.
[22] Ma N, Li W, Brown R, others. Project MONAI. Zenodo, CERN 2021. doi: 10.5281/zenodo.4323058.
[23] Kerfoot E, Clough J, Oksuz I, Lee J, King AP, Schnabel JA. Left-Ventricle Quantification Using Residual U-Net. International workshop on statistical atlases and computational models of the heart, 2018, p. 371–80. doi: 10.1007/978-3-030-12029-0_40.
[24] Jaderberg M, Simonyan K, Zisserman A, Kavukcuoglu K. Spatial Transformer Networks. Adv Neurol 2015;28. URL:https://proceedings.neurips.cc/paper_files/paper/2015/file/33ceb07bf4eeb3da587e268d663aba1a-Paper.pdf.
[25] Pérez-García F, Sparks R, Ourselin S. TorchIO: a Python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning. Comput Methods Programs Biomed 2021:106236. doi: 10.1016/j.cmpb.2021.106236.
[26] McCormick M, Liu X, Jomier J, Marion C, Ibanez L. ITK: enabling reproducible research and open science. Front Neuroinform 2014;8:13. https://doi.org/10.3389/fninf.2014.00013.
[27] Eppenhof KAJ, Lafarge MW, Veta M, Pluim JPW. Progressively Trained Convolutional Neural Networks for Deformable Image Registration. IEEE Trans Med Imaging 2019;39:1594–604. https://doi.org/10.1109/TMI.2019.2953788.
[28] Fu Y, Brown NM, Saeed SU, Casamitjana A, Baum ZMC, Delaunay R, et al. DeepReg: a deep learning toolkit for medical image registration. J Open Source Softw 2020;5:2705. https://doi.org/10.21105/joss.02705.
[29] Hu Y, Modat M, Gibson E, Li W, Ghavami N, Bonmati E, et al. Weakly-supervised convolutional neural networks for multimodal image registration. Med Image Anal 2018;49:1–13. https://doi.org/10.1016/j.media.2018.07.002.
[30] Huang S, Cheng Z, Lai L, Zheng W, He M, Li J, et al. Integrating multiple MRI sequences for pelvic organs segmentation via the attention mechanism. Med Phys 2021;48:7930–45. https://doi.org/10.1002/mp.15285.
[31] Sanders JW, Lewis GD, Thames HD, Kudchadker RJ, Venkatesan AM, Bruno TL, et al. Machine Segmentation of Pelvic Anatomy in MRI-Assisted Radiosurgery (MARS) for Prostate Cancer Brachytherapy. Int J Radiat Oncol Biol Phys 2020;108:1292–303. https://doi.org/10.1016/j.ijrobp.2020.06.076.
[32] Savenije MHF, Maspero M, Sikkes GG, van der Voort van Zyp JRN, Kotte ANTJ, Bol GH, et al. Clinical implementation of MRI-based organs-at-risk auto-segmentation with convolutional networks for prostate radiotherapy. Radiat Oncol 2020;15:1–12. doi: 10.1186/s13014-020-01528-0.
[33] Luo W, Li Y, Urtasun R, Zemel R. Understanding the Effective Receptive Field in Deep Convolutional Neural Networks. Adv Neurol 2016;29. doi: 10.48550/arXiv.1701.04128.
[34] Li S, Sui X, Luo X, Xu X, Liu Y, Goh R. Medical Image Segmentation using Squeeze-and-Expansion Transformers. In: Zhou Z-H, editor. Proceedings of the thirtieth international joint conference on artificial intelligence, IJCAI-21, International Joint Conferences on Artificial Intelligence Organization; 2021, p. 807–15. doi: 10.24963/ijcai.2021/112.
[35] Chen J, Frey EC, He Y, Segars WP, Li Y, Du Y. Transmorph: Transformer for unsupervised medical image registration. Med Image Anal 2022;82:102615. https://doi.org/10.1016/j.media.2022.102615.
[36] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is All you Need. NIPS 2017. https://doi.org/10.48550/arXiv.1706.03762.
[37] Klymenko T, Kim ST, Lauber K, Kurz C, Landry G, Navab N, et al. Butterfly-Net: Spatial-Temporal Architecture for Medical Image Segmentation. 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI), 2021, p. 616–20. doi: 10.1109/ISBI48211.2021.9433939.

5.4 Paper IV

The study described in Paper 4 aimed to delve deeper into the personalized auto-segmentation methods introduced in Paper 2. The new aspects included continuous fine-tuning of personalized models with segmented fraction images and testing the necessity of population models as a starting point for personalized networks. The target group included patients diagnosed with a tumor in the abdomen that were irradiated at the MRIdian MR-LINAC. The abdomen region was selected due to the large database of available patients and the high variability of OARs. They include organs with changing shape and filling (bowel, duodenum, stomach), long organs segmented only in the proximity of the target volume rather than as a whole (aorta, spinal canal), and organs not varying much and having a rather simple shape (liver, kidneys).
This is the first study investigating the influence of DL population models on personalized models and their integration with progressive training, offering a comparison with methods exclusively relying on individual patient data.
In the first experiment, the population auto-segmentation models were trained and subsequently fine-tuned with the patient's segmented planning MRI to generate personalized models. The second experiment examined the benefit of further fine-tuning the personalized models with new segmented fraction MRIss. The third experiment investigated the necessity of population models for patient-specific networks, i.e., the latter were trained using data only from the single patient of interest. All methods were evaluated using geometric metrics, including DSC and HD. Additionally, a radiation oncologist conducted a qualitative analysis to rate the clinical usability of the predicted contours.
The results of the first experiment confirmed the conclusions of the previous prostate study. That is, personalized networks generated by fine-tuning population models with the segmented planning image of a given patient improve segmentation quality. The second experiment found that additional training with the subsequent fraction images further improves the predictions, yet the gain is small and not always statistically significant. The third experiment demonstrated that personalized models trained from scratch showed similar geometric performance as the population models but resulted in slightly more outliers and a larger spread of values. However, the oncologist rated contours predicted by models trained from scratch as clearly worse than those predicted by the population models. They were shown to require more manual corrections before being usable for treatment adaptation. It is worth noting that not all organs benefited equally from the patient-specific fine-tuning. The organs commonly considered challenging to segment, including the duodenum, bowel, and stomach, showed the biggest improvements. In the case of the aorta and spinal canal, personalized training helped to determine the superior and inferior ends of the contours. The least profiting organs were the kidneys and liver. It takes less than an hour to train a personalized model, making it feasible for implementation in clinical routines.

Received: 22 January 2024 | Revised: 27 October 2024 | Accepted: 29 November 2024

DOI: 10.1002/mp.17580

RESEARCH ARTICLE

MEDICAL PHYSICS

Personalized deep learning auto-segmentation models for adaptive fractionated magnetic resonance-guided radiation therapy of the abdomen

Maria Kawula[1] | **Sebastian Marschner**[1] | **Chengtao Wei**[1] | **Marvin F. Ribeiro**[1] | **Stefanie Corradini**[1] | **Claus Belka**[1,2,3] | **Guillaume Landry**[1] | **Christopher Kurz**[1]

[1]Department of Radiation Oncology, LMU University Hospital, LMU Munich, Munich, Germany

[2]German Cancer Consortium (DKTK), partner site Munich, a partnership between DKFZ and LMU University Hospital Munich, Munich, Germany

[3]Bavarian Cancer Research Center (BZKF), Munich, Germany

Correspondence
Christopher Kurz, Department of Radiation Oncology, LMU University Hospital, LMU Munich 81377, Munich, Germany.
Email: Christopher.Kurz@med.uni-muenchen.de

Funding information
Wilhelm Sander-Stiftung, Grant/Award Number: 2019.162.2

Abstract

Background: Manual contour corrections during fractionated magnetic resonance (MR)-guided radiotherapy (MRgRT) are time-consuming. Conventional population models for deep learning auto-segmentation might be suboptimal for MRgRT at MR-Linacs since they do not incorporate manual segmentation from treatment planning and previous fractions.

Purpose: In this work, we investigate patient-specific (PS) auto-segmentation methods leveraging expert-segmented planning and prior fraction MR images (MRIs) to improve auto-segmentation on consecutive treatment days.

Materials and Methods: Data from 151 abdominal cancer patients treated at a 0.35 T MR-Linac (151 planning and 215 fraction MRIs) were included. Population baseline models (BMs) were trained on 107 planning MRIs for one-class segmentation of the aorta, bowel, duodenum, kidneys, liver, spinal canal, and stomach. PS models were obtained by fine-tuning the BMs using the planning MRI (PS_{BM}). Maximal improvement by continuously updating the PS models was investigated by adding the first four out of five fraction MRIs (PS^{F4}_{BM}). Similarly, PS models without BM were trained (PS_{noBM} and PS^{F4}_{noBM}). All hyperparameters were optimized using 23 patients, and the methods were tested on the remaining 21 patients. Evaluation involved Dice similarity coefficient (DSC), average (HD_{avg}) and the 95th percentile (HD_{95}) Hausdorff distance. A qualitative contour assessment by a radiation oncologist was performed for BM, PS_{BM}, and PS_{noBM}.

Results: PS^{F4}_{BM} and PS_{BM} networks had the best geometric performance. PS_{noBM} and BMs showed similar DSC and HDs values, however PS^{F4}_{noBM} models outperformed BMs. PS_{BM} predictions scored the best in the qualitative evaluation, followed by the BMs and PS_{noBM} models.

Conclusion: Personalized auto-segmentation models outperformed the population BMs. In most cases, PS_{BM} delineations were judged to be directly usable for treatment adaptation without further corrections, suggesting a potential time saving during fractionated treatment.

KEYWORDS
auto-segmentation, MR-Linac, patient-specific transfer learning

1 | INTRODUCTION

Magnetic resonance linear accelerators (MR-Linacs) enable online adaptive MR-guided radiation therapy (MRgRT).[1,2] This technology allows for the monitoring of anatomical changes prior to and during patient irradiation, with no additional imaging dose. As a consequence, the ablative dose can be delivered in fewer fractions and with reduced safety margins.[3–5]

The currently available commercial solution for deep learning (DL) auto-segmentation at MR-Linacs[6] and the majority of published research[7–10] employ population-based DL models trained on larger datasets of expert-segmented MR images (MRIs). By design, the networks learn common features shared among a wide range of patients, thus generating segmentations that are combinations of the examples they were trained on. However, this may be a sub-optimal solution for fractionated treatment, that consists of a pre-treatment planning phase and the subsequent series of irradiations called fractions.[11] During irradiation, the manually segmented planning MRI, as well as images from previous fractions with contours approved by radiation oncologists, are available but not integrated into the population models.

Previous studies have examined whether utilizing a patient's segmented planning MRI for fine-tuning a population model[12,13] or training from scratch[14] can enhance the auto-segmentation performance of fraction images in prostate patients. A similar 2D method for patients diagnosed with cancer in the abdomen region was presented by Li et al.,[15] where the personalized models were updated daily with newly acquired data. Nevertheless, each of the presented studies exclusively focused on a single method, and all of them come with their limitations. The first two were conducted for prostate cases, where patient anatomy is relatively simple and stable. The last one included only six patients, permitting only a limited evaluation of the presented methods.

The goal of this work was to perform an investigation of approaches leveraging prior knowledge available in MRgRT in order to enhance the quality of abdominal organs-at-risk (OARs) auto-segmentation on fraction MRIs. Four methods of generating patient-specific (PS) models were compared to population baseline models (BMs) via geometric metrics and a qualitative evaluation by a trained radiation oncologist. The BMs were fine-tuned using either only the segmented planning image or the planning and first four out of five fraction images of a specific patient. Furthermore, personalized models were trained from scratch instead of using BMs as a starting point, with only the planning or the planning and first four fraction MRIs. To the best of our knowledge, this is the first study to investigate the impact of population BMs on personalized segmentation models, combining them with progressive training and comparing these methods with approaches relying solely on individual patient data.

2 | MATERIAL AND METHODS

2.1 | Dataset and data pre-processing

The dataset was collected retrospectively and comprised 151 cases, including 84 males and 67 females. The median of patient's age was 68 years, ranging from 34 to 91 years old. These patients were treated at the 0.35 T MR-Linac (MRIdian, ViewRay Inc, Cleveland, Ohio) at the Department of Radiation Oncology of the LMU Munich University Hospital between January 2020 and December 2022. Tumor sites included the pancreas, liver, and lesions in the abdomen. All MRIs were acquired with a balanced steady-state free precession (bSSFP) sequence with an in-plane resolution of 1.5 mm × 1.5 mm and 1.5 or 3 mm thickness of the axial slices. For each case, there were one planning and between 1 (single-shot treatment) and 5 fraction MRIs included (in total 151 planning and 215 fraction images). Figure S1 in the supplementary material shows segmented MRIs of three exemplary patients on all irradiation days and the planning day. Informed written consent was obtained from all patients, and the study was carried out in accordance with relevant ethics guidelines and regulations (ethics project number 20-291).

Planning MRIs were segmented manually by different trained oncologists several days prior to the irradiation as part of the clinical routine. They were used as ground truth planning contours. During each fraction, the planning and daily MRIs were registered in the treatment planning system, and the resulting vector field was used to deform the planning OAR contours to the anatomy of the day. The deformed structures were adjusted and approved by experienced radiation oncologists. In this work, they served as ground truth fraction contours. The original images and contours were stored in Digital Imaging and Communications in Medicine (DICOM) and Radiotherapy Structure (RT-Struct) formats, respectively. For the purpose of this work, they were converted to voxelized MetaImage format (mha) using plastimatch.[16]

Depending on the exact tumor location, different OARs were delineated for each patient. In this study, the most frequently segmented ones were considered: the aorta, bowel, duodenum, kidneys, liver, spinal canal, and stomach. Table 1 reports the number of MRIs with specific OAR segmentations and patient split into three sets. Set 1 was used for the BM training. Set 2 was used to validate BM training and for PS hyperparameter search. Set 3 was utilized only for testing. The patient demographic was well-balanced across all three sets. The median patient age in all three groups was between 65 and 68 years old. The male-to-female ratio in all three groups was between 0.8 and 1.4.

The pre-processing of 3D MRIs and contours consisted of three steps. First, the contours and MRIs acquired with a 3 mm slice thickness were re-sampled to 1.5 mm slice thickness using nearest neighbor and

TABLE 1 Number of organ-at-risk (OAR) contours used in the study for Set 1, Set 2, and Set 3.

OAR	Set 1	Set 2	Set 3
Aorta	81	20	21
Bowel	95	23	21
Duodenum	77	22	21
Kidney left	77	19	20
Kidney right	83	21	21
Liver	101	23	21
Spinal canal	101	22	20
Stomach	95	22	21
Total MRIs	107	23	21
Median age	68 (34–91)	66 (48–91)	65 (54–83)
M/F ratio	1.4	0.8	1.1
Patient IDs	Pat_{001}-Pat_{107}	Pat_{108}-Pat_{130}	Pat_{131}-Pat_{151}

Note: The median age with range, male-to-female (M/F) ratio and patients' IDs are given for each set.

linear interpolation, respectively. Second, for the 3D models, all images were cropped centrally or zero-padded to dimensions of $256 \times 256 \times 256$. The same was applied for the 2D network data except for no padding/cropping along the superior-inferior axis. The number of patients with images having 160 or 288 slices in the axial direction was 27 and 80 in set 1, 6 and 17 in set 2, and 6 and 15 in set 3. Third, the image intensities were normalized to values between 0 and 1, with clipping applied at the 99th percentile of the image intensity to account for potential MR artifacts with high intensities.

2.2 | Baseline model

For benchmarking and as a basis for the subsequent personalized models, state-of-the-art one-class 3D U-Nets were trained to obtain conventional population models, that is, models trained on large datasets that generalize effectively to unseen examples. Our prior experience showed that one-class model performance surpasses the multi-class models. The networks were trained on planning images from 107 randomly selected patients (Set 1, Pat_{001}-Pat_{107}) and validated on planning images from 23 patients (Set 2, Pat_{108}-Pat_{130}). The remaining 21 patients (Set 3, Pat_{131}-Pat_{151}) were used as an independent test set (for the exact numbers of MRIs for each organ, please refer to Table 1). From here on, these models will be referred to as the BMs. The BM training began with a random initialization of the 3D U-Net parameters. The initial learning rate (lr) was set to 0.001 and decreased to 0.0005 and 0.0001 after 100 and 200 epochs, respectively. The BMs were trained over 300 epochs with a batch size of 1. The lr values and epochs at which the changes were applied were determined empirically based on observations of the validation and training learning curves. Details on data augmentation and hyperparameter search are provided in the supplementary material.

2.3 | Personalized models

Since the personalized models must be trained before the first fraction, no validation data is available to monitor the training progress. Therefore, all hyperparameters and the training duration must be known in advance. In order to determine these, different combinations of hyperparameters were investigated for patients Pat_{108}-Pat_{130} (Set 2). The final set of hyperparameters was selected based on the highest mean Dice similarity coefficient (DSC) achieved among these patients (for more details on the hyperparameter search, we refer to the supplementary material). The final testing was carried out using these fixed parameters for patients Pat_{131}-Pat_{151} (Set 3). Figure 1 presents the PS approaches that have been investigated:

PS with BM (PS_{BM}):
In this method, the personalized models were generated by fine-tuning the BM with a given patient's segmented planning MRI. The patient's 5th fraction image was used to validate the model's performance. The initial lr was set to 0.0001 but reduced to 0.00005 and 0.00001 after 300 and 400 epochs, respectively. These models were trained over 500 epochs with a batch size of 1. Figure 2 shows exemplary validation curves from PS_{BM} training for the aorta, bowel, and right kidney.

PS training from scratch (PS_{noBM}):
This method investigated the importance of BMs for personalized models. Instead of having BMs as a starting point, PS_{noBM} networks were randomly initialized and trained from scratch with the segmented planning MRI of a given patient. For this purpose, 2D models were implemented instead of 3D ones, as the latter proved unreliable in preliminary experiments. Using 2D models reduced network complexity and increased the number of training examples by treating each axial slice as an independent image. The models were validated on the corresponding 5th fraction data. The initial lr was set to 0.0001 but was reduced to 0.00005 and 0.00001 after 400 and 800 epochs, respectively. PS_{noBM} models were trained over 1280 epochs with a batch size of 3.

Progressive training of PS models:
In the last experiment, the potential benefits of including fraction data in PS training have been investigated. Instead of using only the planning data of a given patient, the PS models could be updated further after each fraction with the newly segmented MRI. In this work, the presumed upper limit of this approach has been tested for patients undergoing five fractions, which is a common fractionation scheme at MR-Linacs.[17–19] The BMs or 2D randomly-initialized networks were

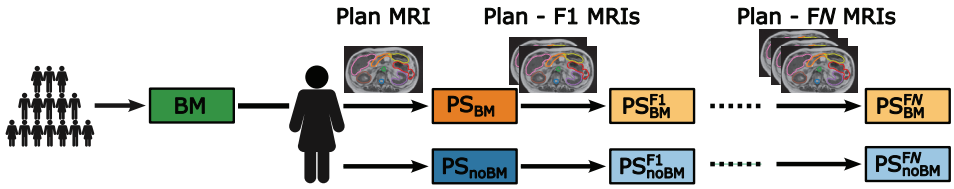

FIGURE 1 Workflow of the investigated training strategies. The boxes represent the investigated models, while the arrows indicate the process of training or fine-tuning. The organ-specific one-class population BMs were trained on a cohort of 107 patients. Subsequently, BMs were fine-tuned by PS training either with the planning (Plan MRI) or the planning and the first $N = 4$ F images yielding PS_{BM} and PS^{F4}_{BM} models, respectively. Repeating the process without the BMs for initializing the model weights and biases resulted in PS_{noBM} and PS^{F4}_{noBM} models, respectively. BMs, baseline models; F, fraction; MRI, magnetic resonance imaging; PS, patient-specific.

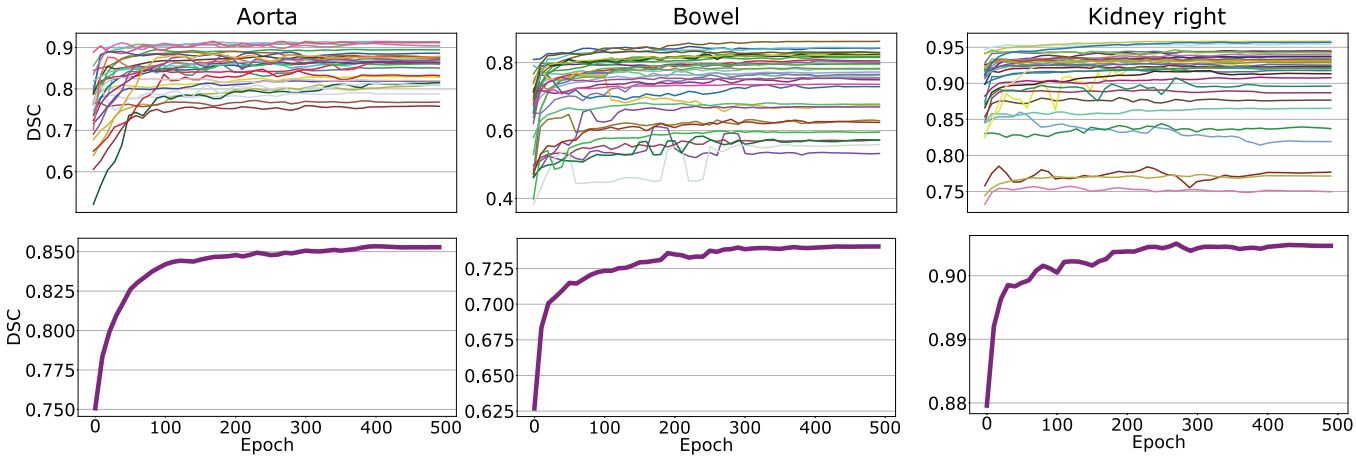

FIGURE 2 Exemplary validation curves for PS_{BM} training for the aorta, bowel, and right kidney. The upper panel displays individual DSC curves for each patient across the training epochs. In the lower panel, cumulative curves depict average DSC scores across all validation patients. DSC for epoch 0 corresponds to the BM performance. BM, baseline model; DSC, dice similarity coefficient.

fine-tuned with the planning and the first four fraction images, resulting in PS^{F4}_{BM} and PS^{F4}_{noBM} models, respectively. For PS^{F4}_{BM} training, the initial lr was set to 0.0001 and decreased to 0.00005 and 0.00001 after 60 and 80 epochs, respectively. PS^{F4}_{BM} models were trained over 100 epochs with a batch size of 1, which resulted in the same number of network updates as for the PS_{BM} models. For PS^{F4}_{noBM} training, the initial lr was set to 0.0001 and decreased to 0.00005 and 0.00001 after 80 and 160 epochs, respectively. PS^{F4}_{noBM} models were trained over 256 epochs with a batch size of 3, which resulted in the same number of network updates as for the PS_{noBM} models.

2.4 | Implementation and technical details

The 2D and 3D MONAI[20] implementations of the residual U-Net developed by Kerfoot et al.[21] were used based on our previous work.[9,12] The networks had 4 resolution levels, each comprising two convolutions with $3 \times 3(\times 3)$ kernels, followed by instance normalization[22] and parametric rectified linear unit (PReLU)[23] activation with an initial slope of 0.2 for negative arguments. For down-sampling in the encoding arm and up-sampling in the decoding arm, a convolution with a stride of 2 and up-convolution were employed, respectively. The output layer of the network featured soft-max activation[24] and thresholding at 0.5. Due to the low foreground-to-background pixel ratio, a DSC-based loss function was chosen[25] for training.

All trainings were performed on Nvidia Quadro RTX 8000 or Nvidia RTX A6000 GPUs.

2.5 | Data evaluation

Since the PS^{F4}_{BM} and PS^{F4}_{noBM} models were trained on data from fractions 0–4, their test set was limited to the 5th fraction image. To ensure a fair comparison between all the investigated methods, the outcomes presented in this study will focus on the predictions on the 5th fractions alone (Set 3, 21 test patients).

Network-predicted contours were compared to the ground truth segmentation used clinically via DSC, the 95th percentile Hausdorff distance (HD_{95}) and the average Hausdorff distance (HD_{avg}). For two binary images,

A and B, each having N voxels the DSC is defined as:

$$\mathrm{DSC} = \frac{2\sum_{i=0}^{N} a_i b_i}{\sum_{i=0}^{N} a_i^2 + \sum_{i=0}^{N} b_i^2} \quad (1)$$

where a_i and b_i are binary pixel values belonging to images A and B, respectively. The Hausdorff distance (HD) is defined as:

$$\begin{aligned} \mathrm{HD} &= \max(\mathrm{hd}(A, B), \mathrm{hd}(B, A)) \quad \text{and} \quad \mathrm{hd}(A, B) \\ &= \max_{a \in \partial A} \min_{\in \partial B} ||\vec{r}(a) - \vec{r}(b)||_2. \end{aligned} \quad (2)$$

where ∂A and ∂B denote the boundary voxels within the structure ($a_i = 1, b_i = 1$) of images A and B, respectively, while $\vec{r}(\cdot)$ is a position vector for image voxels. All metrics were calculated in 3D. In addition, a senior radiation oncologist working clinically at the MR-Linac for over 3 years assessed the usefulness of the predicted OAR contours for plan adaptation. The grades, ranging from 0 to 4, corresponded to the following statements: 0-ideal, 1-clinically acceptable, 2-minor corrections required, 3-major corrections required, and 4-unusable. The contour sets were presented in a random order, withholding their origin. Additionally, the planning target volume (PTV) and its 3 cm isotropic expansion were shown to indicate the high-dose region, mimicking the clinical practice. The radiation oncologist reviewed the auto-segmented 5$^{\text{th}}$ fraction MRIs of the 21 test patients. In this analysis, the BM, $\mathrm{PS_{BM}}$, and $\mathrm{PS_{noBM}}$ were included as methods suitable for all fractionation schemes and from the first fraction onwards.

Statistical analysis:
HDs and $1 - \mathrm{DSC}$ values from the 5$^{\text{th}}$ fractions of the test patients were combined into vectors for each network and organ. Following that, the non-parametric Friedman test[26] was carried out. Since the latter indicated statistically significant differences among the methods for all organs, a post-hoc Nemenyi test[27] was conducted to calculate p-scores for all pairs of methods. Values of $p < 0.05$ were assumed to indicate statistically significant differences.

3 | RESULTS

Figure 3 shows axial slices from an exemplary test patient with predictions from all investigated DL models compared to the ground truth segmentation. In this case, all methods segmented the liver, left kidney, and spinal canal similarly well. For the stomach, duodenum, bowel, and aorta $\mathrm{PS_{BM}}$, $\mathrm{PS^{F4}_{BM}}$, and $\mathrm{PS^{F4}_{noBM}}$ performed the best, while predictions from the remaining models showed larger deviations from the clinical ground truth.

Table 2 and Figure 4 present the geometric performance of the investigated methods on the set 3. Fine-tuning the BMs with PS data showed the best results among the investigated approaches and significantly improved the geometric metrics compared to conventional BMs. For the liver/kidneys/stomach the $\mathrm{PS_{BM}}$ models improved BMs median DSC by approximately 0.02 from 0.93/0.91/0.86 to 0.95/0.935/0.88. The improvements were more pronounced for the duodenum/bowel/aorta/spinal canal, where the median DSC increased from 0.51/0.67/0.76/0.75 to 0.74/0.75/0.86/0.83. The median DSC and HDs for the $\mathrm{PS_{noBM}}$ models were comparable to those of the BMs, however, the former exhibited a larger spread of values and produced more outliers.

For both PS methods, whether with or without the BM, incorporating five images from a given patient led to better outcomes when compared to using only the planning MRI for training. This was particularly noticeable for models trained from scratch. Organs that benefited the most from the PS training were the aorta, bowel, duodenum, and spinal canal. In contrast, improvements for the kidneys, liver, and stomach were moderate.

The Friedman test revealed statistically significant differences among the approaches for all organs. Table S2 in the supplementary material presents the p-values from the post-hoc Nemenyi test. Comparison between $\mathrm{PS_{BM}}$, BM, and $\mathrm{PS_{noBM}}$ showed a statistically significant advantage of the former, for all OARs but the left kidney and spinal canal, where $\mathrm{PS_{BM}}$ and $\mathrm{PS_{noBM}}$ performed equally well. In general, BMs and $\mathrm{PS_{noBM}}$ performed similarly and showed statistically significant differences for the left kidney, stomach, and spinal canal. Increasing the number of patient images for personalized training led to statistically significant improvements in $\mathrm{PS^{F4}_{BM}}$ models for all OARs but the aorta and right kidney. In $\mathrm{PS^{F4}_{noBM}}$ models, significant improvements were noted for the duodenum and aorta.

Figure 5 presents the results of the qualitative assessment performed by a radiation oncologist. In the analysis, 70% of $\mathrm{PS_{BM}}$ contours were found directly suitable for treatment adaptation (scores 0 and 1), 25% needing minor adjustments, and the remaining 5% requiring major corrections. BM-generated delineations were also well graded, with 53% of the predictions usable right away, 26% and 16% requiring minor and major improvements, respectively. The remaining 5% were deemed not usable. Despite comparable geometric performance of the BM and $\mathrm{PS_{noBM}}$, the latter were graded clearly lower with 23% of the contours usable directly, 32% and 27% requiring minor and major corrections, respectively, and 18% deemed unusable. The average scores of the three models were 1.02, 1.54, and 2.36 respectively.

4 | DISCUSSION

For personalized auto-segmentation models, fine-tuning populaiton BMs with segmented patient images ($\mathrm{PS_{BM}}$

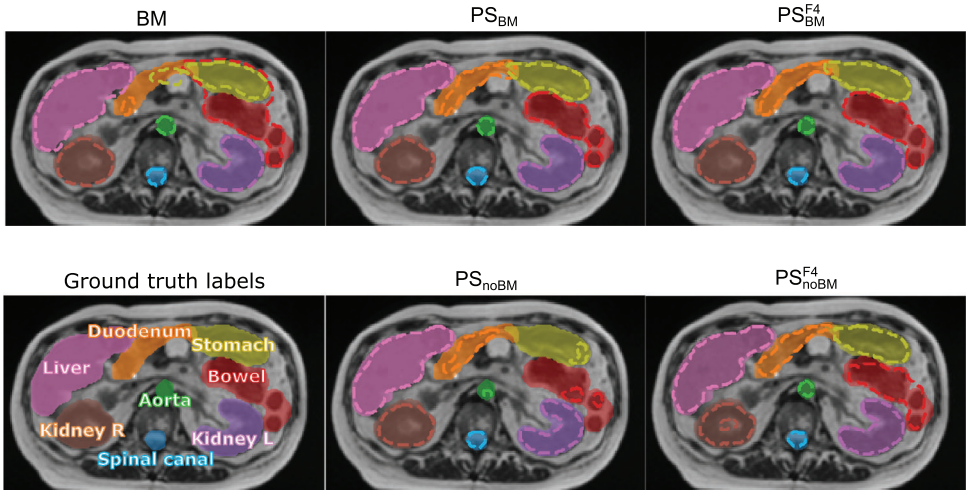

FIGURE 3 Axial view of an exemplary test patient showing predictions of (dashed lines) deep learning models versus (half-transparent background) the clinical ground truth. Predictions of the following models are shown: the BMs, PS models generated by fine-tuning the BMs with the planning (PS_{BM}) or the planning and first four fraction MRIs (PS_{BM}^{F4}), models trained from scratch only with the planning (PS_{noBM}) or with the planning and first four fraction MRIs (PS_{noBM}^{F4}). BMs, baseline models; MRI, magnetic resonance imaging; PS, patient-specific.

TABLE 2 Median and (interquartile range) of Dice similarity coefficient (DSC), 95th percentile (HD_{95}), and average (HD_{avg}) Hausdorff distance for the 5th fractions of the 21 test patients.

Model	Aorta DSC HD_{95}(mm) HD_{avg}(mm)	Bowel DSC HD_{95}(mm) HD_{avg}(mm)	Duodenum DSC HD_{95}(mm) HD_{avg}(mm)	Kidney L. DSC HD_{95}(mm) HD_{avg}(mm)	Kidney R. DSC HD_{95}(mm) HD_{avg}(mm)	Liver DSC HD_{95}(mm) HD_{avg}(mm)	Spinal C. DSC HD_{95}(mm) HD_{avg}(mm)	Stomach DSC HD_{95}(mm) HD_{avg}(mm)
BM	0.76 (0.1)	0.67 (0.14)	0.51 (0.23)	0.91 (0.03)	0.91 (0.1)	0.93 (0.03)	0.75 (0.05)	0.86 (0.07)
	22 (14)	25 (26)	16 (14)	4.9 (2.4)	5.7 (9.0)	7.3 (3.1)	14 (14)	7.7 (12)
	3.7 (3.5)	6.2 (8.8)	4.7 (5.3)	1.7 (0.6)	1.8 (2.5)	2.5 (0.9)	3.0 (2.2)	2.2 (1.7)
PS_{BM}	0.86 (0.06)	0.75 (0.14)	0.74 (0.17)	**0.94 (0.03)**	0.93 (0.05)	**0.95 (0.02)**	0.83 (0.05)	0.88 (0.06)
	6.0 (5.1)	14 (25)	8.7 (8.9)	2.9 (1)	**2.6 (1.9)**	5.7 (3.5)	6.0 (8.4)	6.1 (3.4)
	1.4 (0.8)	4.0 (7.3)	2.5 (2.5)	1.1 (0.4)	1.2 (0.6)	1.9 (0.7)	1.6 (1.2)	1.8 (0.7)
PS_{BM}^{F4}	**0.88 (0.06)**	**0.82 (0.08)**	**0.78 (0.1)**	**0.94 (0.02)**	**0.94 (0.05)**	**0.95 (0.01)**	**0.85 (0.03)**	**0.9 (0.03)**
	3.0 (2.7)	**11 (7)**	**6.2 (5.1)**	**2.6 (0.9)**	**2.6 (2.1)**	**4.5 (1.9)**	**3.1 (2.5)**	**4.0 (2.7)**
	1.1 (0.5)	**3.7 (1.5)**	**2.0 (1.2)**	**1.0 (0.2)**	**1.0 (0.8)**	**1.6 (0.7)**	**1.2 (0.3)**	**1.5 (0.7)**
PS_{noBM}	0.76 (0.15)	0.61 (0.28)	0.56 (0.33)	0.91 (0.06)	0.87 (0.18)	0.94 (0.08)	0.82 (0.13)	0.71 (0.33)
	12 (13)	32 (51)	19 (17)	3.3 (6.9)	5.7 (6.1)	6.3 (13.3)	4.5 (8.8)	15 (23)
	2.7 (2.7)	8.4 (15)	4.9 (7.1)	1.4 (1.3)	1.9 (2.0)	2.0 (3.7)	1.4 (1.5)	4.2 (9.6)
PS_{noBM}^{F4}	0.83 (0.11)	0.69 (0.22)	0.65 (0.20)	0.92 (0.04)	0.90 (0.20)	0.93 (0.04)	0.82 (0.09)	0.83 (0.08)
	8.5 (11.5)	49 (54)	14 (10)	3.1 (2.9)	4.5 (8.9)	6.9 (9.6)	11 (17)	7.3 (4.3)
	2.0 (1.8)	13 (14)	3.4 (2.6)	1.3 (0.6)	1.7 (2.3)	2.6 (2.0)	2.1 (2.5)	2.2 (0.9)

Note: For all organs-at-risk the performance of the following models are compared: the baseline models (BM), patient-specific models generated by fine-tuning the BMs with one (PS_{BM}) and five MRIs (PS_{BM}^{F4}), as well as patient-specific models trained from scratch with one (PS_{noBM}) and with 5 MRIs (PS_{noBM}^{F4}). The best metrics achieved are given in bold.

and PS_{BM}^{F4} models) was shown to significantly improve the accuracy of BM predictions for all investigated OARs. This was demonstrated by the best DSC and HDs values, as well as qualitative assessment by a trained radiation oncologist. In fact, PS_{BM} predictions were considered ready to use 30% more often than the contours generated by BMs.

The necessity of BMs for PS models has been investigated by training PS_{noBM} models. While they achieved geometric accuracy similar to BMs, the clinical

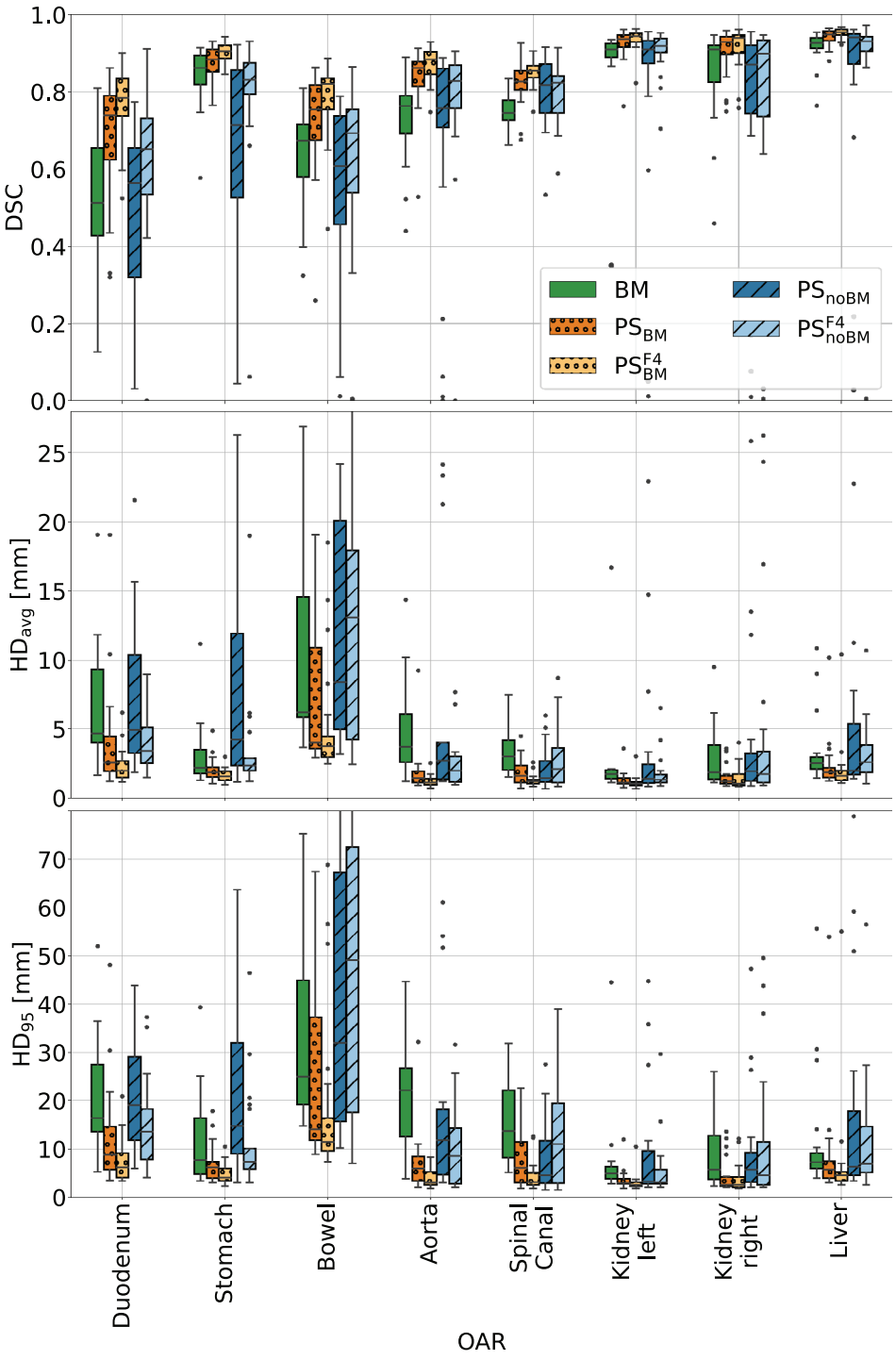

FIGURE 4 Box plots presenting DSC, 95th percentile (HD$_{95}$), and average (HD$_{avg}$) Hausdorff distance for the 5th fractions of the 21 test patients. For all organs-at-risk the performance of the following models are compared: the BMs, PS models generated by fine-tuning the BMs with one (PS$_{BM}$) and five MRIs (PS$^{F4}_{BM}$), as well as PS models trained from scratch with one (PS$_{noBM}$) and with 5 MRIs (PS$^{F4}_{BM}$) of a given patient. BMs, baseline models; HD, Hausdorff distance; MRI, magnetic resonance imaging; PS, patient-specific.

evaluation clearly favored the latter. PS$_{noBM}$ had only a quarter of clinically acceptable predictions and generated the highest percentage of unusable contours. Despite their relatively good overlap with the ground truth, the irregular borders of the PS$_{noBM}$ predictions would still require tedious adjustments. In contrast, BM delineations had smoother borders and easier-to-correct errors, for example, misclassified volume of surrounding tissue that could be quickly deleted.

For both PS$^{F4}_{BM}$ and PS$^{F4}_{noBM}$ models, training with more patient images further enhanced the performance of PS$_{BM}$ and PS$_{noBM}$, respectively. This was especially

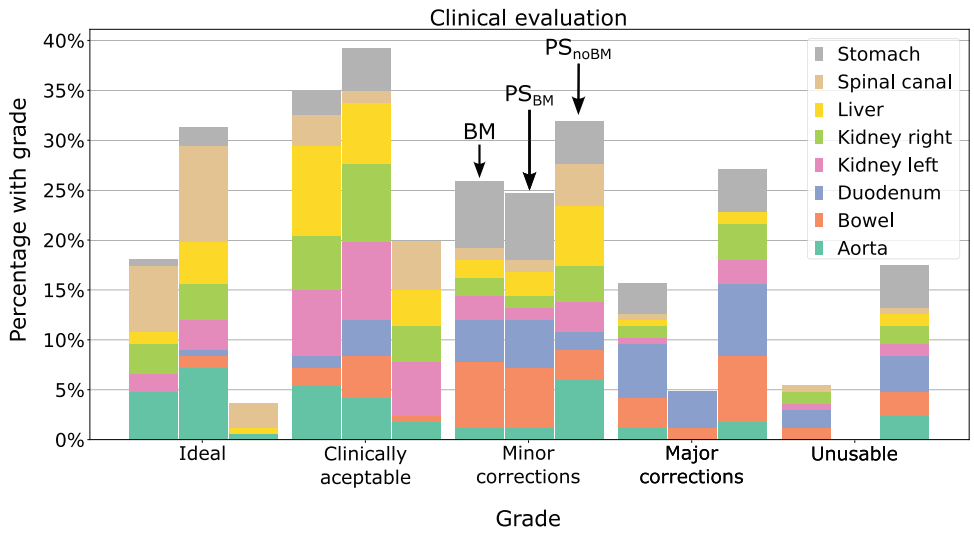

FIGURE 5 Bar plots displaying radiation oncologist's grading of predictions generated by the BMs, PS models fine-tuning the BM with the planning MRI (PS_{BM}) and models trained from scratch with the planning MRI (PS_{noBM}). The grades range from "ideal" to "unusable". BMs, baseline models; MRI, magnetic resonance imaging; PS, patient-specific.

notable for models trained from scratch. More images not only improved DSC and HDs but also resulted in contours with smoother borders. The trend of improving PS models with more patient images (up to five fraction images) was also observed in Li et al.'s study.[15]

The BMs yielded satisfactory results for most OARs, comparable to prior studies. Our DSC values for kidneys, liver, and stomach were in agreement with Fu et al.'s work,[8] but their results for bowel and duodenum surpassed ours. However, in our study we used clinical contours directly, whereas in Fu et al.'s study, the contours were refined by multiple trained professionals using dedicated software for accurate contour corrections. In comparison to Liang et al.'s study,[7] we achieved higher DSCs for kidneys but a lower one for the liver. Li et al.'s[15] results for PS models were better than ours, achieving DSCs above 0.9 for most OARs. However, their evaluation's robustness is limited by testing on only six patients. Despite all studies focused on MRIs from MR-Linac treatments, the testing cohorts differed, introducing limitations to the comparison. Additionally, institutional guidelines, contouring styles, and the effort put into correcting the fraction contours might have influenced the ground truth quality. Therefore, the presented comparison with other studies should be taken with caution.

In this study, 2D U-Nets were explored as potential candidates for the BMs and PS_{BM} models, just as 3D U-Nets were explored as candidates for PS_{noBM} models. However, the 3D BM and PS_{BM} models showed higher DSC than their 2D counterparts, while 2D PS_{noBM} networks outperformed their 3D PS_{noBM} counterparts. Consequently, 3D architectures were selected as the final models for testing in the case of BM, PS_{BM}, and PS_{BM}, whereas 2D architectures were chosen for PS_{noBM} and PS^{F4}_{noBM}. The superior performance of the 2D PS_{noBM} models can likely be attributed to the lower complexity of 2D models compared to their 3D counterparts in scenarios with limited data. Using 2D data effectively increases the size of the training set, as each axial slice is treated as an independent image. Selecting 2D instead of 3D networks while training with little data has also been done in the work of Fransson et al.[14] and Li et al.[15]

In addition to the U-Net architecture developed by Kerfoot et al.,[21] which was used throughout this work, a preliminary study was conducted using the nnUnet-v2 model.[28] The duodenum and aorta were included in this exploratory study, both benefiting considerably from PS training. The analysis based on the self-configuring single-label 3D nnUnet yielded the same conclusions: PS training enhances BM performance, and adding more images to the training set further improves outcomes. This indicates that the PS training strategies proposed in this study benefit not only a "conventional" U-Net but also the state-of-the-art nnUNet. Nevertheless, more studies are necessary to explore the advantages of this approach fully.

The predictions of models investigated in this work were also compared to predictions obtained by TotalSegmentator MRI.[29,30] The latter is a ready-to-use nnUNet that has been trained on a wide range of diagnostic MRIs from different scanners, institutions, and protocols and is therefore expected to perform well on most MR images. However, the resulting contours were worse than the predictions of all models in this study. This was partially related to differences in contouring styles, but could also be attributed to the

application to an unseen MRI domain. This indicates that the TotalSegmentator MRI might require additional investigation in the future in the scope of 0.35 T MRIs from the investigated MR-Linac, also regarding potential PS training schemes.

Not all OARs benefited equally from PS fine-tuning or more patient images. In fact, they fell into three categories. The first included kidneys and liver, which have rather stable shapes and clear boundaries. For them, PS_{BM} and PS_{BM}^{F4} models corrected larger BM misclassifications, but the overall improvements were moderate. The second group comprised the aorta, spinal canal, and bowel. Due to their large vertical extent, radiation oncologists segment only axial slices around the PTV. PS training encoded information on the superior-inferior segmentation ends into each personalized model, resulting in a higher geometric agreement between the predictions and the clinical ground truth. The third subgroup included the stomach, duodenum, and again bowel, organs prone to large volume changes during the course of treatment. Notably, PS improvements were the most pronounced in this group.

In this study, we concentrated on the abdominal OARs of patients treated with MRgRT. The abdomen is known for its complexity in auto-segmentation, making it an ideal evaluation scenario for the methods under investigation. Nevertheless, there are no conceptual limitations to employing these methods for other anatomical sites. Moreover, there are no constraints restricting their use to MRgRT. They might also be employed for other fractionated treatments, for example, in the scope of cone beam computed tomography (CBCT)-guided adaptive radiotherapy.[31]

Training of a single PS model requires less than 1 h. This is sufficiently short to be performed before the first irradiation as well as between fractions. Although preparing PS models for individual patients demands clearly more effort than training population BMs, it reduces the need for manual corrections in a critical moment, while patients are already in treatment position.

The study has its limitations. Firstly, in clinical practice, contouring radiation oncologists may not adjust OARs located further away from the PTV, considering their minimal impact on dose calculation. Consequently, while these clinical ground truth contours are sufficiently accurate for treatment adaptation, they might be suboptimal for network development. Secondly, all oncologists that generated the ground truth contours belonged to one institution. However, our previous study on OAR auto-segmentation for prostate cancer patients[12] showed no significant differences between cohorts from different institutions, suggesting the generalizability of the presented methods. Thirdly, involving multiple radiation oncologists would enhance the robustness of the clinical evaluation. Finally, the time saved for manual corrections by using PS contours instead of the population ones has not been measured. However, the better score in the clinical evaluation suggests considerable time-saving.

Future work directions include exploring PS transfer learning by freezing the BM parameters and adding new trainable layers.[32] Another alternative might be to use the segmented planning MRI as an additional input to help auto-segmentation of fraction images.[33]

5 | CONCLUSIONS

In this study, we demonstrated the advantages of personalized segmentation models in fractionated MRgRT of the abdomen region compared to conventional BMs. Particularly, PS models generated through fine-tuning the BM with patient data performed the best, while training from scratch performed worse than the BMs. Moreover, progressive fine-tuning of the personalized models with new segmented fraction MRIs was shown to further enhance the performance of the models. Physician assessment showed that fine-tuning BMs with only the planning MRI generates delineations that, in most cases, can be used directly for plan adaptation, and only a few require major corrections.

ACKNOWLEDGMENTS

The first author wishes to thank Martin Rädler for proofreading the manuscript and discussions throughout the study. This work was funded by the Wilhelm Sander-Stiftung (2019.162.2).

CONFLICT OF INTEREST STATEMENT

The Department of Radiation Oncology of the University Hospital of LMU Munich has research agreements with Elekta and Brainlab.

REFERENCES

1. Tijssen RH, Philippens ME, Paulson ES, et al. MRI commissioning of 1.5 T MR-linac systems–a multi-institutional study. *Radiother Oncol*. 2019;132:114-120.
2. Klüter S. Technical design and concept of a 0.35 T MR-Linac. *Clin Transl Oncol*. 2019;18:98-101.
3. Winkel D, Bol GH, Kroon PS, et al. Adaptive radiotherapy: the Elekta Unity MR-linac concept. *Clin Transl Oncol*. 2019;18:54-59.
4. Henke L, Contreras JA, Green OL, et al. Magnetic resonance image-guided radiotherapy (MRIgRT): a 4.5-year clinical experience. *Clin Oncol*. 2018;30:720-727.
5. Corradini S, Alongi F, Andratschke N, et al.MR-guidance in clinical reality: current treatment challenges and future perspectives. *Radiat Oncol*. 2019:14:1-12.
6. Nachbar M, lo Russo M, Gani C, et al. Automatic AI-based contouring of prostate MRI for online adaptive radiotherapy. *Z Med Phys*. 2023.
7. Liang F, Qian P, Su K-H, 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.
8. Fu Y, Mazur TR, Wu X, et al. A novel MRI segmentation method using CNN-based correction network for MRI-guided adaptive radiotherapy. *Med Phys*. 2018;45:5129-5137.

9. Ribeiro MF, Marschner S, Kawula M, et al. Deep learning based automatic segmentation of organs-at-risk for 0.35 T MRgRT of lung tumors. *Radiat Oncol*. 2023;18:135.
10. Harrison K, Pullen H, Welsh C, Oktay O, Alvarez-Valle J, Jena R. Machine learning for auto-segmentation in radiotherapy planning. *Clin Oncol*. 2022;34:74-88.
11. Hunt A, Hansen V, Oelfke U, Nill S, Hafeez S. Adaptive radiotherapy enabled by MRI guidance. *Clin Oncol*. 2018;30:711-719.
12. Kawula M, Hadi I, Nierer L, et al.Patient-specific transfer learning for auto-segmentation in adaptive 0.35 T MRgRT of prostate cancer: a bi-centric evaluation. *Med Phys*. 2023;50:1573-1585.
13. Chen X, Ma X, Yan X, et al.Personalized auto-segmentation for magnetic resonance imaging–guided adaptive radiotherapy of prostate cancer. *Med Phys*. 2022:49:4971-4979.
14. Fransson S, Tilly D, Strand R. Patient specific deep learning based segmentation for magnetic resonance guided prostate radiotherapy. *Phys Imaging Radiat Oncol*. 2022;23:38-42.
15. Li Z, Zhang W, Li B, et al. Patient-specific daily updated deep learning auto-segmentation for MRI-guided adaptive radiotherapy. *Radiother Oncol*. 2022;177:222-230.
16. Sharp GC, Li R, Wolfgang J, et al. Plastimatch: an open source software suite for radiotherapy image processing. In: Proceedings of the XVI'th International Conference on the use of Computers in Radiotherapy (ICCR). IEEE; 2010.
17. Teoh S, Ooms A, George B, et al. Evaluation of hypofractionated adaptive radiotherapy using the MR Linac in localised pancreatic cancer: protocol summary of the Emerald-Pancreas phase 1/expansion study located at Oxford University Hospital, UK. *BMJ Open*. 2023;13:e068906.
18. Chuong MD, Bryant J, Mittauer KE, et al. Ablative 5-fraction stereotactic magnetic resonance–guided radiation therapy with on-table adaptive replanning and elective nodal irradiation for inoperable pancreas cancer. *Pract Radiat Oncol*. 2021;11:134-147.
19. Haas YP, Ludwig R, Dal Bello R, Tanadini-Lang S, Unkelbach J. Adaptive fractionation at the MR-linac. *Phys Med Biol*. 2023;68:035003.
20. Cardoso MJ, et al. MONAI: An open-source framework for deep learning in healthcare. arXiv preprint arXiv:2211.02701 2022.
21. Kerfoot E, Clough J, Oksuz I, Lee J, King AP, Schnabel JA. Left-ventricle quantification using residual U-Net. In: *International Workshop on Statistical Atlases and Computational Models of the Heart*. Springer; 2018:371-380.
22. Ulyanov D, Vedaldi A, Lempitsky V. Instance Normalization: The Missing Ingredient for Fast Stylization. *arXiv preprint arXiv:1607.08022* 2016.
23. Ding B, Qian H, Zhou J. Activation functions and their characteristics in deep neural networks. In: *2018 Chinese control and decision conference (CCDC)*. IEEE; 2018:1836–1841.
24. Bridle JS. Training stochastic model recognition algorithms as networks can lead to maximum mutual information estimation of parameters. In: Proceedings of the 2nd International Conference on Neural Information Processing Systems. MIT Press; 1990:211–217.
25. Milletari F, Navab N, Ahmadi S-A. V-net: fully convolutional neural networks for volumetric medical image segmentation. In: *2016 fourth international conference on 3D vision (3DV)*. IEEE; 2016:565-571.
26. Friedman M. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *J Stat Data Sci Educ*. 1937;32:675-701.
27. Nemenyi PB. *Distribution-free multiple comparisons*. Princeton University, 1963.
28. 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;18:203-211.
29. Wasserthal J, Breit H-C, Meyer MT, et al. TotalSegmentator: robust segmentation of 104 anatomic structures in CT images. *Radiology: Artificial Intelligence*. 2023;5:e230235.
30. D'Antonoli TA, Berger LK, Indrakanti AK, et al. TotalSegmentator MRI: Sequence-Independent Segmentation of 59 Anatomical Structures in MR images. 2024.
31. Schiff JP, Stowe HB, Price A, et al. In silico trial of computed tomography-guided stereotactic adaptive radiation therapy (CT-STAR) for the treatment of abdominal oligometastases. *Int J Radiat Oncol Biol Phys*. 2022;114:1022-1031.
32. Talo M, Baloglu UB, Yıldırım Ö, Acharya UR. Application of deep transfer learning for automated brain abnormality classification using MR images. *Cogn Syst Res*. 2019;54:176-188.
33. Klymenko T, Kim ST, Lauber K, et al. Butterfly-Net: spatial-temporal architecture for medical image segmentation. In: *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*. IEEE; 2021:616–620.
34. Pérez-García F, Sparks R, Ourselin S. TorchIO: a Python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning. *Comput Methods Programs Biomed*. 2021:106236.
35. Kawula M, Purice D, Li M, et al. Dosimetric impact of deep learning-based CT auto-segmentation on radiation therapy treatment planning for prostate cancer. *Radiat Oncol*. 2022;17:1-12.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Kawula M, Marschner S, Wei C, et al. Personalized deep learning auto-segmentation models for adaptive fractionated magnetic resonance-guided radiation therapy of the abdomen. *Med Phys*. 2024;1-10. https://doi.org/10.1002/mp.17580

Supplementary material

Patient anatomy over the treatment days

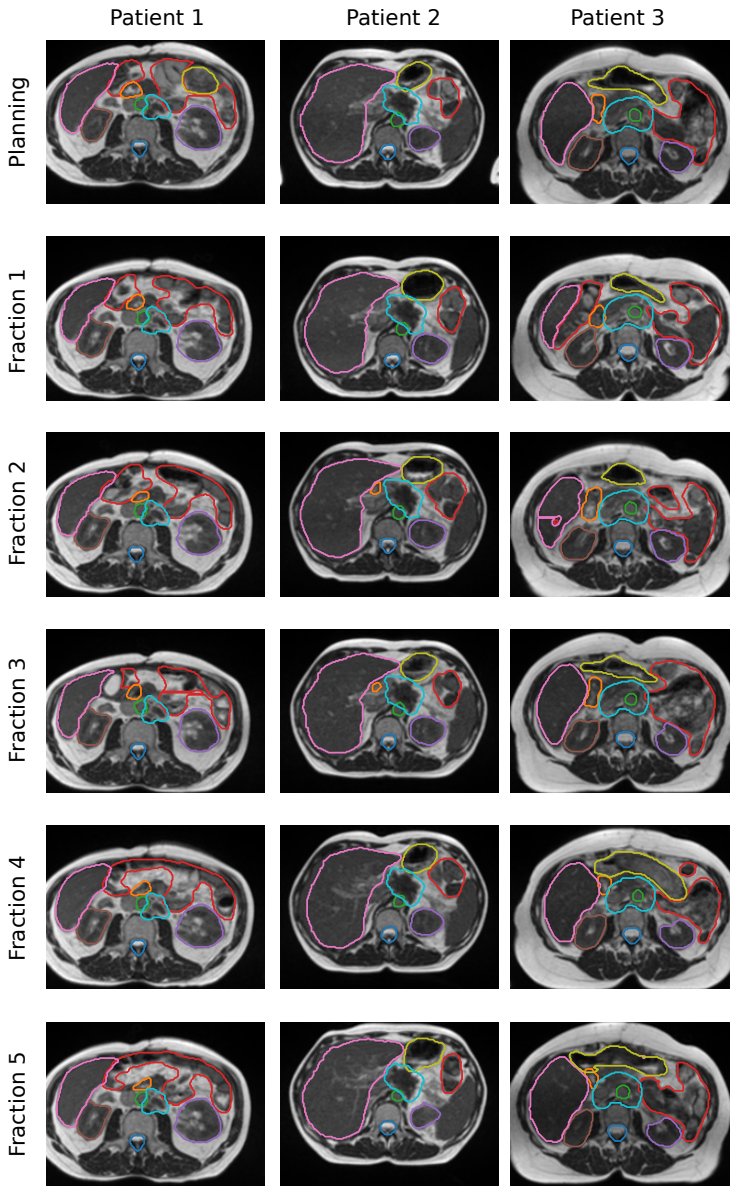

Figure S1: Axial view of three exemplary test patients showing the clinical ground truth contours over the course of treatment. The following organs and structures are shown: (pink) liver, (red) bowel, (yellow) stomach, (orange) duodenum, (green) aorta, (brown) right kidney, (purple) left kidney, (dark blue) spinal canal, and (cyan) planning target volume.

Data augmentation

In this section, we will describe the data augmentation scheme used for training of the BMs and all investigated PS models.

The on-the-fly data augmentation applied during all trainings included spatial and intensity based transformations. The spatial deformations consisted of affine transformations (with maximum rotation angle α_{max} and maximum shift Δ_{max}), zooming (with zoom and cropping ratio z_{min}, z_{max}), and B-Spline elastic deformation (with number of control points n_{cp} and maximum displacement d). The intensity-based augmentation included MR-related motion artifacts (with maximum translation m_{Δ}, rotation angle m_{α}), random blur (with standard deviation σ_{blur} of the Gaussian kernels used to blur the image), and random noise (with mean μ, standard deviation σ of a normal distribution for noise sampling). The intensity-based augmentations, along with B-spline elastic deformations, were imported from TorchIO[34], while other functions were from the MONAI library[20].

The choice of augmentation functions and initial parameter values were based on our previous studies[9,12,35]. Spatial augmentation, especially affine and elastic deformation, was found to be the most impactful strategy. Therefore, we performed a hyperparameter search for these two only. Since the impact of the remaining functions was only minor, we adopted values from our prior research without further optimization. For affine transformation we explored α_{max} within the range of 10 to 20 degrees, and Δ_{max} from 15 to 30 mm. For elastic deformation with B-Splines, we kept the number of control points constant, but varied the maximum displacement values d from 15 to 60 mm. Table S1 summarizes the information on the applied data augmentation.

In this work it was important to find hyperparameters giving optimal performance while keeping training times reasonably low for daily re-training of personalized models. Whereas BMs require only one training, PS_{BM} and PS_{noBM} models need to be created individually for each patient and $\text{PS}^{\text{F4}}_{\text{BM}}$ and $\text{PS}^{\text{F4}}_{\text{noBM}}$ require daily fine-tuning. Currently, data augmentation is the most time-consuming step in the training pipeline, with MR-specific transforms being notably the slowest. For this reason, we trained BMs with the full augmentation, but for PS models we reduced it to spatial deformations and one intensity-base transformation (random noise).

Table S1: Data augmentation parameters and final values used for training.

Function	Parameter	Final value				
		BM	PS_{BM}	PS^{F4}_{BM}	PS_{noBM}	PS^{F4}_{noBM}
Probability	p_{agm}	0.85	0.95	0.95	0.95	0.95
Rotation	α_{max}	10°	10°	10°	20°	20°
Translation	Δ_{max}	15 mm	15 mm	15 mm	30 mm	30 mm
Deformation	n_{cp}	10	10	10	10	10
	d	15 mm	30 mm	30 mm	45 mm	45 mm
Zooming	z_{min}, z_{max}	0.9, 1.1	0.9, 1.1	0.9, 1.1	0.9, 1.1	0.9, 1.1
Motion	m_α	20°	-	-	-	-
	m_Δ	75 mm	-	-	-	-
Random blur	σ_{blur}	0.6	-	-	-	-
Noise	σ	0.25	0.25	0.25	0.25	0.25
	μ	0	0	0	0	0

Test of statistical significance

Table S2: P-values obtained from the post-hoc Nemenyi test for the testing set for all possible pairwise model comparisons. Significant p-values (< 0.05) are denoted with an asterisk.

Method 1	Method 2	Aorta	Bowel	Duodenum	Kidney L.	Kidney R.	Liver	Spinal Canal	Stomach
BM	PS_{BM}	0.001*	0.001*	0.001*	0.001*	0.001*	0.001*	0.001*	0.001*
BM	PS_{BM}^{F4}	0.001*	0.001*	0.001*	0.001*	0.001*	0.001*	0.001*	0.001*
BM	PS_{noBM}	0.7	0.9	0.45	0.01*	0.08	0.9	0.001*	0.001*
BM	PS_{noBM}^{F4}	0.001*	0.62	0.14	0.007*	0.06	0.57	0.15	0.9
PS_{BM}	PS_{BM}^{F4}	0.06	0.001*	0.001*	0.001*	0.09	0.001*	0.01*	0.001*
PS_{BM}	PS_{noBM}	0.001*	0.001*	0.001*	0.08	0.001*	0.001*	0.9	0.001*
PS_{BM}	PS_{noBM}^{F4}	0.12	0.002*	0.02*	0.12	0.001*	0.001*	0.1	0.001*
PS_{BM}^{F4}	PS_{noBM}	0.001*	0.001*	0.001*	0.001*	0.001*	0.001*	0.001*	0.001*
PS_{BM}^{F4}	PS_{noBM}^{F4}	0.001*	0.001*	0.001*	0.001*	0.001*	0.001*	0.001*	0.001*
PS_{noBM}	PS_{noBM}^{F4}	0.002*	0.42	0.001*	0.9	0.9	0.84	0.5	0.06

Chapter 6

Conclusion

MRI guidance and daily treatment adaptation have transformed how fractionated radiation therapy is administered. Already for several years, it has been possible to perform MRgRT using MR-LINACs. They allow daily patient imaging to create up-to-date treatment plans. Additionally, fast imaging during irradiation enables gated treatment. The combination of gating and the possibility of daily plan adaptation gives the confidence to reduce the safety margins. This results in less normal tissue being irradiated, leading to lower toxicity and fewer side effects. Simultaneously, higher daily doses can be delivered to the tumor in fewer fractions, which shortens the total treatment time.
Every optimization and re-optimization of a treatment plan requires delineation of target volumes and OARs, naturally raising interest in auto-segmentation. The most promising techniques developed recently are based on DL. An underlying assumption in DL is that neural networks must perform well on new and diverse examples. However, this might not hold true for auto-segmentation at MR-LINACs. Fraction images do not have to be considered entirely new and random, as segmented planning MRIs exist for each patient. In fact, every fraction image is highly similar to the planning and previous fraction MRIs. Therefore, teaching personalized models features that are specific to a given patient but not necessarily shared among the general population might result in better performance compared to population models.
The primary goal of this dissertation was to address this particular aspect of auto-segmentation in fractionated treatment, namely leveraging the existence of previously segmented images at the time of fraction image segmentation. Four main projects were carried out in the framework of this thesis. The first (introductory) project was carried out for radiotherapy at conventional LINACs, while the remaining three studies focused entirely on MRgRT at the MRIdian MR-LINACs. These three studies were specifically designed for fractionated treatments. They aimed to create models optimized for repeatedly segmenting the same patient rather than optimizing average performance over a set of unknown patients. In particular, the studies aimed to use prior knowledge present in MRgRT to enhance segmentation on fraction days.

The study described in Paper 1 was one of the first to analyze the impact of DL contours on dose optimization during treatment planning. It considered prostate cancer patients undergoing radiotherapy at conventional LINACs. Additionally, it investi-

gated whether there is a positive correlation between the geometric contour quality and the quality of treatment plans optimized using these DL contours. The study included commonly used geometric metrics and dosimetric parameters employed in clinics on a daily basis.
Auto-segmentation networks trained for this study achieved performance comparable to other state-of-the-art studies [128–130]. Utilizing DL contours in treatment plan optimization led to dose distributions that did not overdose the neighboring OARs, provided sufficient target volume coverage in all but one case, and did not create any undesirable hot or cold dose spots. This led to the conclusion that although human review is still required, implementing DL contours has the potential to speed up treatment planning by eliminating the need for fully manual contouring without compromising the plan quality.
The only moderately positive correlation was found between the prostate's DSC and gamma index. It was observed that despite some contours showing the same geometric performance, the quality of dose distributions optimized using these contours differed. While the location of misclassified pixels does not change the DSC and has only a limited impact on Hausdorff distance, it influences dose optimization. This observation demonstrated the importance of contour evaluation beyond geometric metrics.

The research described in Paper 2 investigated three types of auto-segmentation models: first, conventional population models that were trained to perform well on an average patient; second, models adjusted to patients from a specific facility; third, patient-specific models optimized to perform well for a patient seen multiple times. It was the first work that considered patient-specific models for prostate cancer patients treated at the MRIdian MR-LINACs. Moreover, it was the first study to compare DL-generated contours with the current solution implemented at the MRIdian TPS.
The study demonstrated that population models had geometric performance comparable to similar studies in the literature [116]. A radiation oncologist graded the OAR contours predicted by the population models as more useful for treatment adaptation than the structures suggested by the TPS using DIR. The study showed the benefit of facility-specific transfer learning for prostate CTVs, which could account for differences in institutional delineation guidelines between the LMU and Gemelli Hospital. There was no benefit of facility-specific transfer learning for OARs. This suggested the transferability of OAR models between different institutions. Finally, the personalized fine-tuning was shown to be advantageous for all investigated structures, which aligned with other studies published independently around the same time [115, 116]. Patients with uncommon anatomies benefited the most from patient-specific transfer learning. Moreover, personalized models could account for the fact that CTV segmentation did not always follow any visible organ boundaries. In fact, it was frequently defined based on factors that could not be captured by imaging (e.g., blood test outcomes and prior medical history). The weakness of personalized models was their lack of robustness to larger changes with respect to the planning day. Additionally, personalized models could propagate errors to the fraction images if the planning segmentation was inaccurate.

The study described in Paper 3 analyzed networks combining DIR and segmenta-

tion for prostate cancer patients. These networks were trained to register the planning and fraction MRIs and deform planning contours to the daily anatomy. Unlike the personalized models developed in Paper 2, the registration approach requires only one model to be trained for all patients. Patient-specific fine-tuning is unnecessary since the planning contours are an additional network input. This study was the first to combine image registration and contour propagation for OARs.
The study showed that registration networks could model only small anatomical changes regardless of the training scheme. The presumed reason for predicting only limited deformations was the large dimensionality of the expected output in relation to the training data set size. Registration networks were found unsuitable for organs that undergo substantial changes, such as the bladder and rectum. However, these networks showed potential for prostate CTV segmentation, as the latter should not change substantially to ensure proper tumor dose coverage throughout the treatment.

The fourth study explored alternatives for training personalized auto-segmentation models with and without the population baseline models. One new strategy involved continuous (also called progressive) training, adjusting models to the latest patient anatomy after every fraction. Another investigated method examined whether a single training example is sufficient for creating personalized auto-segmentation models from scratch. The target group comprised patients diagnosed with lesions in the abdomen, which is a region that includes a wide variety of OARs and is subject to pronounced inter-fractional anatomical changes. Thanks to this variety, some conclusions from this study could likely be extrapolated to other anatomical sites. This was the first work investigating the impact of DL population models on personalized models and their integration with progressive training. Moreover, it offered a comparison with methods relying solely on individual patient data.
The study concluded that personalized models developed by fine-tuning population models with a given patient's data predicted contours that could, in most cases, be used clinically without additional corrections. Conversely, models trained with a single image generated contours that often required substantial manual corrections before being useful for treatment adaptation. Patient-specific networks profited from progressive learning in both scenarios, with and without baseline models. Personalized training was particularly beneficial for organs commonly considered challenging to segment (e.g., duodenum) and elongated organs that are only segmented in the high-dose region (e.g., aorta and spinal canal). Improvements with respect to population models in organs commonly considered easy to segment (e.g., kidneys and liver) were small and not always statistically significant.

The presented studies have limitations. First, the quality of clinical data used as ground truth throughout this work might not have been optimal for training auto-segmentation networks. The clinical contours were prepared with the assumption of being sufficiently accurate for dose optimization. Therefore, OAR parts in the low-dose regions typically receive less attention, and the main focus is on the organ parts within 3 cm around the PTV. This problem is more pronounced for the fraction than planning contours, as the former are always created under time pressure when a patient waits for the irradiation in treatment position. While small segmentation inconsistencies,

especially outside of the high-dose region, do not affect the dose calculation, they can considerably decrease DSC and HD during testing and validation.
Another limitation was the lack of measuring the exact time saved when employing the DL contours. Measuring correction times was not feasible due to the high clinical workload of the radiation oncologists in combination with the high number of investigated methods and the dataset size. Instead, a surrogate quantity was used, i.e., a four- or five-point scale indicating how many corrections the grading oncologist would have to apply to make the contour clinically usable.
Another constraint of this study was that it did not test the methods on public datasets, allowing for fair comparison with other studies. To the best of the author's knowledge at the time of writing, there were no publicly available datasets of annotated MRIdian images for any entity, nor any longitudinal sets of planning and fraction images. The only external dataset was of prostate cancer patients and acquired at the Gemelli Hospital.

Several future studies are planned to further investigate the topic of auto-segmentation for fractionated MRgRT. One project involves using the segmented planning MRI as an additional network input when generating predictions. U-Nets used for that purpose have one encoding arm for the image to segment (here, fraction image) and an additional arm to encode the planning information. The information from both arms is combined at the bottleneck and fed into the common decoder arm. The advantage of this method is that it uses the planning MRI at the time of segmenting fraction images, eliminating the need for patient-specific fine-tuning.
Another possible direction for future studies is to replace convolutional networks with transformer architectures for combined image registration and segmentation. A limitation of convolutional networks is their inability to model long-distance relationships between image voxels. Transformer architectures with a built-in attention mechanism, e.g., TransMorph [122], have the potential to overcome this limitation. This project would require collaborations with institutions operating the same MR-LINAC or using pre-trained models since transformers need large training sets of hundreds of images.
Another aspect worth investigating is time-saving measurements: the potential time saved by replacing manual planning segmentation with population models and replacing the TPS fraction contours with the personalized models. It would be ideal to conduct these experiments prospectively, where each new patient is randomly assigned to either the DL or the non-DL group. This approach ensures that the perception of necessary corrections is based on using the contours for actual treatment rather than just for scientific purposes. In fact, a similar time-saving experiment for lung cancer patients is currently carried out at the LMU University Hospital. The development of networks used for these time-saving measurements [131] was based on the implementation of the prostate models described in the second paper of this dissertation. Preliminary results suggest that using DL contours leads to time savings, but more statistics is necessary.
Possible future projects involve utilizing Monte Carlo dropout to estimate segmentation uncertainty. This method generates several predictions for the same input image, each with a random subset of network connections removed. Subsequently, a prediction frequency map is generated to illustrate which voxels the network classified

consistently and for which it showed variations. This map would highlight regions requiring more attention from the contouring physician.
An evaluation of contour quality at the dose level could also be carried out. For this purpose, the methodology used in the first paper for conventional LINACs could be implemented for the MRgRT at the MRIdian MR-LINAC.

In summary, this work focused on prior-knowledge aware DL auto-segmentation methods for fractionated MRgRT at the MRIdian MR-LINAC. Personalized models created by fine-tuning population models with segmented planning (or prior fraction) images performed best among investigated alternatives. Training times of personalized models were short enough to be implemented clinically. Therefore, personalized models can potentially shorten treatment adaptation in MRgRT at the MR-LINACs. A shorter adaptation translates to less time patients spend on the treatment couch, increasing patient compliance and comfort while decreasing the likelihood of further anatomical changes. Moreover, this would enhance the MR-LINAC throughput, allowing more patients to benefit from MRgRT.

Acknowledgements

Completing this thesis has been a challenging yet rewarding journey, and I would not have made it without the support of many people. I am grateful to everyone who contributed to this work, whether through scientific guidance, collaboration, or personal encouragement.

First and foremost, I would like to thank Dr. Christopher Kurz for supervising and mentoring me. I am deeply grateful for your scientific guidance, insightful discussions, curiosity, encouragement, never losing your patience, and meticulous reading of my manuscripts. I truly appreciate that you always made time for me— whether it was for lengthy revisions, the fundamentals of MRI, or simply the next steps in my research. Beyond being a great supervisor, I am also grateful for your involvement outside of work. From celebrating together to sharing good times, your support and kindness have meant a lot to me. It was truly a pleasure to have you as my supervisor.

I would like to sincerely thank Prof. Guillaume Landry, the head of the LMU Adaptive Radiation Therapy Lab, for your genuine involvement in my work and for your valuable and critical feedback on my projects. I also appreciate your efforts in shaping our group, building external collaborations, and encouraging us to participate in national and international conferences. Thank you for organizing both formal and informal gatherings — all of which have helped strengthen our team. It has been a pleasure to be part of a group led by you.

I consider myself very fortunate to have conducted my PhD projects at the Radiation Oncology Department of LMU University Hospital. Being near patient oncological treatment, observing the work of physicians and medical physicists, and always having the opportunity to ask questions that drive our research has been invaluable. I am deeply grateful to Prof. Claus Belka and Prof. Stefanie Corradini for creating such an inspiring environment. I also want to thank Dr. Sebastian Marschner and Dr. Indrawati Hadi for their support of my research at the MR-Linac. Many thanks to the medical physics team, especially Dr. Jan Hofmaier, Dr. Vanessa da Silva Mendes, Tobias Winderl, Lukas Nierer, and Catrin Rodenberg, for your valuable insights on MR-Linac operation and treatment planning.

Special thanks to our collaborators from Gemelli University Hospital in Rome, with whom I had the pleasure of conducting several studies. I truly appreciate the contributions of Dr. Davide Cusumano, Dr. Luca Boldrini, Dr. Lorenzo Placidi, and Marica Vagni — thank you!

I would like to thank our HiWis, Antonio and Samira, for their many hours of meticulous work in exporting data for my research. Your efforts are truly appreciated!

I would like to express my gratitude to the Wilhelm Sander Stiftung for funding my

work. Thank you very much for your trust and support, which enabled me to carry out this research. I hope that this project contributes a valuable step in the right direction for cancer treatment.
Over these three years, I was fortunate to be surrounded by an amazing group of people who made this time truly special. Ivy, thank you for being both a great friend and my sport buddy. Jackie, I appreciate your cool-headed advice whenever I was freaking out. Lili, thank you for your honesty. Nikos, I'm grateful for your help with IT-related issues. Adrian, thanks for being such a welcoming host for our after-work gatherings. Elia, I always enjoyed our interesting lunchtime conversations. Chengtao, thank you for your help with the revision of my last manuscript. Henning, it was fun to go to conferences with you. And Moritz, I truly benefited from your experience: from selecting the best color maps for plots to dissertation writing. Thank you all!
On a personal note, I would like to thank my family in Poland and Germany for their love, support, and unwavering faith in me. Thanks to my siblings, parents, and friends for always being there. Special thanks to Christine for proofreading parts of this dissertation. An equally special thanks to Maciej for reading the entire document and ensuring it sounds more like a scientific work rather than a novel about people strolling through Krakowskie Przedmieście.
Last but not least, I would like to thank Martin. Thank you for your love and for being a steady source of support. I deeply appreciate being able to discuss every detail of my work with you, feeling understood, and receiving your valuable feedback. Thank you for proofreading this document and for your help in creating high-quality vector graphics for this dissertation. I am deeply grateful to be sharing these busy, yet exciting and meaningful years with you.

Bibliography

[1] Rebecca L Siegel, Kimberly D Miller, Nikita Sandeep Wagle, and Ahmedin Jemal. Cancer statistics, 2023. *CA: A Cancer Journal for Clinicians*, 73(1):17–48, 2023.

[2] WHO GCO. Cancer Today. `https://gco.iarc.fr/today/en/dataviz/bars`. Accessed: 06.03.2024.

[3] Cerise M Siamof, Shreya Goel, and Weibo Cai. Moving beyond the pillars of cancer treatment: perspectives from nanotechnology. *Frontiers in Chemistry*, 8:598100, 2020.

[4] Deborah E Citrin. Recent developments in radiotherapy. *New England Journal of Medicine*, 377(11):1065–1075, 2017.

[5] Konrad Mohnike, Jens Ricke, and Stefanie Corradini. *Manual on Image-Guided Brachytherapy of Inner Organs: Technique, Indications and Evidence*. Springer, 2021.

[6] Charles M Washington and Dennis T Leaver. *Principles and practice of radiation therapy-e-book*. Elsevier Health Sciences, 2015.

[7] Thomas Bortfeld. IMRT: a review and preview. *Physics in Medicine & Biology*, 51(13):R363, 2006.

[8] Marcel Van Herk. Errors and margins in radiotherapy. In *Seminars in radiation oncology*, volume 14, pages 52–64. Elsevier, 2004.

[9] Sebastian Klüter. Technical design and concept of a 0.35 T MR-Linac. *Clinical and Translational Radiation Oncology*, 18:98–101, 2019.

[10] Dennis Winkel, Gijsbert H Bol, Petra S Kroon, Bram van Asselen, Sara S Hackett, Anita M Werensteijn-Honingh, Martijn PW Intven, Wietse SC Eppinga, Rob HN Tijssen, Linda GW Kerkmeijer, et al. Adaptive radiotherapy: the Elekta Unity MR-linac concept. *Clinical and Translational Radiation Oncology*, 18:54–59, 2019.

[11] Xin Liu, Zhenjiang Li, and Yong Yin. Clinical application of MR-Linac in tumor radiotherapy: a systematic review. *Radiation Oncology*, 18(1):52, 2023.

[12] Francesco Cuccia, Stefanie Corradini, Rosario Mazzola, Luigi Spiazzi, Michele Rigo, Marco Lorenzo Bonù, Ruggero Ruggieri, Michela Buglione di Monale e Bastia, Stefano Maria Magrini, and Filippo Alongi. MR-guided hypofractionated radiotherapy: current emerging data and promising perspectives for localized prostate cancer. *Cancers*, 13(8):1791, 2021.

[13] Görkem Güngör, İlkay Serbez, Bilgehan Temur, Gökhan Gür, Namık Kayalılar, Teuta Zoto Mustafayev, Latif Korkmaz, Gökhan Aydın, Bülent Yapıcı, Banu Atalar, et al. Time analysis of online adaptive magnetic resonance–guided radiation therapy workflow according to anatomical sites. *Practical Radiation Oncology*, 11(1):e11–e21, 2021.

[14] Alain Jungo, Raphael Meier, Ekin Ermis, Marcela Blatti-Moreno, Evelyn Herrmann, Roland Wiest, and Mauricio Reyes. On the effect of inter-observer variability for a reliable estimation of uncertainty of medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part I*, pages 682–690. Springer, 2018.

[15] Carlos E Cardenas, Jinzhong Yang, Brian M Anderson, Laurence E Court, and Kristy B Brock. Advances in auto-segmentation. In *Seminars in radiation oncology*, volume 29, pages 185–197. Elsevier, 2019.

[16] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25, 2012.

[17] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.

[18] Mohammad Hesam Hesamian, Wenjing Jia, Xiangjian He, and Paul Kennedy. Deep learning techniques for medical image segmentation: achievements and challenges. *Journal of Digital Imaging*, 32:582–596, 2019.

[19] Thorsten M Buzug. Computed tomography. In *Springer handbook of medical technology*, pages 311–342. Springer, 2011.

[20] Ervin B Podgoršak et al. *Radiation physics for medical physicists*, volume 1. Springer, 2006.

[21] Wolfgang Schlegel and Josef Bille. *Medizinische Physik*. Springer, 2018.

[22] Shivaramu. Effective atomic numbers for photon energy absorption and photon attenuation of tissues from human organs. *Medical Dosimetry*, 27(1):1–9, 2002.

[23] Glenn F Knoll. *Radiation detection and measurement*. John Wiley & Sons, 2010.

[24] Frank Herbert Attix. *Introduction to radiological physics and radiation dosimetry*. John Wiley & Sons, 2008.

[25] Brent Burbridge. *Undergraduate Diagnostic Imaging Fundamentals*. Distance Education Unit, University of Saskatchewan, 11 2017.

[26] Gavin Poludniowski, Guillaume Landry, François Deblois, PM Evans, and Frank Verhaegen. SpekCalc: a program to calculate photon spectra from tungsten anode x-ray tubes. *Physics in Medicine & Biology*, 54(19):N433, 2009.

[27] Harrison H Barrett and William Swindell. *Radiological imaging: the theory of image formation, detection, and processing*. Academic press, 1996.

[28] Moritz Rabe. *Investigation of time-resolved volumetric MRI to enhance MR-guided radiotherapy of moving lung tumors*. PhD thesis, lmu, 2022.

[29] Anne T Davis, Antony L Palmer, and Andrew Nisbet. Can CT scan protocols used for radiotherapy treatment planning be adjusted to optimize image quality and patient dose? A systematic review. *The British Journal of Radiology*, 90(1076):20160406, 2017.

[30] R.C. Gonzalez and R.E. Woods. *Digital Image Processing*. Prentice Hall, 2008.

[31] Robert W Brown, Y-C Norman Cheng, E Mark Haacke, Michael R Thompson, and Ramesh Venkatesan. *Magnetic resonance imaging: physical principles and sequence design*. John Wiley & Sons, 2014.

[32] Eite Tiesinga, Peter J Mohr, David B Newell, and Barry N Taylor. CODATA recommended values of the fundamental physical constants: 2018. *Journal of Physical and Chemical Reference Data*, 50(3), 2021.

[33] Joseph Larmor. LXIII. On the theory of the magnetic influence on spectra; and on the radiation from moving ions. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 44(271):503–512, 1897.

[34] A.D. Elster and J.H. Burdette. *Questions & Answers in Magnetic Resonance Imaging*. Mosby, 2001.

[35] William Rooney. MRI: from picture to proton, 2003.

[36] Felix Bloch. Nuclear induction. *Physical Review*, 70(7-8):460, 1946.

[37] Axel Haase. Snapshot FLASH MRI. Applications to T1, T2, and chemical-shift imaging. *Magnetic Resonance in Medicine*, 13(1):77–89, 1990.

[38] Matt A Bernstein, Kevin F King, and Xiaohong Joe Zhou. *Handbook of MRI pulse sequences*. Elsevier, 2004.

[39] John F Schenck. The role of magnetic susceptibility in magnetic resonance imaging: MRI magnetic compatibility of the first and second kinds. *Medical Physics*, 23(6):815–850, 1996.

[40] John S Ginn, Nzhde Agazaryan, Minsong Cao, Umar Baharom, Daniel A Low, Yingli Yang, Yu Gao, Peng Hu, Percy Lee, and James M Lamb. Characterization of spatial distortion in a 0.35 T MRI-guided radiotherapy system. *Physics in Medicine & Biology*, 62(11):4525, 2017.

[41] Kristy K Brock, Sasa Mutic, Todd R McNutt, Hua Li, and Marc L Kessler. Use of image registration and fusion algorithms and techniques in radiotherapy: Report of the AAPM Radiation Therapy Committee Task Group No. 132. *Medical Physics*, 44(7):e43–e76, 2017.

[42] J.V. Hajnal and D.L.G. Hill. *Medical Image Registration*. Biomedical Engineering. CRC Press, 2001.

[43] Kristy K Brock. *Image processing in radiation therapy*. CRC press, 2013.

[44] A. Frangi, J. Prince, and M. Sonka. *Medical Image Analysis*. The MICCAI Society book Series. Elsevier Science, 2023.

[45] John Ashburner and Karl J Friston. Rigid body registration. *Statistical Parametric Mapping: The Analysis of Functional Brain Images*, pages 49–62, 2007.

[46] Ruzena Bajcsy and Stane Kovačič. Multiresolution elastic matching. *Computer Vision, Graphics, and Image Processing*, 46(1):1–21, 1989.

[47] Gary E Christensen and Hans J Johnson. Consistent image registration. *IEEE Transactions on Medical Imaging*, 20(7):568–582, 2001.

[48] Magnus R Hestenes, Eduard Stiefel, et al. Methods of conjugate gradients for solving linear systems. *Journal of Research of the National Bureau of Standards*, 49(6):409–436, 1952.

[49] John E Dennis Jr and Robert B Schnabel. *Numerical methods for unconstrained optimization and nonlinear equations*. SIAM, 1996.

[50] John A Nelder and Roger Mead. A simplex method for function minimization. *The Computer Journal*, 7(4):308–313, 1965.

[51] A Joubert and N Foray. Intrinsic radiosensitivity and dna double-strand breaks in human cells. *Cancer Radiotherapie: Journal de la Societe Francaise de Radiotherapie Oncologique*, 11(3):129–142, 2007.

[52] H Rodney Withers. The four R's of radiotherapy. In *Advances in radiation biology*, volume 5, pages 241–271. Elsevier, 1975.

[53] Frank Pajonk, Erina Vlashi, and William H McBride. Radiation resistance of cancer stem cells: the 4 R's of radiobiology revisited. *Stem Cells*, 28(4):639–648, 2010.

[54] Gerd Muehllehner and Joel S Karp. Positron emission tomography. *Physics in Medicine & Biology*, 51(13):R117, 2006.

[55] International Commission on Radiation Units and Measurements. 4. Definition of Volumes. *Journal of the ICRU*, 10(1):41–53, 2010. PMID: 24173326.

[56] Marcel Van Herk, Peter Remeijer, Coen Rasch, and Joos V Lebesque. The probability of correct target dosage: dose-population histograms for deriving treatment margins in radiotherapy. *International Journal of Radiation Oncology* Biology* Physics*, 47(4):1121–1135, 2000.

[57] Luciana Caravatta, Gabriella Macchia, Gian Carlo Mattiucci, Aldo Sainato, Nunzia LV Cernusco, Giovanna Mantello, Monica Di Tommaso, Marianna Trignani, Antonino De Paoli, Gianni Boz, et al. Inter-observer variability of clinical target volume delineation in radiotherapy treatment of pancreatic cancer: a multi-institutional contouring experience. *Radiation Oncology*, 9:1–9, 2014.

[58] Philip P Connell and Samuel Hellman. Advances in radiotherapy and implications for the next century: a historical perspective. *Cancer Research*, 69(2):383–392, 2009.

[59] John F Fowler. The linear-quadratic formula and progress in fractionated radiotherapy. *The British Journal of Radiology*, 62(740):679–694, 1989.

[60] Stephen Joseph McMahon. The linear quadratic model: usage, interpretation and challenges. *Physics in Medicine & Biology*, 64(1):01TR01, 2018.

[61] CM Van Leeuwen, AL Oei, J Crezee, A Bel, NAP Franken, LJA Stalpers, and HP Kok. The alfa and beta of tumours: a review of parameters of the linear-quadratic model, derived from clinical radiotherapy studies. *Radiation Oncology*, 13:1–11, 2018.

[62] Lawrence B Marks, Ellen D Yorke, Andrew Jackson, Randall K Ten Haken, Louis S Constine, Avraham Eisbruch, Søren M Bentzen, Jiho Nam, and Joseph O Deasy. Use of normal tissue complication probability models in the clinic. *International Journal of Radiation Oncology* Biology* Physics*, 76(3):S10–S19, 2010.

[63] Steve Webb. The physical basis of IMRT and inverse planning. *The British Journal of Radiology*, 76(910):678–689, 2003.

[64] Anders Brahme, J-E Roos, and Ingemar Lax. Solution of an integral equation encountered in rotation therapy. *Physics in Medicine & Biology*, 27(10):1221, 1982.

[65] Karl Otto. Volumetric modulated arc therapy: IMRT in a single gantry arc. *Medical Physics*, 35(1):310–317, 2008.

[66] Anders Ahnesjö. Collapsed cone convolution of radiant energy for photon dose calculation in heterogeneous media. *Medical Physics*, 16(4):577–592, 1989.

[67] Chang-Ming Ma, JS Li, T Pawlicki, SB Jiang, J Deng, MC Lee, T Koumrian, M Luxton, and S Brain. A Monte Carlo dose calculation tool for radiotherapy treatment planning. *Physics in Medicine & Biology*, 47(10):1671, 2002.

[68] Kavitha Srinivasan, Mohammad Mohammadi, and Justin Shepherd. Applications of linac-mounted kilovoltage Cone-beam Computed Tomography in modern radiation therapy: A review. *Polish Journal of Radiology*, 79:181, 2014.

[69] Vanessa Da Silva Mendes, Michael Reiner, Lili Huang, Daniel Reitz, Katrin Straub, Stefanie Corradini, Maximilian Niyazi, Claus Belka, Christopher Kurz, Guillaume Landry, et al. ExacTrac Dynamic workflow evaluation: Combined surface optical/thermal imaging and X-ray positioning. *Journal of Applied Clinical Medical Physics*, 23(10):e13754, 2022.

[70] Jenny Bertholet, Antje Knopf, Björn Eiben, Jamie McClelland, Alexander Grimwood, Emma Harris, Martin Menten, Per Poulsen, Doan Trang Nguyen, Paul Keall, et al. Real-time intrafraction motion monitoring in external beam radiotherapy. *Physics in Medicine & Biology*, 64(15):15TR01, 2019.

[71] Arash Navran, Wilma Heemsbergen, Tomas Janssen, Olga Hamming-Vrieze, Marcel Jonker, Charlotte Zuur, Marcel Verheij, Peter Remeijer, Jan-Jakob Sonke, Michiel van den Brekel, et al. The impact of margin reduction on outcome and toxicity in head and neck cancer patients treated with image-guided volumetric modulated arc therapy (VMAT). *Radiotherapy and Oncology*, 130:25–31, 2019.

[72] Abrahim Al-Mamgani, Rob Kessels, Tomas Janssen, Arash Navran, Suzanne van Beek, Casper Carbaat, Willem H Schreuder, Jan-Jakob Sonke, and Corrie AM Marijnen. The dosimetric and clinical advantages of the GTV-CTV-PTV margins reduction by 6 mm in head and neck squamous cell carcinoma: Significant acute and late toxicity reduction. *Radiotherapy and Oncology*, 168:16–22, 2022.

[73] Tobias Finazzi, Miguel A Palacios, Femke OB Spoelstra, Cornelis JA Haasbeek, Anna ME Bruynzeel, Ben J Slotman, Frank J Lagerwaard, and Suresh Senan. Role of on-table plan adaptation in MR-guided ablative radiation therapy for central lung tumors. *International Journal of Radiation Oncology* Biology* Physics*, 104(4):933–941, 2019.

[74] Andrew J McPartlin, XA Li, Lucy E Kershaw, U Heide, L Kerkmeijer, C Lawton, Usman Mahmood, F Pos, N Van As, Marcel Van Herk, et al. MRI-guided prostate adaptive radiotherapy–A systematic review. *Radiotherapy and Oncology*, 119(3):371–380, 2016.

[75] Kelly Kisling, Timothy D Keiper, Daniela Branco, Grace Gwe-Ya Kim, Kevin L Moore, and Xenia Ray. Clinical commissioning of an adaptive radiotherapy platform: Results and recommendations. *Journal of Applied Clinical Medical Physics*, 23(12):e13801, 2022.

[76] Gary P Liney, B Whelan, B Oborn, Michael Barton, and P Keall. MRI-linear accelerator radiotherapy systems. *Clinical Oncology*, 30(11):686–691, 2018.

[77] S Valdenaire, O Riou, N Aillères, P Fenoglietto, D Azria, and P Debuire. Acceptance, commissioning and quality assurance of the MRIdian®: Site experience and three years follow-up. *Cancer/Radiothérapie*, 27(4):303–311, 2023.

[78] Oliver Bieri. Spoiled & Balanced Gradient Echo Methods. In *Proceedings of the International Society for Magnetic Resonance in Medicine*, 2012.

[79] Oliver Bieri and Klaus Scheffler. Fundamentals of balanced steady state free precession MRI. *Journal of Magnetic Resonance Imaging*, 38(1):2–11, 2013.

[80] Karla L Miller. FMRI using balanced steady-state free precession (SSFP). *Neuroimage*, 62(2):713–719, 2012.

[81] Jochen Leupold. *Neue Methoden zur frequenzselektiven Kernspintomographie mit TrueFISP-Sequenzen*. PhD thesis, Verlag nicht ermittelbar, 2005.

[82] Paul Rogowski, Rieke von Bestenbostel, Franziska Walter, Katrin Straub, Lukas Nierer, Christopher Kurz, Guillaume Landry, Michael Reiner, Christoph Josef Auernhammer, Claus Belka, et al. Feasibility and early clinical experience of online adaptive MR-guided radiotherapy of liver tumors. *Cancers*, 13(7):1523, 2021.

[83] Stefanie Corradini, Filippo Alongi, Nicolaus Andratschke, C Belka, Luca Boldrini, Francesco Cellini, J Debus, M Guckenberger, J Hörner-Rieber, FJ Lagerwaard, et al. MR-guidance in clinical reality: current treatment challenges and future perspectives. *Radiation Oncology*, 14:1–12, 2019.

[84] Elia Lombardo, Jennifer Dhont, Denis Page, Cristina Garibaldi, Luise A Künzel, Coen Hurkmans, Rob HN Tijssen, Chiara Paganelli, Paul ZY Liu, Paul J Keall, et al. Real-time motion management in MRI-guided radiotherapy: Current status and AI-enabled prospects. *Radiotherapy and Oncology*, page 109970, 2023.

[85] B.J. Copeland. Encyclopedia Britannica. `https://www.britannica.com/technology/artificial-intelligence`. Accessed: 08.04.2024.

[86] Li Deng, Dong Yu, et al. Deep learning: methods and applications. *Foundations and Trends® in Signal Processing*, 7(3–4):197–387, 2014.

[87] Saad Albawi, Tareq Abed Mohammed, and Saad Al-Zawi. Understanding of a convolutional neural network. In *2017 international conference on engineering and technology (ICET)*, pages 1–6. Ieee, 2017.

[88] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[89] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. pmlr, 2015.

[90] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016.

[91] Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. Large language models in medicine. *Nature Medicine*, 29(8):1930–1940, 2023.

[92] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

[93] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[94] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. Ieee, 2016.

[95] Davood Karimi and Septimiu E Salcudean. Reducing the hausdorff distance in medical image segmentation with convolutional neural networks. *IEEE Transactions on Medical Imaging*, 39(2):499–513, 2019.

[96] Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *International conference on machine learning*, pages 1139–1147. PMLR, 2013.

[97] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(7), 2011.

[98] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[99] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986.

[100] Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):62–66, 1979.

[101] Anil K Jain. *Fundamentals of digital image processing*. Prentice-Hall, Inc., 1989.

[102] James A Sethian et al. *Level set methods and fast marching methods*, volume 98. Cambridge Cambridge UP, 1999.

[103] Juan Eugenio Iglesias and Mert R Sabuncu. Multi-atlas segmentation of biomedical images: a survey. *Medical Image Analysis*, 24(1):205–219, 2015.

[104] Kunihiko Fukushima. Visual feature extraction by a multilayered network of analog threshold elements. *IEEE Transactions on Systems Science and Cybernetics*, 5(4):322–333, 1969.

[105] Eric Kerfoot, James Clough, Ilkay Oksuz, Jack Lee, Andrew P King, and Julia A Schnabel. Left-ventricle quantification using residual U-Net. In *Statistical Atlases and Computational Models of the Heart. Atrial Segmentation and LV Quantification Challenges: 9th International Workshop, STACOM 2018, Held in Conjunction with*

MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers 9, pages 371–380. Springer, 2019.

[106] Lei Huang, Jie Qin, Yi Zhou, Fan Zhu, Li Liu, and Ling Shao. Normalization techniques in training dnns: Methodology, analysis and application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[107] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018.

[108] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[109] Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R Roth, and Daguang Xu. Swin UNETR: Swin transformers for semantic segmentation of brain tumors in MRI images. In *International MICCAI Brainlesion Workshop*, pages 272–284. Springer, 2021.

[110] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, 2021.

[111] Abdel Aziz Taha and Allan Hanbury. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. *BMC Medical Imaging*, 15:1–28, 2015.

[112] Ankur Joshi, Saket Kale, Satish Chandel, and D Kumar Pal. Likert scale: Explored and explained. *British Journal of Applied Science & Technology*, 7(4):396–403, 2015.

[113] Daniel A Low, William B Harms, Sasa Mutic, and James A Purdy. A technique for the quantitative evaluation of dose distributions. *Medical physics*, 25(5):656–661, 1998.

[114] Katherine Mackay, David Bernstein, B Glocker, K Kamnitsas, and A Taylor. A review of the metrics used to assess auto-contouring systems in radiotherapy. *Clinical Oncology*, 35(6):354–369, 2023.

[115] Jaehee Chun, Justin C Park, Sven Olberg, You Zhang, Dan Nguyen, Jing Wang, Jin Sung Kim, and Steve Jiang. Intentional deep overfit learning (IDOL): A novel deep learning strategy for adaptive radiation therapy. *Medical Physics*, 49(1):488–496, 2022.

[116] Xinyuan Chen, Xiangyu Ma, Xuena Yan, Fei Luo, Siran Yang, Zekun Wang, Runye Wu, Jianyang Wang, Ningning Lu, Nan Bi, et al. Personalized auto-segmentation for magnetic resonance imaging–guided adaptive radiotherapy of prostate cancer. *Medical Physics*, 49(8):4971–4979, 2022.

[117] Samuel Fransson, David Tilly, and Robin Strand. Patient specific deep learning based segmentation for magnetic resonance guided prostate radiotherapy. *Physics and Imaging in Radiation Oncology*, 23:38–42, 2022.

[118] Zhenjiang Li, Wei Zhang, Baosheng Li, Jian Zhu, Yinglin Peng, Chengze Li, Jennifer Zhu, Qichao Zhou, and Yong Yin. Patient-specific daily updated deep learning auto-segmentation for MRI-guided adaptive radiotherapy. *Radiotherapy and Oncology*, 177:222–230, 2022.

[119] Grant Haskins, Uwe Kruger, and Pingkun Yan. Deep learning in medical image registration: a survey. *Machine Vision and Applications*, 31:1–18, 2020.

[120] Yabo Fu, Yang Lei, Tonghe Wang, Walter J Curran, Tian Liu, and Xiaofeng Yang. Deep learning in medical image registration: a review. *Physics in Medicine & Biology*, 65(20):20TR01, 2020.

[121] Haonan Xiao, Xinzhi Teng, Chenyang Liu, Tian Li, Ge Ren, Ruijie Yang, Dinggang Shen, and Jing Cai. A review of deep learning-based three-dimensional medical image registration methods. *Quantitative Imaging in Medicine and Surgery*, 11(12):4895, 2021.

[122] Junyu Chen, Eric C Frey, Yufan He, William P Segars, Ye Li, and Yong Du. Transmorph: Transformer for unsupervised medical image registration. *Medical Image Analysis*, 82:102615, 2022.

[123] Daniel Rueckert, Luke I Sonoda, Carmel Hayes, Derek LG Hill, Martin O Leach, and David J Hawkes. Nonrigid registration using free-form deformations: application to breast MR images. *IEEE Transactions on Medical Imaging*, 18(8):712–721, 1999.

[124] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. *Advances in Neural Information Processing Systems*, 28, 2015.

[125] Guha Balakrishnan, Amy Zhao, Mert R Sabuncu, John Guttag, and Adrian V Dalca. VoxelMorph: a learning framework for deformable medical image registration. *IEEE Transactions on Medical Imaging*, 38(8):1788–1800, 2019.

[126] Koen AJ Eppenhof, Maxime W Lafarge, Mitko Veta, and Josien PW Pluim. Progressively trained convolutional neural networks for deformable image registration. *IEEE Transactions on Medical Imaging*, 39(5):1594–1604, 2019.

[127] Alessa Hering, Lasse Hansen, Tony CW Mok, Albert CS Chung, Hanna Siebert, Stephanie Häger, Annkristin Lange, Sven Kuckertz, Stefan Heldmann, Wei Shao, et al. Learn2Reg: comprehensive multi-task medical image registration challenge, dataset and evaluation in the era of deep learning. *IEEE Transactions on Medical Imaging*, 42(3):697–712, 2022.

[128] Anjali Balagopal, Samaneh Kazemifar, Dan Nguyen, Mu-Han Lin, Raquibul Hannan, Amir Owrangi, and Steve Jiang. Fully automated organ segmentation in male pelvic CT images. *Physics in Medicine & Biology*, 63(24):245015, dec 2018.

[129] Nuo Tong, Shuiping Gou, Shuzhe Chen, Yao Yao, Shuyuan Yang, Minsong Cao, Amar Kishan, and Ke Sheng. Multi-task edge-recalibrated network for male pelvic multi-organ segmentation on CT images. *Physics in Medicine & Biology*, 66(3):035001, jan 2021.

[130] Sharmin Sultana, Adam Robinson, Daniel Y. Y. Song, and Junghoon Lee. Automatic multi-organ segmentation in computed tomography images using hierarchical convolutional neural network. *Journal of Medical Imaging*, 7(5):055001, 2020.

[131] Marvin F Ribeiro, Sebastian Marschner, Maria Kawula, Moritz Rabe, Stefanie Corradini, Claus Belka, Marco Riboldi, Guillaume Landry, and Christopher Kurz. Deep learning based automatic segmentation of organs-at-risk for 0.35 T MRgRT of lung tumors. *Radiation Oncology*, 18(1):135, 2023.