Dosimetric Impact of Deep Learning-Based CT Auto-Segmentation on IMRT Treatment Planning for Prostate Cancer

Maria Kawula

University Hospital Munich: Klinikum der Universitat Munchen <https://orcid.org/0000-0002-5039-8278>

Dinu Purice

University Hospital Munich: Klinikum der Universitat Munchen

Minglun Li

University Hospital Munich: Klinikum der Universitat Munchen

Gerome Vivar

German Center for Vertigo and Balance Disorder, LMU Munich

Seyed-Ahmad Ahmadi

German center for Vertigo and Balance Disorder, LMU Munich

Katia Parodi

Ludwig-Maximilians-Universität München

Claus Belka

University Hospital Munich: Klinikum der Universitat Munchen

Guillaume Landry

University Hospital Munich: Klinikum der Universitat Munchen

Christopher Kurz (✉ Christopher.Kurz@med.uni-muenchen.de)

University Hospital Munich: Klinikum der Universitat Munchen

Research

Keywords: 3D U-Net, automatic segmentation, IMRT, prostate cancer, neural networks, deep learning

Posted Date: August 2nd, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-718965/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

RESEARCH

Dosimetric impact of deep learning-based CT auto-segmentation on IMRT 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*}

*Correspondence:

Christopher.Kurz@med.uni-muenchen.de

¹Department of Radiation Oncology, University Hospital, LMU Munich, Munich, Germany
Full list of author information is available at the end of the article

Abstract

Background: The evaluation of the automatic segmentation algorithms is commonly performed using geometric metrics, yet an evaluation based on dosimetric parameters might be more relevant in clinical practice but is still 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 intensity-modulated radiation therapy (IMRT) for prostate patients for the first time.

Methods: A database of 69 computed tomography (CT) 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%/3mm gamma-criterion with 10% dose cut-off. A Pearson correlation coefficient between DSC and dosimetric metrics, 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 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 DVH parameters for prostate and prostate + 3mm margin (surrogate clinical target volume) showed good target coverage for the 3D U-Net segmentation with the exception of one case. The average gamma pass-rate was 85%. A comparison between geometric and dosimetric metrics showed no strong statistically significant correlation between these metrics.

Conclusions: The 3D U-Net developed for this work achieved state-of-the-art geometrical performance. The study highlighted the importance of dosimetric evaluation on top of standard geometric parameters and concluded that the automatic segmentation is sufficiently accurate to assist the physicians in manually contouring organs in CT images of the male pelvic region, which is an important step towards a fully automated workflow in IMRT.

Keywords: 3D U-Net; automatic segmentation; IMRT; prostate cancer; neural networks; deep learning

Background

The anatomical structure of the male pelvic region with prostate, surrounded by seminal vesicles, bladder, and rectum, makes intensity-modulated radiation therapy (IMRT) 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 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: a) a pixel-wise comparison of ground-truth and predicted segmentation and b) 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. 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. Since in patients with low-grade prostate cancer tumor is located only in the prostate, seminal vesicles were not considered for segmentation. Realistic VMAT plans with the same optimization settings were created for all test cases using manual segmentations and the automatic segmentations obtained from the 3D U-Net. This allowed to infer for the first time also the dosimetric impact of deep learning delineations.

To this aim, the quality of the treatment plans optimized on the automatically generated contours was compared with the reference plans in terms of dose volume-histograms (DVH) parameters, conformity index (CI) and gamma pass rate. In addition, a standard contour-based analysis based on DSC as well as an average and 95% 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 of Munich. Patients with substantial CT artifacts, e.g., due to fiducial markers or metal bone implants, 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. 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. Images and segmentations (converted into binary masks) have been resampled to a $1 \times 1 \times 1$ mm³ spacing 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 $1.074 \times 1.074 \times 3$ mm³ grid, and binary masks, defined on a $1 \times 1 \times 1$ mm³ grid, we found that employing 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.

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 process is characterized by a grid of $n \times n \times n$ control points and random shifts defined by a Gaussian $[\mu, \sigma]$. 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: a) data augmentation related parameters such as maximal translation shifts, rotation angles, zooming and soft-tissue window limits, B-Spline deformation parameters, augmentation probability and b) 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 a desired network performance was achieved, which is defined as a stage in which the loss function evaluated on the validation data reaches a stable minimum.

Treatment planning

For all test cases, single arc photon VMAT treatment plans were generated from scratch 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 $V_{95\%} = 100\%$ was achieved, while dose to OARs was below the recommendations of the QUANTEC report [20]. Care was additionally taken to choose optimization settings producing 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. For all test cases deviations in the considered OAR and target DVH parameter (see following section) were below $\pm 2\%$ for the perturbed scenario. These settings were then used to optimize a treatment plan using the 3D U-Net segmentations without further user interaction. Table 1 summarizes the goals of the treatment plan along with the importance of each factor.

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

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

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 the colon misclassification. The entire data evaluation 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 [21] 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 prostate + 3 mm (surrogate CTV), values of $D_{98\%}$, $D_{2\%}$ and $V_{95\%}$ were determined. Similarly, for the rectum $V_{50/65/70\text{ Gy}}$ and for the bladder

$V_{60/65/70 \text{ Gy}}$ 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 investigate the correlation between the dosimetric and geometric metrics, the Pearson correlation coefficient [22] 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: $p_{\text{aug}} = 0.93$, rotation angles $\alpha = 20^\circ$, $\beta = \gamma = 10^\circ$, translation shifts $x_{\text{trans}} = y_{\text{trans}} = z_{\text{trans}} = 10\text{mm}$, $l_{\text{min}} = 0.9$, $l_{\text{max}} = 1.1$, $I_{\text{min}} = -150\text{HU}$, $I_{\text{max}} = 150 \text{ HU}$, 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 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 DSC (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.

Table 2: Contour based metrics: DSC, average Hausdorff distance (HD_{avg}) and 95% Hausdorff distance ($\text{HD}_{95\%}$) of all test patients. The last row presents the mean and standard deviation (STD) over all test cases.

	DSC			$(\text{HD}_{\text{avg}}/\text{95\%})$ [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)

Dosimetric analysis

Figure 2 illustrates dose distributions of three exemplary patients with the highest, the average, and the lowest gamma pass-rate. 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. One can observe higher differences approximately 6 mm from the prostate contour corresponding to the PTV margin.

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 0.88 with the 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 the average of 0.78(0.06). The gamma-pass rates (3mm, 3%) were between 71% and 94%, with an average value of 85%.

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

	CI_{man}	$CI_{\text{3D U-Net}}$	(3mm, 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)

Figure 3 illustrates differences between clinically relevant DVH parameters of the two optimized dose distributions. For rectum and bladder, all the differences are below 5% and 2%, respectively. 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 3Gy/2% for D_{98} , D_2 and V_{95} , apart for one outlier (patient 59) 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%. 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 prostate + 3 mm, 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 ($p < 0.05$ confidence level).

Discussion

In this work a 3D U-Net has been successfully trained and applied for CT-based organ segmentation in the male pelvic area. In contrast to current literature, the evaluation of the network's performance was based not only on 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 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 with regard to the correctly classified central part of this organ less relevant. The low contrast of the prostate on the CT images makes its segmentation most challenging. The boundaries of the rectum were satisfactory, with the exception of one case (Pat. 32) in which a substantial portion of the colon was misclassified as part of the rectal contour. 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 bladder and 0.84 for rectum. The edge-calibrated multitask network by Tong *et al.* [16] showed an overall bladder, rectum and prostate segmentation performance of DSC = 0.89. The U-Net-GAN hybrid architecture by Sultana *et al.* [17] achieved at the same time DSC = 0.90. A more detailed comparison is shown in Tab. 4. In all studies, bladder achieved the highest segmentation accuracy, followed by prostate and rectum.

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

		Present work	Balagopal <i>et al.</i> [14]	Sultana <i>et al.</i> [17]	Tong <i>et al.</i> [16]
Prostate	DSC	0.87±0.03	0.90±0.02	0.90±0.05	0.86±0.06
	HD _{avg}	1.6±0.4	-	1.56±0.37	1.01±0.65
	HD _{95%}	4±1	-	5.21±1.2	3.51±1.66
Bladder	DSC	0.96±0.01	0.95±0.02	0.95±0.02	0.96±0.02
	HD _{avg}	0.95±0.2	-	0.95±0.15	0.97±0.53
	HD _{95%}	2.5±0.5	-	4.37±0.56	3.17±3.61
Rectum	DSC	0.88±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

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, especially a trend of lower D_{98} and V_{95} was observed as shown in Fig. 3. In only one case (patient 59), major deviations were observed, which can be explained by an incorrect prostate contouring that is shifted towards the bladder, as can be seen in Fig. 1.

DVH parameters for OARs also showed satisfactory results. Treatment plans optimized on 3D U-Net-generated contours resulted in a slightly lower dose to the rectum with the greatest deviations of $\pm 4\%$, measured by $V_{50/65/70 \text{ Gy}}$. No considerable differences were found for the bladder, showing deviations of the order of $\pm 2\%$ for $V_{60/65/70 \text{ Gy}}$. Consequently, results suggest, that plans optimized on automatically generated contours do not overdose the neighboring OARs, i.e. bladder and rectum.

The average value of the CI for 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 index confirms slightly worse target coverage. More pronounced deviations between the dose distributions were captured by the gamma index analysis, showing values of 71-94% with a mean value of 85%.

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 even $DSC_{Pat.44}=0.88$, $DSC_{Pat.81}=0.91$ while $\gamma_{Pat.44} = 94$ and $\gamma_{Pat.81} = 87$. This leads to the conclusion, that a high geometric similarity between contours 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) used by the contouring software as well as the TPS, and the binary masks required for CNN training. The use of linear interpolation in the conversion pipeline caused the original contours to shrink, which led to differences in structural volume and subsequently to underdosage in the PTV. The nearest neighbors interpolation instead 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 shows 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 good target volume coverage in most 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.

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.

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>.

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.

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.

Authors' contributions

MK 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 & AA provided the core part of the 3D U-Net implementation. KP & CB reviewed the manuscript and helped to finalize it. GL & CK designed the study, participated in all stages of this work from data preparation, network training, data analysis and writing the manuscript.

Acknowledgements

Not applicable.

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
- $HD_{95\%}$ - 95th percentile Hausdorff distance
- HU- Hounsfield unit
- IMRT- Intensity-modulated radiation therapy
- MR- magnetic resonance
- OAR(s)- Organ volume(s)
- 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

Authors' information

Not applicable.

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.

References

1. Hummel, S., Simpson, E., Hemingway, P., Stevenson, M., Rees, A.: Intensity-modulated radiotherapy for the treatment of prostate cancer: a systematic review and economic evaluation. *Health Technol Assess* **14**(47), 1–108 (2010)
2. Guckenberger, M., Flentje, M.: Intensity-modulated radiotherapy (imrt) of localized prostate cancer. *Strahlentherapie und Onkologie* **183**(2), 57–62 (2007)
3. Chen, M.J., Weltman, E., Hanriot, R.M., Luz, F.P., Cecílio, P.J., Da Cruz, J.C., Moreira, F.R., Santos, A.S., Martins, L.C., Nadalin, W.: Intensity modulated radiotherapy for localized prostate cancer: rigid compliance to dose-volume constraints as a warranty of acceptable toxicity? *Radiation Oncology* **2**(1), 1–7 (2007)
4. Wu, Q.J., Thongphiew, D., Wang, Z., Mathayomchan, B., Chankong, V., Yoo, S., Lee, W.R., Yin, F.-F.: On-line re-optimization of prostate imrt plans for adaptive radiation therapy. *Physics in Medicine & Biology* **53**(3), 673 (2008)
5. McVicar, N., Popescu, I.A., Heath, E.: Techniques for adaptive prostate radiotherapy. *Physica Medica* **32**(3), 492–498 (2016)
6. Choi, H., Kim, Y., Lee, S., Lee, Y., Park, G., Jung, J., Cho, B., Park, S., Ahn, H., Kim, C.-S., Yi, S.Y., Ahn, S.: Inter- and intra-observer variability in contouring of the prostate gland on planning computed tomography and cone beam computed tomography. *Acta oncologica (Stockholm, Sweden)* **50**, 539–46 (2011). doi:[10.3109/0284186X.2011.562916](https://doi.org/10.3109/0284186X.2011.562916)
7. Nyholm, T., Jonsson, J., Söderström, K., Bergström, P., Carlberg, A., Frykholm, G., Behrens, C.F., Geertsen, P.F., Trepikas, R., Hanvey, S., *et al.*: Variability in prostate and seminal vesicle delineations defined on magnetic resonance images, a multi-observer,-center and-sequence study. *Radiation oncology* **8**(1), 1–12 (2013)
8. Wong, J., Fong, A., McVicar, N., Smith, S., Giambattista, J., Wells, D., Kolbeck, C., Giambattista, J., Gondara, L., Alexander, A.: Comparing deep learning-based auto-segmentation of organs at risk and clinical target volumes to expert inter-observer variability in radiotherapy planning. *Radiotherapy and Oncology* **144**, 152–158 (2020)
9. Cardenas, C.E., Yang, J., Anderson, B.M., Court, L.E., Brock, K.B.: Advances in auto-segmentation. In: *Seminars in Radiation Oncology*, vol. 29, pp. 185–197 (2019). Elsevier
10. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. *nature* **521**(7553), 436–444 (2015)
11. Savenije, M.H., Maspero, M., Sikkes, G.G., van der Voort van Zyp, J.R., T.J Kotte, A.N., Bol, G.H., T. van den Berg, C.A.: Clinical implementation of mri-based organs-at-risk auto-segmentation with convolutional networks for prostate radiotherapy. *Radiation Oncology* **15**, 1–12 (2020)
12. Chung, S.Y., Chang, J.S., Choi, M.S., Chang, Y., Choi, B.S., Chun, J., Keum, K.C., Kim, J.S., Kim, Y.B.: Clinical feasibility of deep learning-based auto-segmentation of target volumes and organs-at-risk in breast cancer patients after breast-conserving surgery. *Radiation Oncology* **16**(1), 1–10 (2021)
13. Milletari, F., Navab, N., Ahmadi, S.-A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: *2016 Fourth International Conference on 3D Vision (3DV)*, pp. 565–571 (2016). IEEE
14. Balagopal, A., Kazemifar, S., Nguyen, D., Lin, M.-H., Hannan, R., Owringi, A., Jiang, S.: Fully automated organ segmentation in male pelvic ct images. *Physics in Medicine & Biology* **63**(24), 245015 (2018)

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. *Medical image analysis* **54**, 168–178 (2019)
16. Tong, N., Gou, S., Chen, S., Yao, Y., Yang, S., Cao, M., Kishan, A., Sheng, K.: Multi-task edge-recalibrated network for male pelvic multi-organ segmentation on ct images. *Physics in Medicine & Biology* **66**(3), 035001 (2021)
17. Sultana, S., Robinson, A., Song, D.Y., Lee, J.: Automatic multi-organ segmentation in computed tomography images using hierarchical convolutional neural network. *Journal of Medical Imaging* **7**(5), 055001 (2020)
18. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. *arXiv preprint arXiv:1406.2661* (2014)
19. Taha, A.A., Hanbury, A.: Metrics for evaluating 3d medical image segmentation: analysis, selection, and tool. *BMC medical imaging* **15**(1), 1–28 (2015)
20. Marks, L.B., Yorke, E.D., Jackson, A., Ten Haken, R.K., Constine, L.S., Eisbruch, A., Bentzen, S.M., Nam, J., Deasy, J.O.: Use of normal tissue complication probability models in the clinic. *International Journal of Radiation Oncology* Biology* Physics* **76**(3), 10–19 (2010)
21. Paddick, I.: A simple scoring ratio to index the conformity of radiosurgical treatment plans. *Journal of neurosurgery* **93**(supplement.3), 219–222 (2000)
22. Benesty, J., Chen, J., Huang, Y., Cohen, I.: Pearson correlation coefficient. In: *Noise Reduction in Speech Processing*, pp. 1–4. Springer, ??? (2009)

Figures

Figure 1: Axial slices showing (solid lines) the ground truth contours and (dashed lines) predictions generated by the 3D U-Net. (Red) prostate, (green) rectum, and (blue) bladder delineations are presented for three test patients showing (left) the worst, (middle) the average, and (right) the best agreement with the ground truth by means of DSC for prostate.

Figure 2: Dose distributions 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 (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.

Figure 3: 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.

Figures

Figure 1

Axial slices showing (solid lines) the ground truth contours and (dashed lines) predictions generated by the 3D U-Net. (Red) prostate, (green) rectum, and (blue) bladder delineations are presented for three test patients showing (left) the worst, (middle) the average, and (right) the best agreement with the ground truth by means of DSC for prostate.

Figure 2

Dose distributions 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 (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.

Figure 3

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.