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

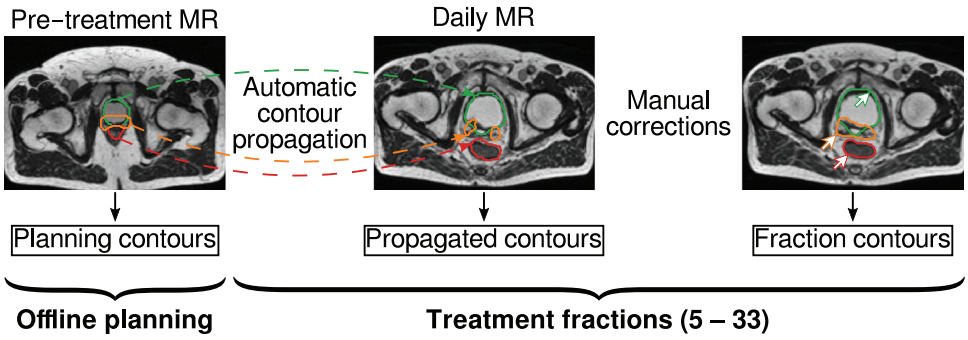

FIGURE 1 Illustration of the adaptive radiotherapy workflow at the MRIdian presenting the different types of contours incorporated in the study.

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

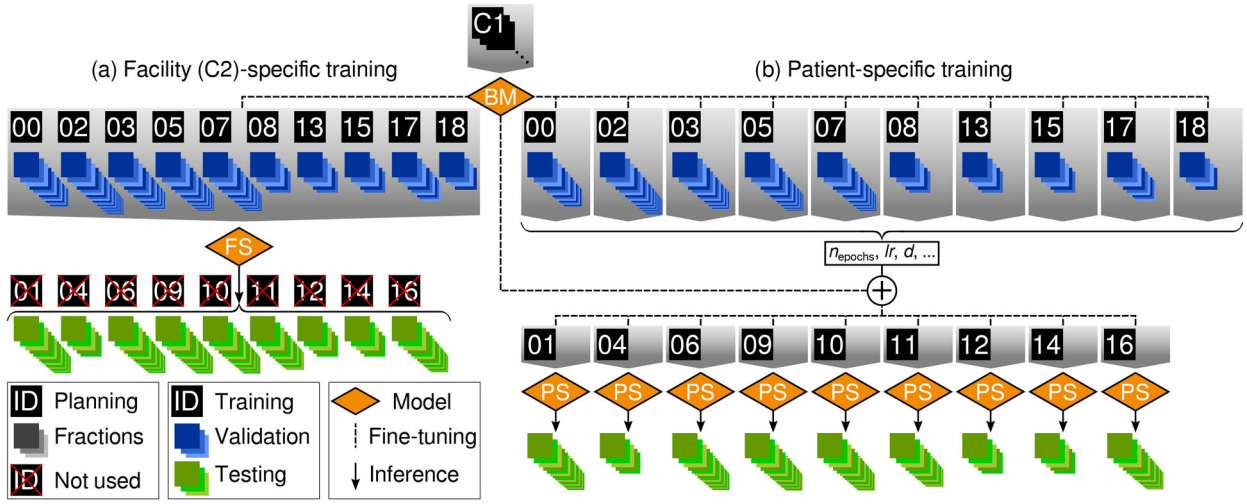

FIGURE 2 Representation of the training scheme as well as the patients (ID) split for (a) the facility-specific (FS) and (b) the patient-specific (PS) training. The gray background indicates images considered together for model training and validation. Both variants share the same test set. The depicted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

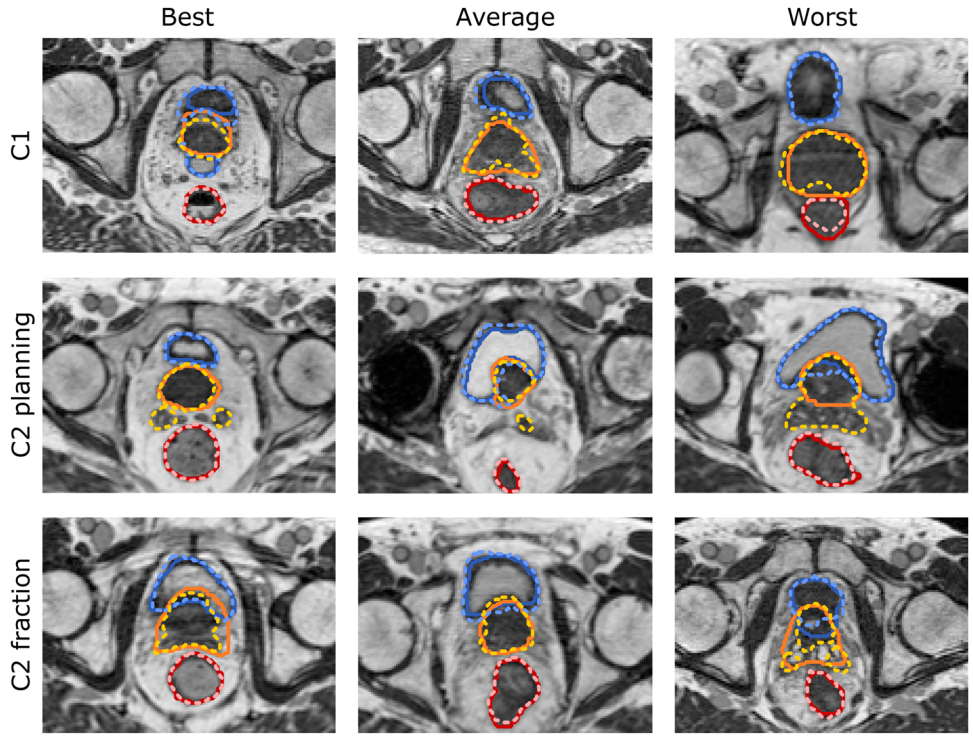

FIGURE 3 Image slices showing (left) one of the best, (middle) average, and (right) worst, baseline model performance. Image slices from (top) C1, (middle) C2 planning, and (bottom) C2 fraction MRs are shown. The (solid line, saturated colors) ground truth and (dashed, faded counterparts) network predictions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

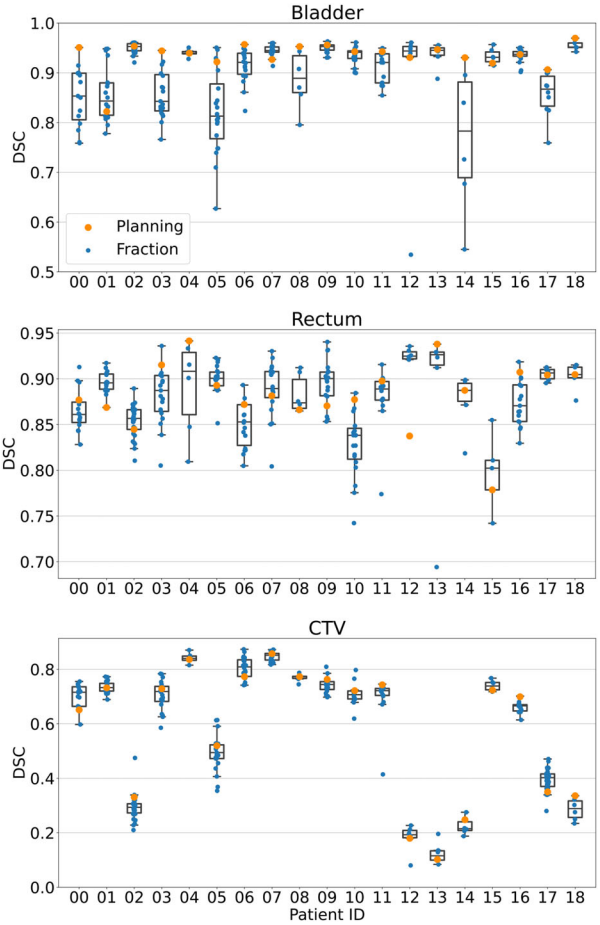

FIGURE 4 The baseline model outcomes. Dice similarity coefficient (DSC) for the bladder, rectum, and clinical target volume (CTV) segmentation for all 19 C2 patients separately. For each patient (horizontal black line) the median value, (orange) performance on the planning data, and (blue) performance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

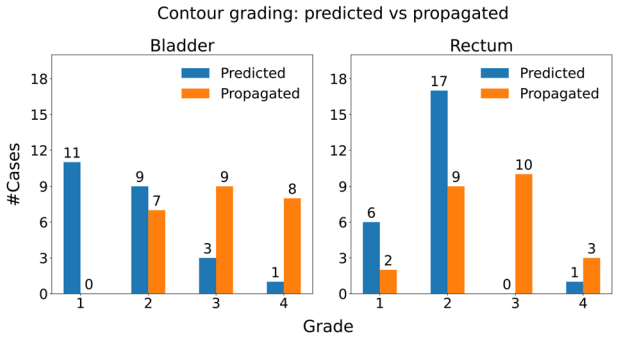

FIGURE 5 Bar plots showing physician's grading of the network predictions (baseline models) and the treatment planning system (TPS)-propagated delineations. The grading is defined as follows: 1—ready-to-use, 2—small corrections, 3—major corrections required, and 4—not useful.

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 predicted	24	0.91(0.09)	0.81(0.02)$^{(ss)}$
		1.5(0.9)	2.2(2.9)$^{(ss)}$
		4.1(2.6)	5.8(6.6)$^{(ss)}$
TPS-propagated	24	0.91(0.1)	0.88(0.16)$^{(ss)}$
		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.

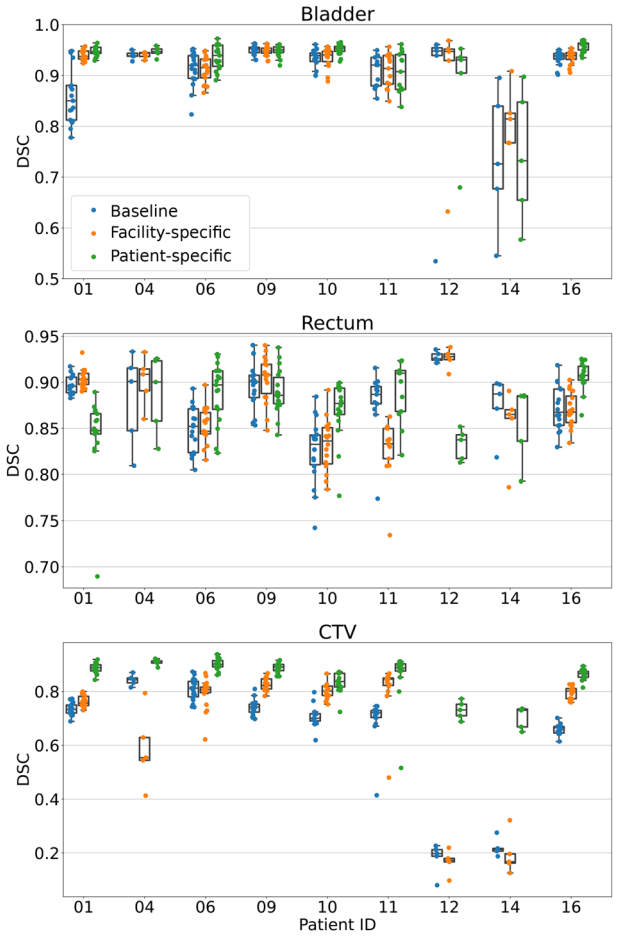

FIGURE 6 Box plots comparing the outcomes of the (blue) baseline, the (orange) facility-, and the (green) patient-specific training for the nine test patients. A single point on the plot represents dice similarity coefficient (DSC) of a predicted fraction contour.

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).

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

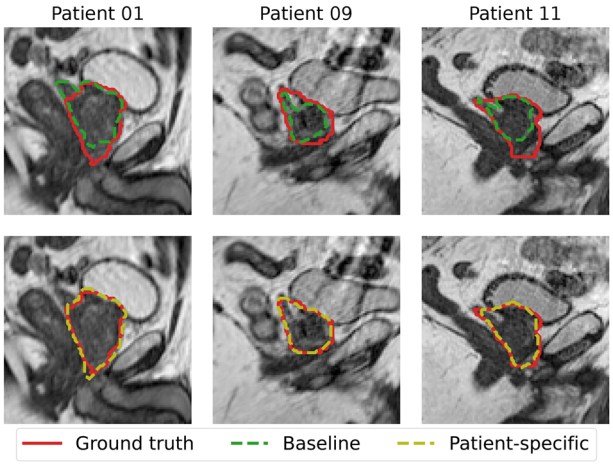

FIGURE 7 Image slices showing the comparison between clinical target volume (CTV) segmentation performed by the (top) baseline and (bottom) patient-specific models.

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