


RESEARCH

Open Access



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

Maria Kawula¹ , Dinu Purice^{1,2}, Minglun Li¹, Gerome Vivar³, Seyed-Ahmad Ahmadi³, Katia Parodi², 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, $V_{60/65/70\text{Gy}}$ for the bladder and $V_{50/65/70\text{Gy}}$ for the rectum, showed agreement between dose distributions within $\pm 5\%$ and $\pm 2\%$, 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



© The Author(s) 2022. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

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 Aquilion LB CT scanner (Canon Medical Systems, Japan) using 512×512 pixels in the axial plane and a variable number of slices. Voxel size was $1.074 \times 1.074 \times 3$ mm³. 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 $1 \times 1 \times 1 \text{ mm}^3$ 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 $1.074 \times 1.074 \times 3 \text{ mm}^3$ grid, and binary masks, defined on a $1 \times 1 \times 1 \text{ mm}^3$ 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 downsampling) 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 $128 \times 128 \times 128$ 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 $\alpha = 20^\circ$, $\beta = \gamma = 10^\circ$, 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

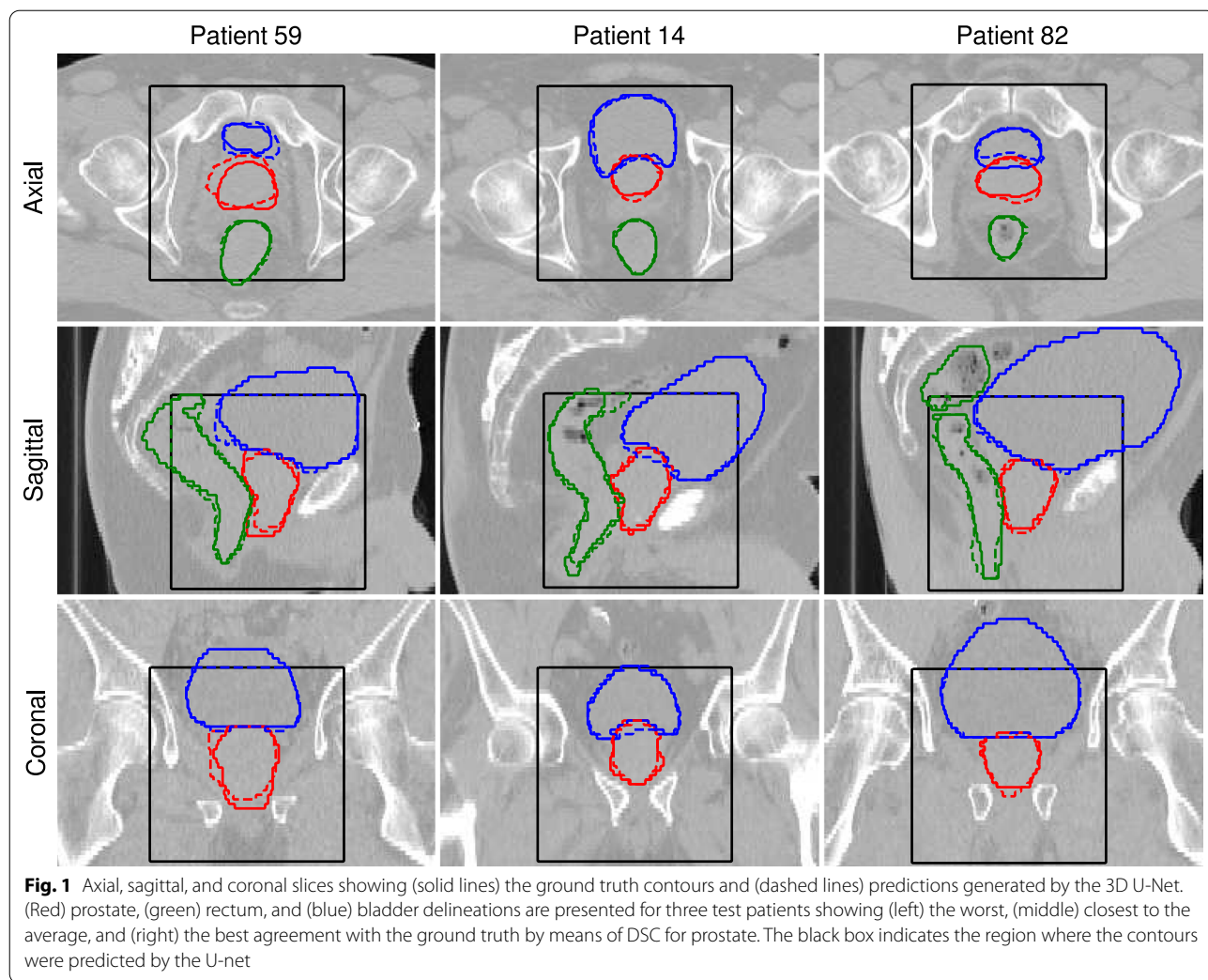


Table 2 Contour based metrics: DSC, average Hausdorff distance (HD_{avg}) and 95% Hausdorff distance ($HD_{95\%}$) of all test patients

	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.

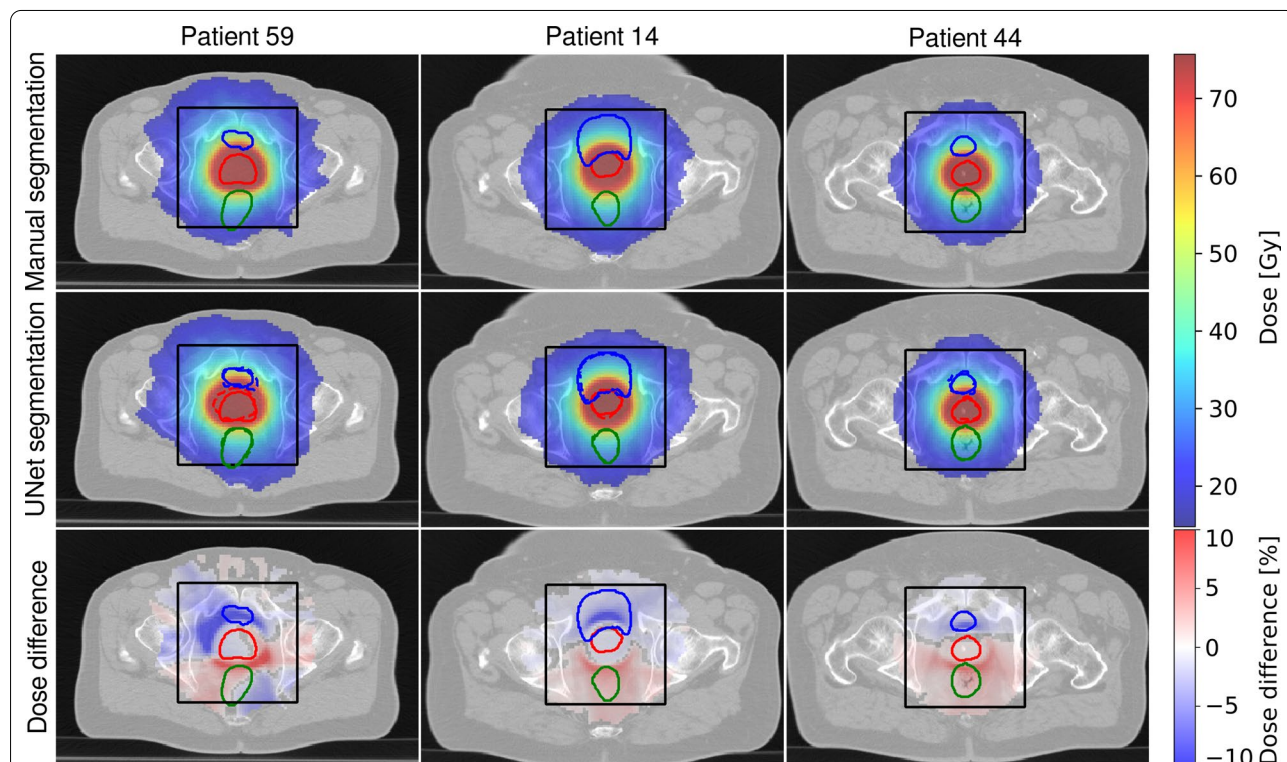
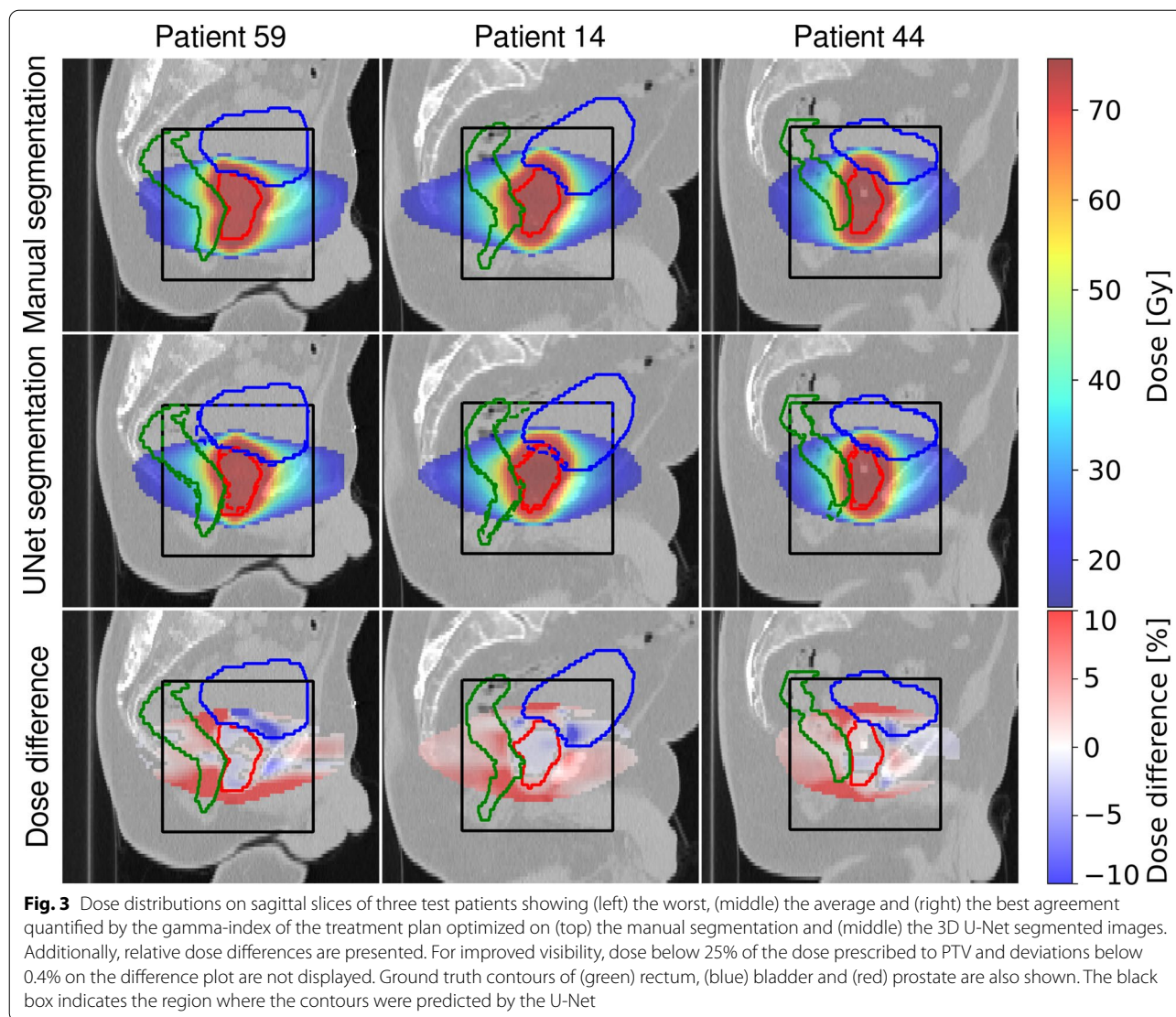


Fig. 2 Dose distributions on axial slices of three test patients showing (left) the worst, (middle) the average and (right) the best agreement quantified by the gamma-index of the treatment plan optimized on (top) the manual segmentation and (middle) the 3D U-Net segmented images. Additionally, relative 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



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

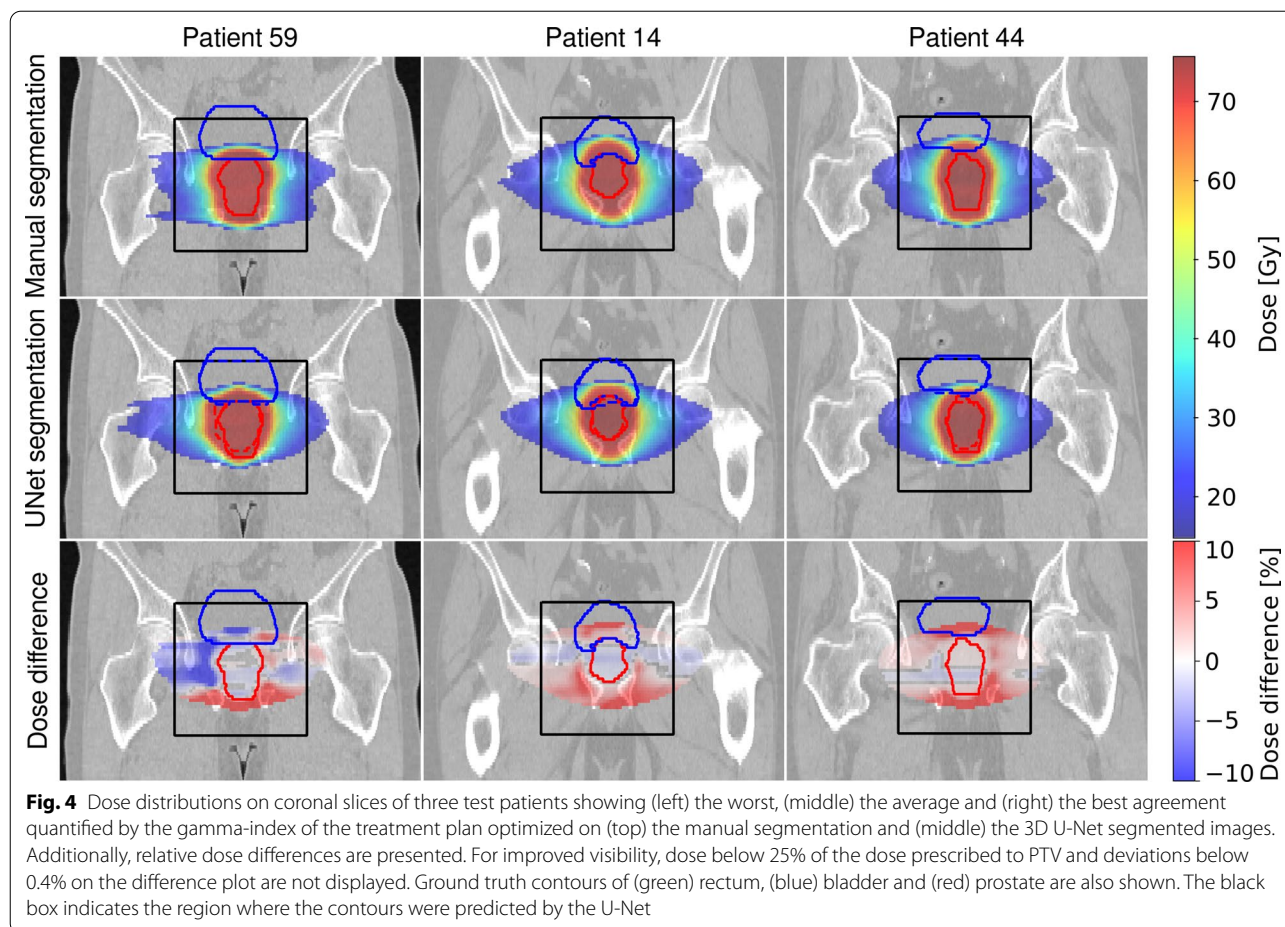


Table 3 Gamma pass rate (3 mm, 3%) and conformity index calculated for plans optimized on manual (CI_{man}) and 3D U-Net ($CI_{3DU-Net}$) generated segmentations of all test patients

	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

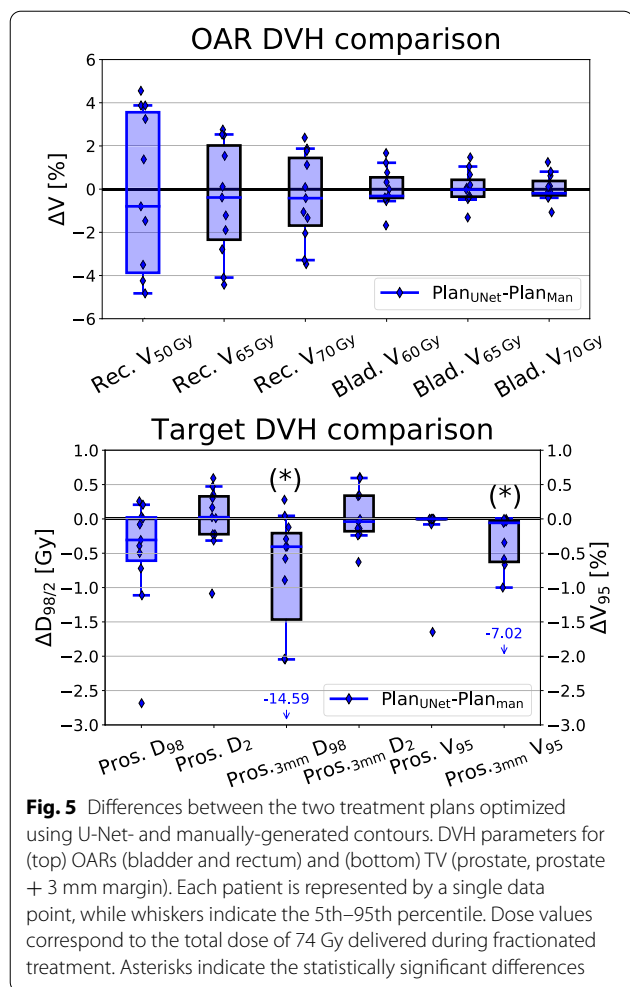


Fig. 5 Differences between the two treatment plans optimized using U-Net- and manually-generated contours. DVH parameters for (top) OARs (bladder and rectum) and (bottom) TV (prostate, prostate + 3 mm margin). Each patient is represented by a single data point, while whiskers indicate the 5th–95th percentile. 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	Balogopal et al. [14]	Sultana et al. [17]	Tong et al. [16]
<i>Prostate</i>				
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
<i>Bladder</i>				
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
<i>Rectum</i>				
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 shown 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 $V_{50/65/70}$ 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. $DSC_{Pat.43} = DSC_{Pat.90} = 0.85$ while $\gamma_{Pat.43} = 93$ and $\gamma_{Pat.90} = 74$ or $DSC_{Pat.44} = 0.88$, $DSC_{Pat.81} = 0.91$ while $\gamma_{Pat.44} = 94$ and $\gamma_{Pat.81} = 87$. 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 research agreements with Elekta Inc., Brainlab GmbH and ViewRay Inc.

Author details

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

Received: 26 July 2021 Accepted: 10 January 2022

Published online: 31 January 2022

References

- 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.
- Guckenberger M, Flentje M. Intensity-modulated radiotherapy (IMRT) of localized prostate cancer. *Strahlenther Onkol*. 2007;183(2):57–62.
- Chen MJ, Weltman E, Hanriot RM, Luz FP, Cecilio 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.
- 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.
- McVicar N, Popescu IA, Heath E. Techniques for adaptive prostate radiotherapy. *Physica Med*. 2016;32(3):492–8.
- 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;50(5):539–46.
- 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.
- 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.
- Cardenas CE, Yang J, Anderson BM, Court LE, Brock KB. Advances in auto-segmentation. In: *Seminars in radiation oncology*. Elsevier; 2019. p. 185–97.
- LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521(7553):436–44.
- 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. Plasti-match: 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, Constone 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, Diodato 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

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

