Physica Medica 125 (2024) 104486

Contents lists available at ScienceDirect

Physica Medica

journal homepage: www.elsevier.com/locate/ejmp

Original paper

Automated segmentation in pelvic radiotherapy: A comprehensive evaluation of ATLAS-, machine learning-, and deep learning-based models

B. Bordigoni [a], S. Trivellato [a], R. Pellegrini [b], S. Meregalli [c], E. Bonetto [c], M. Belmonte [c,d], M. Castellano [c,d], D. Panizza [a,d], S. Arcangeli [c,d,*], E. De Ponti [a,d]

[a] Medical Physics, Fondazione IRCCS San Gerardo dei Tintori, Monza, Italy
[b] Medical Affairs, Elekta AB, Stockholm, Sweden
[c] Radiation Oncology, Fondazione IRCCS San Gerardo dei Tintori, Monza, Italy
[d] School of Medicine and Surgery, University of Milano Bicocca, Milano, Italy

ARTICLE INFO

Keywords:
Artificial intelligence
Machine learning
Deep learning
Automated contouring

ABSTRACT

Artificial intelligence can standardize and automatize highly demanding procedures, such as manual segmentation, especially in an anatomical site as common as the pelvis. This study investigated four automated segmentation tools on computed tomography (CT) images in female and male pelvic radiotherapy (RT) starting from simpler and well-known atlas-based methods to the most recent neural networks-based algorithms. The evaluation included quantitative, qualitative and time efficiency assessments. A mono-institutional consecutive series of 40 cervical cancer and 40 prostate cancer structure sets were retrospectively selected. After a preparatory phase, the remaining 20 testing sets per each site were auto-segmented by the atlas-based model STAPLE, a Random Forest-based model, and two Deep Learning-based tools (DL), MVision and LimbusAI. Setting manual segmentation as the Ground Truth, 200 structure sets were compared in terms of Dice Similarity Coefficient (DSC), Hausdorff Distance (HD), and Distance-to-Agreement Portion (DAP). Automated segmentation and manual correction durations were recorded. Expert clinicians performed a qualitative evaluation. In cervical cancer CTs, DL outperformed the other tools with higher quantitative metrics, qualitative scores, and shorter correction times. On the other hand, in prostate cancer CTs, the performance across all the analyzed tools was comparable in terms of both quantitative and qualitative metrics. Such discrepancy in performance outcome could be explained by the wide range of anatomical variability in cervical cancer with respect to the strict bladder and rectum filling preparation in prostate Stereotactic Body Radiation Therapy (SBRT). Decreasing segmentation times can reduce the burden of pelvic radiation therapy routine in an automated workflow.

1. Introduction

In recent decades, artificial intelligence (AI) significantly affected several areas of health care. Within clinical oncology, AI may have the potential to transform the radiotherapy workflow, resulting in improved quality, standardization, safety, accuracy, and timeliness of radiotherapy delivery [1,2].

Manual image segmentation is a time-consuming task routinely performed by radiation oncologists or radiation therapists to identify each patient's targets and adjacent organs-at-risk (OARs). Additionally, the radiotherapy plan efficacy and safety require segmentation as accurate as possible. However, even if segmentation is performed according to the same guidelines, inter- and intra-observers' inconsistencies and large heterogeneity may still exist and strongly affect treatment outcomes [1,3,4,5]. A recent review across five trials studied major delineation deviations [6], and Cox et al registered 'unacceptable' or 'major' deviations in 2.9–13.4 % of cases [7]. Furthermore, radiomics or treatment plan analyses, such as dose-volume histograms (DVH) evaluation, can be affected by manual image segmentation being strictly related to contouring accuracy. Automated segmentation of targets and normal tissues might address all these challenges.

The field of AI-based medical image segmentation has seen accelerated growth also offering a solution to overcome several image-related problems to get an accurate and efficient automated segmentation [1]. Firstly, medical images are affected by noise that can influence each

* Corresponding author at: Radiation Oncology, Fondazione IRCCS San Gerardo dei Tintori, Monza, Italy.
E-mail address: stefano.arcangeli@unimib.it (S. Arcangeli).

https://doi.org/10.1016/j.ejmp.2024.104486
Received 16 December 2023; Received in revised form 20 June 2024; Accepted 17 July 2024

voxel intensity [8]. Secondly, tissues within a patient typically exhibit intensity non-uniformity, meaning that voxel intensities within a single tissue may gradually vary over the image extent [9]. Furthermore, especially in abdomen and pelvis CT scans, the scarce contrast between soft tissue organs and the high anatomical variability usually limits automated segmentation performances [9]. Lastly, images are reconstructed during acquisition to have a predefined voxel size leading to partial volume averaging [10].

In recent years, taking full advantage of Convolutional Neural Network, automated segmentation tools have migrated from the research domain to commercially available products. Although such products may represent ideal solutions, their clinical deployment raises significant considerations, including quality assurance and validation. Guidance from the European Society for Radiotherapy and Oncology (ESTRO) physics workshop on the AI implementation suggested commissioning vendor's validation report with combination of quantitative and qualitative evaluation metrics by comparing automated to manual segmentation set as Ground Truth (GT) [11–13]. Additionally, the comparison with the Institution specific GT (Institution specific) allows to deeply understand which are the pros and cons in the clinical use of a specific tool.

Commercial deep learning (DL)-based tools were widely investigated in the OARs automated image segmentation in several districts by now [14–17]. In contrast, few examples were reported in the literature investigating DL in the pelvic district [18–20]. Simpler algorithms-based systems were preferred, especially the atlas-based in the female pelvis [21]. In male pelvis, automated segmentation in prostate RT were widely evaluated with atlas-based tools [22–24] and, recently, DL models have been compared to manual segmentation [25–28]. To our knowledge, this is the first study investigating and comparing four different algorithms for OAR automated segmentation in both female and male pelvis radiotherapy (RT), bringing a broad comparison moving from the earliest widely used automated techniques to the actual sophisticated AI-based tools. The first one is an atlas-based model (STAPLE, Simultaneous Truth And Performance Level Estimation, Atlas-Based Autosegmentation (ABAS), v2.01.01, Elekta AB, Sweden). The second one is a conventional machine learning-based model which relies on a combination of decision trees as a random forest (Admire, v 3.47, Elekta AB, Sweden). The most recent algorithms are two DL-based tools MVision (v1.2.2, MVision AI, Finland) and LimbusAI (v1.7.0-B3, Limbus AI Inc, Canada). A comprehensive quantitative evaluation has been performed including the Distance-to-Agreement Portion (DAP), the Dice Similarity Coefficient (DSC) and Hausdorff Distance (HD). The study comprehended a time-based evaluation collecting automated segmentation times and radiation oncologists' correction times. Finally, to clinically assess automated tools, a qualitative evaluation of automated contours has been performed.

2. Materials and methods

2.1. Automated segmentation tools

STAPLE is a statistical method of combining atlases from multiple segmentations into a single result implemented in the Elekta ABAS tool. This method is applied for each structure separately, taking as input the structure set of selected atlases. It automatically estimates multiple classifiers for each atlas and then it estimates the ground truth segmentation, as a weighted combination of such classifiers. For this purpose, a not so large atlases sample size is necessary to correctly autosegment CT images. Published studies proved that segmentation performances improve with an increasing number of atlases, reaching a plateau at about 5–10 atlases [24,29,30].

Random Forest is a conventional machine learning-based model relying on a decision trees-combination implemented in the Elekta Admire tool. Unlike STAPLE, the label decision is interpreted as a classification problem operating on ambiguous voxel labels by using the intensity information from the previously deformed atlas images. Each tree assesses a set of features to determine the class label and the majority voting is used to decide the final class label. Both STAPLE and Random Forest run on a workstation with NVIDIA Quadro GP100 and produced structure sets including the same structures included in the used atlases.

On the other hand, both DL-tools, MVision and Limbus, use a Convolutional Neural Networks structure relying on a U-Net architecture, widely used thanks to its segmentation capabilities [31–33]. Limbus application run on a workstation with an NVIDIA Quadro P2000 and it allows to choose the desired anatomical structures. On the other hand, MVision is a web-based tool with no need for local installation and it only proposes the anatomical district choice.

2.2. Patient cohort

A mono-institutional cohort of 80 patients, treated from November 2018 to November 2022, was retrospectively selected. 40 of them underwent post-operative cervical cancer External Beam Radiotherapy (EBRT) treatment (48–55 Gy in 25–27 fractions). The remaining 40 patients underwent prostate cancer SBRT. 20 out of 40 prostate cancer patients were treated with 36.25 Gy in 5 fractions and 20 received 42.7 Gy in 7 fractions. All patients underwent a CT simulation in the supine position with 120 kVp and images were reconstructed in a 512 × 512 matrix (0.98 × 0.98 mm^2 pixel size) with the standard filtered back-projection, a slice thickness of 3 mm and 1 mm and a median number of slices of 128 [84–180] and 315 [229–414] for cervical cancer and prostate cancer patients, respectively.

Cervical cancer patients received bladder and rectum preparation indications to be followed at home. On the other hand, prostate cancer patients followed a strict SBRT preparation protocol at the Department that included drinking 0.5 L of water and rectal micro-enema 30 min before the CT images acquisition. Patients' original structure sets were manually segmented on CT images by two cervical cancer and a prostate cancer expert radiation oncologists at the Monaco TPS (v5.51.11, Elekta AB, Sweden). Manual-segmented structure sets have been set as the Ground Truth to evaluate OAR-automated segmentation performances: rectum, bladder, and femoral heads for both cervical and prostate cancer structure sets with the addition of bowel bag for the first ones mentioned.

2.3. Datasets preparation

Selected structure sets have been subdivided into a preparatory and a testing dataset. 20 cervical cancer patients and 20 prostate cancer patients had been carefully selected with a sufficient size and shape variability for each structure and included in the preparatory set for ATLAS-based and Random Forest models. The inclusion of standard patients only could affect the algorithm robustness: it is in fact necessary to cover the anatomical variations as much as possible (i.e. thin/obese, empty/full bladder or rectum).

To avoid any bias, the remaining 20 cervical cancer and 20 prostate cancer sets constituted the so-called testing dataset, used to evaluate all the four selected tools.

Firstly, STAPLE and Random Forest underwent a validation phase aiming to determine the minimum numerosity of the preparatory set in a closed-loop evaluation (one-in-one-out) of automated segmentation results. This meant to determine the minimum number of atlases for STAPLE and the minimum available training sample size for Random Forest. While STAPLE results were not improved by the increasing atlas number, the Random Forest-based algorithm showed an accuracy gain moving from 10 to 20 cases. This led to the use of the whole available preparatory set for both algorithms.

Instead, it was not necessary to perform a training phase for the two selected Convolutional Neural Networks-based software: these commercial tools were trained by producers taking advantage of very large available training image datasets [34,35].

2.4. Evaluation metrics

2.4.1. Time metrics

Firstly, automated segmentation times per set of structures were registered for each of the four analyzed tools. 5 out of 20 cervical cancer structure sets and 5 out of 20 prostate cancer structure sets were randomly selected to register the contouring correction times needed by two cervical cancer-expert radiation oncologists and a prostate cancer-expert radiation oncologist, respectively.

2.4.2. Quantitative metrics

All metrics have been calculated through Golden Rule software (Canis Lupus LLC, 2023). In particular, the algorithm performances were assessed by computing the slice-wise DSC and the volumetric HD, being the most popular metrics for the evaluation of segmentation [12]. As institutional criteria, an acceptable DSC score was set at 0.9 and 0.8 for bony structures and soft tissues, respectively [12,36]. Furthermore, performances were also compared by means of the DAP, i.e., the percentage of automatic contours standing 1 mm-, 3 mm-, and 5 mm-far from the manual Ground Truth (DAP_{1mm}, DAP_{3mm}, DAP_{5mm}). The DAP is a Golden Rule original metric which can be compared to the added path length (APL) [37]. Both DAP and APL focus on automated segmentation discrepancies from the manual Ground Truth in terms of shrinkage and expansion.

2.4.3. Qualitative clinical evaluation

Finally, a qualitative evaluation has been performed by two cervical cancer-expert radiation oncologists and a prostate cancer-expert radiation oncologist for a subset of 5 cases for each investigated district by means of a five degrees-score table as follows:

1. GOOD AGREEMENT for acceptable structures to treat “as is”
2. MINOR DIFFERENCES for modest non-critical edits required (10–20 % of the volume) far from the target
3. EDITS REQUIRED for modest non-critical edits required (10–20 % of the volume) in the area of the target
4. MODERATE EDITS REQUIRED for moderate changes required (20–50 % of the volume)
5. GROSS ERROR in case of no resemblance to the clinical structure or 75 % slices needing edit [19].

2.5. Statistical analysis

Quantitative results have been compared to determine the presence of statistical differences between algorithm performances. The Shapiro-Wilk test established whether to perform the parametric *t*-test or the non-parametric Wilcoxon-Mann-Whitney rank-sum test. The Bonferroni correction for multiple tests has been applied with a selected significance level at 5 % ($p = 0.05$). All the statistical tests have been performed using the software Rstudio (2021.09.0).

Finally, the Linear Weighted Cohen’s Kappa coefficient was evaluated with MatLab (R2023b) to assess the inter-rater reliability of the two cervical cancer-experts in the qualitative evaluation. Cohen’s kappa scores were defined as poor ($k < 0.20$), fair ($0.21 < k < 40$), moderate ($0.41 < k < 0.60$), good ($0.61 < k < 0.80$), and excellent ($k > 0.81$) [38].

3. Results

3.1. Time evaluation

The registered automated segmentation time ranges for the testing set were listed in Table 1 along with manual correction times. For technical reasons, the STAPLE automated segmentation times in prostate cancer patients were not recorded for the whole group of 20 structure sets. For the sake of completeness, values were reported for the registered subset. At our Institution, typical expert manual segmentation times for cervical cancer EBRT and prostate cancer SBRT have been estimated at 30–45 min and 90 min, respectively. On the other hand, typical junior radiation oncologists manual segmentation times for pelvic RT can be up to 2 h.

3.2. Quantitative evaluation

The Shapiro test established the non-normality of data distributions: DSC and HD were summarized as median and range values in Table 2 and p-values from Wilcoxon-Mann-Whitney rank-sum test, corrected by Bonferroni, were reported in Table 3. In cervical cancer patients, bladder DSC and HD values improved passing from STAPLE and Random Forest to DL tools. It is worth noticing that the median STAPLE bladder DSC did not reach a DSC > 0.8. The DL’s better automatic segmentation performances were confirmed by the DAP comparison, as reported in Table 4. A similar trend was observed for rectum, with lower values for STAPLE and significantly better DL performance. STAPLE and Random Forest did not reach the soft tissue threshold set to 0.8 for DSC metric in rectum delineation for cervical cancer. As for the bladder, DL models showed higher DAP results (Table 4).

On the other hand, in prostate cancer patients, a good agreement between manual and automatic contours was globally registered in bladder and rectum definition and these results were confirmed at the DAP comparison. No statistically significant differences were registered in algorithm performances. It is worth noticing that the minimum DSC values recorded in rectum DL delineation could be due to the inaccurate distinction of rectum from sigmoid colon.

In cervical cancer cases, bowel bag contouring registered differences in cranio-caudal extent. To reduce this bias and focus on the interesting volume, the extent was redefined as limited to the manual last cranial slice at maximum. The evaluation has then been repeated on cropped automatic contours. The recorded maximum HD values were investigated: discrepancies were mainly in the caudal part of the bowel bag and due to identification correctness of rectum/bladder/bowel borders. As shown in Table 2, Limbus and STAPLE showed a slightly lower

Table 1
Automated segmentation times and manual correction times per structure set (testing set). Median values and [minimum–maximum] ranges are reported.

TIME (minutes)	ST	RF	MV	LI
CC				
AS	21.9 [16.8–36.2]	20.9 [18.1–24.7]	0.7 [0.5–1.1]	1.1 [0.8–1.3]
MC (1)	23.0 [17.0–37.0]	24.0 [13.0–25.0]	10.0 [5.0–16.0]	12.0 [9.0–16.0]
MC (2)	30.0 [20.0–60.0]	30.0 [20.0–30.0]	5.0 [3.0–12.0]	7.0 [5.0–20.0]
PC				
AS	28.7 [24.0–33.0]	22.0 [18.0–28.0]	1.8 [1.2–2.6]	2.0 [1.5–3.0]
MC	16.0 [14.0–23.0]	20.0 [18.0–21.0]	16.0 [14.0–21.0]	15.0 [14.0–19.0]

Abbreviations: *CC: cervical cancer, AS: automated segmentation, MC: manual correction, (1): expert 1, (2): expert 2, PC: prostate cancer, ST: STAPLE, RF: random forest, MV: MVision, LI: Limbus.*

Table 2
Comparison of automatic segmentation performance in terms of quantitative metrics in cervical and prostate cancer patients (testing set). Median values and ranges are reported.

CC METRICS	ST	RF	MV	LI
Bladder				
DSC	0.77 [0.12–0.90]	0.89 [0.58–0.97]	0.95 [0.89–0.98]	0.94 [0.79–0.98]
HD (mm)	25.91 [15.84–173.62]	19.26 [9.52–45.35]	13.30 [5.15–25.82]	13.81 [5.90–53.02]
Rectum				
DSC	0.61 [0.34–0.77]	0.78 [0.47–0.89]	0.87 [0.11–0.93]	0.84 [0.75–0.92]
HD (mm)	37.29 [18.35–106.15]	26.39 [12.32–52.32]	23.81 [6.99–51.84]	23.32 [8.68–37.10]
LFH				
DSC	0.94 [0.89–0.96]	0.95 [0.92–0.97]	0.95 [0.87–0.97]	0.94 [0.89–0.95]
HD (mm)	13.59 [8.89–63.09]	13.21 [8.32–64.08]	16.23 [6.36–66.7]	15.98 [9.16–63.37]
RFH				
DSC	0.94 [0.66–0.96]	0.96 [0.92–0.97]	0.95 [0.84–0.97]	0.95 [0.90–0.96]
HD (mm)	13.48 [9.47–157.22]	13.26 [8.93–56.22]	16.34 [6.77–55.28]	15.45 [6.70–58.62]
Bowel bag				
DSC	0.84 [0.10–0.91]	0.88 [0.75–0.93]	0.88 [0.74–0.94]	0.79 [0.55–0.91]
HD (mm)	40.32 [26.81–133.17]	37.16 [27.58–65.44]	33.07 [25.96–80.92]	48.19 [26.47–118.29]

PC METRICS	ST	RF	MV	LI
Bladder				
DSC	0.91 [0.70–0.97]	0.88 [0.63–0.96]	0.94 [0.91–0.98]	0.93 [0.86–0.98]
HD (mm)	14.27 [6.95–38.55]	17.68 [8.08–40.55]	11.42 [7.49–18.87]	12.43 [4.79–20.58]
Rectum				
DSC	0.85 [0.76–0.89]	0.87 [0.79–0.90]	0.87 [0.78–0.92]	0.87 [0.74–0.91]
HD (mm)	16.50 [9.65–36.46]	16.55 [10.00–30.62]	19.57 [11.86–36.46]	18.64 [7.25–46.37]
LFH				
DSC	0.96 [0.87–0.98]	0.97 [0.88–0.98]	0.97 [0.92–0.98]	0.97 [0.93–0.97]
HD (mm)	11.46 [4.81–35.72]	12.03 [4.91–34.74]	10.55 [4.41–26.54]	7.91 [3.13–20.42]
RFH				
DSC	0.97 [0.88–0.97]	0.97 [0.89–0.98]	0.97 [0.92–0.98]	0.97 [0.95–0.98]
HD (mm)	15.34 [5.25–32.54]	16.73 [4.87–31.95]	8.20 [6.38–25.70]	10.62 [3.70–18.76]

Abbreviations: *CC: cervical cancer, PC: prostate cancer, LFH: Left femoral head, RFH: right femoral head, ST: STAPLE, RF: random forest, MV: MVision, LI: Limbus, DSC: dice similarity coefficient, HD: Hausdorff distance.*

Table 3
Statistical analysis: p-values from Wilcoxon-Mann-Whitney rank-sum test corrected by Bonferroni for multiple tests were reported. Bold: statistical significance.

CC	DSC				HD				**PC**	DSC				HD			
Bladder									**Bladder**								
	ST	RF	MV	LI	ST	RF	MV	LI		ST	RF	MV	LI	ST	RF	MV	LI
ST	/	0.240	**< 0.001**	**< 0.001**	/	1.000	**< 0.001**	**0.003**	ST	/	1.000	0.336	1.000	/	1.000	1.000	1.000
RF	/	/	**0.017**	0.240	/	/	0.180	1.000	RF	/	/	0.768	1.000	/	/	**0.029**	0.288
MV	/	/	/	1.000	/	/	/	1.000	MV	/	/	/	1.000	/	/	/	1.000
LI	/	/	/	/	/	/	/	/	LI	/	/	/	/	/	/	/	/
Rectum									**Rectum**								
	ST	RF	MV	LI	ST	RF	MV	LI		ST	RF	MV	LI	ST	RF	MV	LI
ST	/	**0.003**	**< 0.001**	**< 0.001**	/	0.240	0.300	**0.008**	ST	/	1.000	0.576	1.000	/	1.000	1.000	1.000
RF	/	/	0.240	0.300	/	/	1.000	1.000	RF	/	/	1.000	1.000	/	/	1.000	1.000
MV	/	/	/	1.000	/	/	/	1.000	MV	/	/	/	1.000	/	/	/	1.000
LI	/	/	/	/	/	/	/	/	LI	/	/	/	/	/	/	/	/
LFH									**LFH**								
	ST	RF	MV	LI	ST	RF	MV	LI		ST	RF	MV	LI	ST	RF	MV	LI
ST	/	1.000	1.000	1.000	/	1.000	0.720	0.960	ST	/	1.000	0.672	1.000	/	1.000	1.000	1.000
RF	/	/	1.000	0.180	/	/	0.120	0.240	RF	/	/	1.000	1.000	/	/	1.000	1.000
MV	/	/	/	1.000	/	/	/	1.000	MV	/	/	/	1.000	/	/	/	1.000
LI	/	/	/	/	/	/	/	/	LI	/	/	/	/	/	/	/	/
RFH									**RFH**								
	ST	RF	MV	LI	ST	RF	MV	LI		ST	RF	MV	LI	ST	RF	MV	LI
ST	/	0.120	1.000	1.000	/	1.000	1.000	1.000	ST	/	1.000	1.000	1.000	/	1.000	1.000	1.000
RF	/	/	1.000	0.780	/	/	0.480	0.180	RF	/	/	1.000	1.000	/	/	1.000	1.000
MV	/	/	/	1.000	/	/	/	1.000	MV	/	/	/	1.000	/	/	/	1.000
LI	/	/	/	/	/	/	/	/	LI	/	/	/	/	/	/	/	/
Bowel bag																	
	ST	RF	MV	LI	ST	RF	MV	LI									
ST	/	1.000	0.420	1.000	/	1.000	1.000	1.000									
RF	/	/	1.000	0.059	/	/	1.000	1.000									
MV	/	/	/	**0.003**	/	/	/	0.120									
LI	/	/	/	/	/	/	/	/									

Abbreviations: CC: cervical cancer, PC: prostate cancer, *ST: STAPLE, RF: random forest, MV: MVision, LI: Limbus*.

Table 4
Percentage of automated contours (testing set) standing 1 mm-, 3 mm-, and 5 mm-far from the manual Ground-Truth OARs. Median values are reported.

CC					PC				
DAP	ST	RF	MV	LI	**DAP**	ST	RF	MV	LI
Bladder					**Bladder**				
1 mm (%)	11.2	25.6	45.3	35.2	1 mm (%)	32.3	21.6	45.0	41.2
3 mm (%)	27.1	53.1	74.5	62.4	3 mm (%)	74.6	49.7	84.7	82.5
5 mm (%)	44.2	75.3	87.9	79.0	5 mm (%)	88.1	65.9	94.0	93.4
Rectum					**Rectum**				
1 mm (%)	10.9	21.0	29.4	24.4	1 mm (%)	27.4	31.5	33.9	32.3
3 mm (%)	26.7	44.1	47.4	48.2	3 mm (%)	57.4	64.7	56.1	56.9
5 mm (%)	44.2	61.8	59.2	66.4	5 mm (%)	76.5	77.9	62.5	66.3
Left femoral head					**Left femoral head**				
1 mm (%)	44.8	50.4	40.8	44.2	1 mm (%)	67.1	68.6	78.1	76.5
3 mm (%)	60.1	66.9	54.5	57.3	3 mm (%)	87.4	86.2	92.8	94.3
5 mm (%)	77.2	84.1	68.5	74.0	5 mm (%)	92.7	90.3	97.4	98.6
Right femoral head					**Right femoral head**				
1 mm (%)	47.3	54.3	38.2	45.0	1 mm (%)	70.9	72.0	79.8	73.8
3 mm (%)	59.9	70.1	51.3	61.8	3 mm (%)	87.4	85.6	93.1	91.9
5 mm (%)	75.9	84.4	66.1	78.4	5 mm (%)	92.2	90.3	97.9	96.8
Bowel bag									
1 mm (%)	14.7	15.9	17.1	8.5					
3 mm (%)	34.1	39.6	41.3	19.3					
5 mm (%)	51.2	55.3	58.3	30.3					

Abbreviations: CC: cervical cancer, PC: prostate cancer, DAP: Distance-to-Agreement Portion, *ST: STAPLE, RF: random forest, MV: MVision, LI: Limbus.*

performance. In particular, Limbus recorded a median DSC just below the institutional threshold having the shortest cranial extension.

Similarly, automated segmentation of the femoral heads was limited in the caudal direction to the last slice of manual segmentation at most: all systems registered similar and optimal performances in both cervical and prostate cancer cases.

It is worth noticing the maximum HD values registered in STAPLE performance in cervical cancer patients: a detailed observation led to the identification of the algorithm failure in the contouring of a single obese patient.

3.3. Qualitative clinical evaluation

An automated segmentation example in a cervical and prostate case is reported in Fig. 1. Results of the qualitative evaluation are reported in Fig. 2, globally confirming the quantitative results. DL outperformed STAPLE and Random Forest, showing small edits requirement and MVision registering the best scoring. Cervical cancer-expert radiation oncologists expressed the exact same score in 4 out of 20 evaluations (5 patients contoured with the four evaluated segmentation tools). This led to a poor agreement with a Cohen's Kappa coefficient of 0.2 (poor). However, scores were also gathered in a three-points scale: significant corrections are needed (score 1 or 2), edits required (score 3), no significant corrections (score 4 or 5). This scale reduction proved the existence of a better agreement with a Cohen's Kappa coefficient increase (0.7, good). Furthermore, prostate cancer DL excellent performances have been confirmed, with Limbus proving to have the most desirable contours by having a good agreement in 4 out of 5 cases, although STAPLE and Random Forest showed just small differences to be edited.

4. Discussion

In this study, four different auto-segmentation tools were evaluated in the OAR delineation in cervical cancer EBRT and prostate cancer SBRT covering from the earliest automated segmentation techniques to the more recent AI-based tools. A comprehensive evaluation has been performed including a qualitative assessment to assess automated segmentation tools for clinical use.

In cervical cancer RT, DL showed better results than the atlas-based model and Random Forest in terms of quantitative evaluation metrics and in shorter segmentation times per single structure sets. Previous studies reported segmentation duration shortening thanks to DL tools [15]. Gambacorta et al. evaluated segmentation times by means of automated delineation duration and time required by radiation oncologists to validate and correct automatic contours in prostate RT with 5 mm-thick CT slices [39]. A total segmentation time of about 25 min and 37 min was registered for 2 different automated segmentation tools based on Convolutional Neural Networks. In Table 1, reported data on manual correction times led to an estimation of global segmentation time of about 17 min for prostate SBRT with a slice thickness of 1 mm, suggesting a possible significant shortening of radiation oncologist's contouring burden. Even shorter DL times in cervical cancer sets with respect to prostate cancer sets could be explained by the lower number of slices dictated by the thicker slices. In particular, the MVision automated segmentation times were comparable to Limbus ones despite a higher number of structures to delineate: it has in fact been highlighted how MVision did not allow the selection of desired OARs. DL automated segmentation would shorten cervical cancer contouring duration from about 45 min to 10–15 min.

DL results are certainly due to the proven capabilities of innovative AI technology, thanks to recent technological development and their training on large datasets. Conversely, different studies showed that a wide preparatory dataset is not improving accuracy in atlas-based automated segmentation on CT images [24,29,30]. It is worth noticing that an outlier case in STAPLE automated contours has been found among cervical cancer patients and its addition to atlas Ground Truth dataset could be suggested to include an even larger atlas variability.

On the other hand, model training for image classification with Random Forest would require training datasets of at least 50–100 cases, limiting the here presented performances [22,40]. The preparatory set of only 20 cases could explain the limited performances in female pelvis, comparable with STAPLE and significantly lower than DL algorithms.

It is worth highlighting that performance differences taper off in prostate cancer contouring. This may be due to the thinner slices leading to better cranio-caudal resolution. Furthermore, prostate cancer patients followed a rigid rectum and bladder preparation before image acquisition giving a strongly reduced OAR variability in filling and shapes. Despite preparatory dataset limitations, all these aspects could help STAPLE and Random Forest correctly defining filled OARs, closing the gap with DL capabilities.

Comparing the results presented here with the recent literature, similar STAPLE DSC values were found in a previous study on a mono-institutional cohort of 21 prostate cancer patients undergoing a similar rectum and bladder preparation [41]. Moreover, DL widely

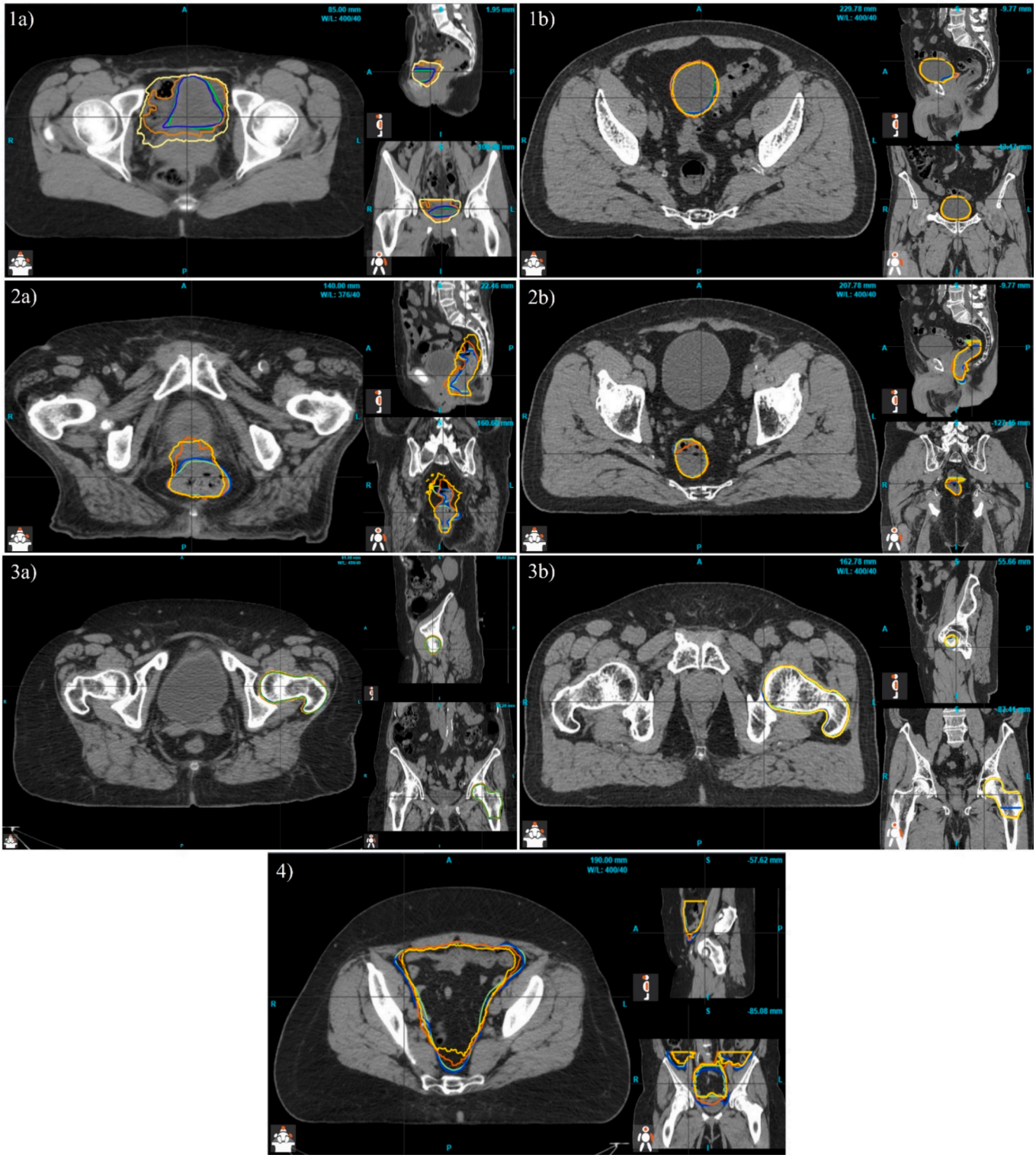

Fig. 1. A comparison example of cervical cancer (a) and prostate cancer (b) automated segmentation of bladder (1), rectum (2), and left femoral head (3). An example of bowel bag segmentation is reported (4). Legend: *Yellow = STAPLE, Orange = Random forest, Blue = MVision, Green = Limbus.* (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

demonstrated its segmentation capabilities on MRI images achieving good results for prostate RT [42–44] and several studies obtained good results in prostate and OAR automated segmentation through UNet algorithms, achieving up to 0.95 and 0.85 DSC values in bladder and rectum, respectively [25]. In very recent studies, two other DL commercial algorithms reached DSC values over 0.90 for both bladder and rectum in prostate radiotherapy on CT images [27,28] and similar DSC results have been achieved by a homemade combination of a multichannel 2D U-Net followed by a 3D U-Net [45]: 0.84 and 0.95 for bladder and rectum, respectively.

In addition, a recent study on cervical cancer segmentation was performed by training a homemade DL-algorithm by selecting 104

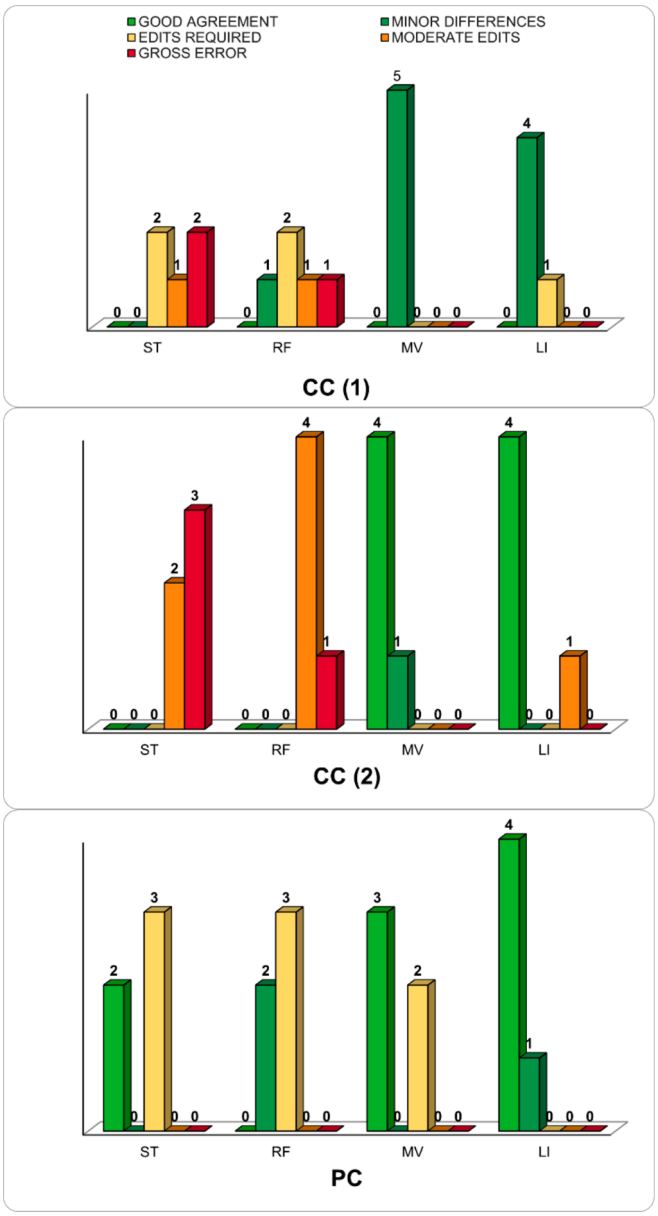

Fig. 2. Qualitative evaluation of automatic contours performed by two cervical cancer experts and a prostate cancer expert. Abbreviations: *CC(1): first cervical cancer expert, CC(2): second cervical cancer expert, PC: prostate cancer expert, ST: Staple, RF: random forest, MV: MVision, LI: Limbus.*

cervical cancer cases [46]. High DSC values have been reached for bladder (0.83) and rectum (0.82). Liu et al. obtained higher bladder DSC values (0.90) exploiting a Convolutional Neural Network architecture for OAR segmentation in locally advanced cervical cancer RT [47].

As already known, bone delineation is an easier task considering the soft tissue-bone interface contrast. Femoral heads automated segmentation succeeded confirming excellent results presented by Wang et al. [48]. The remaining critical point in femoral heads delineation is their caudal extension which may influence the already high DSC value. On the other hand, performances in bowel bag contouring were mainly affected by its cranial extension depending on the extent of training images and the treatment demands, i.e., the target cranial extension. Bowel bag caudal definition is instead affected by rectum and bladder delineation. The bowel bag registered values were comparable with the median DSC median of 0.88 obtained with a Convolutional Neural Network-based tool [49].

It is thus worth noticing that automated segmentation evaluation is strongly dependent on institution specific requirements, i.e., the chosen Ground-Truth, the Institution preparation protocol, and the target extension. From this point of view, atlas-based models benefit from homemade training even if affected by a smaller preparatory sample. In contrast, commercial DLs depend on a dedicated training set, often commercially sealed. All these aspects converge affecting the radiation oncologists' manual correction times. In other words, radiation oncologists will also have to delete delineations in not useful slices spending more time than that required to correct any inaccuracies in the slices of interest.

The last part of this study focused on clinical qualitative evaluation resulting in a globally positive score, especially for automatic contours in prostate cancer cases.

In male pelvis, a four or a five degrees-scale was often used to run an automated segmentation qualitative evaluation [17,50–52]. Gibbons et al. performed a four point-qualitative scale evaluation then grouping them into two categories of clinical and rejectable automatic contours [17]. They observed DL OARs clinical acceptable contours in 81.7 % of pelvic OARs automatic contours. Huyskens et al., for example, proposed to evaluate the number of slices to be corrected in 39 auto-segmented prostate cases: rectum delineation resulted in a not acceptable score in 45 % of the cases [51]. To compare those different four level-scales with the here reported qualitative evaluation scores, it is possible to group scores 4 and 5 as unacceptable. In this case, all here analyzed algorithms resulted in acceptable contours in 100 % of prostate cases. Conversely, a five-level scale to assess the clinical use of DL automatic contours in male pelvis RT has been used by Duan et al. Considering level 3 and above as clinically acceptable, they found that the percentage of automatic contours that has been scored equal to or better than the manual segmentation exceeded 50 % [52].

The here reported optimal scores could be explained in terms of organs filling preparation, as already mentioned in the discussion of quantitative metrics. It is worth noticing that the reduced number of available expert ROs may offer a not largely-shared view on the qualitative contour evaluation. Furthermore, this prevents a robust preliminary analysis of inter-operator variability in manual Ground Truth.

In conclusion, this broad comparison of the earliest automated segmentation and the actual AI-based tools showed how clinically advantageous the use of DL-based models can be for CT images automated segmentation. Especially in female pelvis, DL registered better quantitative results, shorter times and higher QE scores with the expected highest scores in high-contrast bony structures. On the other hand, in male pelvis, a strict filling preparation strongly reduced inter-patient variability helping traditional algorithms getting very good performances.

These excellent performance and significantly reduced segmentation times would allow for faster RT workflow and high-quality treatments.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Hindocha S, Zucker K, Jena R, Banfill K, Mackay K, Price G, et al. Artificial intelligence for radiotherapy auto-contouring: current use, perceptions of and barriers to implementation. Clin Oncol 2023;4:219–26. https://doi.org/10.1016/j.clon.2023.01.014.

[2] Cusumano D, Boldrini L, Dhont J, Fiorino C, Green O, Güngör G, et al. Artificial Intelligence in magnetic Resonance guided Radiotherapy: Medical and physical considerations on state of art and future perspectives. Phys Med 2021;85:175–91. https://doi.org/10.1016/j.ejmp.2021.05.010.

[3] Chung S, Chang J, Kim Y. Comprehensive clinical evaluation of deep learning-based auto-segmentation for radiotherapy in patients with cervical cancer. Front Oncol 2023;13:1119008. https://doi.org/10.3389/fonc.2023.1119008.

[4] Mukesh M, Benson R, Jena R, Hoole A, Roques T, Scrase C, et al. Interobserver variation in clinical target volume and organs at risk segmentation in post-parotidectomy radiotherapy: can segmentation protocols help? Br J Radiol 2012; 85:e530–6. https://doi.org/10.1259/bjr/66693547.

[5] Perna L, Cozzarini C, Maggiulli E, Fellin G, Rancati T, Valdagni R, et al. Inter-observer variability in contouring the penile bulb on CT images for prostate cancer treatment. Radiat Oncol 2011;6:123. https://doi.org/10.1186/1748-717x-6-123.

[6] Lin D, Lapen K, Sherer MV, Kantor J, Zhang Z, Boyce LM, et al. A systematic review of contouring guidelines in radiation oncology: analysis of frequency, methodology, and delivery of consensus recommendations. Int J Radiat Oncol Biol Phys 2020;107:827–35. https://doi.org/10.1016/j.ijrobp.2020.04.011.

[7] Cox S, Cleves A, Clementel E, Miles E, Staffurth J, Gwynne S. Impact of deviations in target volume delineation – Time for a new RTQA approach? Radiother Oncol 2019;137:1–8. https://doi.org/10.1016/j.radonc.2019.04.012.

[8] Liu X, Qu L, Xie Z, Zhao J, Shi Y, Song Z. Towards more precise automatic analysis: a systematic review of deep learning-based multi-organ segmentation. Biomed Eng Online 2024;23(52). https://doi.org/10.1186/s12938-024-01238-8.

[9] Tong N, Xu Y, Zhang J, Gou S, Li M. Robust and efficient abdominal CT segmentation using shape constrained multi-scale attention network. Phys Med 2023;110:102595. https://doi.org/10.1016/j.ejmp.2023.102595.

[10] Chukwujindu E, Faiz H, AI-Douri S, Faiz K, De Sequeira A. Role of artificial intelligence in brain tumor imaging. Eur J Radiol 2024;176:111509. https://doi.org/10.1016/j.ejrad.2024.111509.

[11] Vandewinckele L, Claessens M, Dinkla A, Brouwer C, Verellen D, et al. Overview of artificial intelligence-based applications in radiotherapy: Recommendations for implementation and quality assurance. Radiother Oncol 2020;153:55–66. https://doi.org/10.1016/j.radonc.2020.09.008.

[12] Mackay K, Bernstein D, Glocker B, Kamnitsas K, Taylor A. A review of the metrics used to assess auto-contouring systems in radiotherapy. Clin Oncol 2023;35: 354–69. https://doi.org/10.1016/j.clon.2023.01.016.

[13] Coffey A, Moreno J, Lenards N, Hunzeker A, Tobler M. A survey of medical dosimetrists' perceptions of efficiency and consistency of auto-contouring software. Med Dosim 2022;47:312–7. https://doi.org/10.1016/j.meddos.2022.05.003.

[14] Nikolov S, Blackwell S, Zverovitch A, Mendes R, Livne M, De Fauw J, et al. Clinically applicable segmentation of head and neck anatomy for radiotherapy: deep learning algorithm development and validation study. J Med Internet Res 2021;23(7):e26151. https://doi.org/10.2196/26151.

[15] Byun H, Chang J, Choi M, Chun J, Jung J, Jeong C, et al. Evaluation of deep learning-based autosegmentation in breast cancer radiotherapy. Radiat Oncol 2021;16:203. https://doi.org/10.1186/s13014-021-01923-1.

[16] Strolin S, Santoro M, Paolani G, Ammendolia I, Arcelli A, Benini A, et al. How smart is artificial intelligence in organs delineation? Testing a CE and FDA-approved Deep-Learning tool using multiple expert contours delineated on planning CT images. Front Oncol 2023;13:1089807. https://doi.org/10.3389/fonc.2023.1089807.

[17] Gibbons E, Hoffmann M, Westhuyzen J, Hodgson A, Chick B, Last A. Clinical evaluation of deep learning and atlas-based autosegmentation for critical organs at risk in radiation therapy. J Med Radiat Sci 2023;2022(70):15–25. https://doi.org/10.1002/jmrs.618.

[18] Rhee D, Jhingran A, Rigaud B, Netherton T, Cardenas CE, Zhang L, et al. Automatic contouring system for cervical cancer using convolutional neural networks. Med Phys 2020;47:5648–58. https://doi.org/10.1002/mp.14467.

[19] Nachbar M, Lo Russo M, Gani C, Boeke S, Wegener D, Paulsen F, et al. Automatic AI-based contouring of prostate MRI for online adaptive radiotherapy. J Med Phys 2023;23. https://doi.org/10.1016/j.zemedi.2023.05.001.

[20] Rayn K, Gokhroo G, Jeffers B, Gupta V, Chaudhari S, Clark R, et al. Multicenter study of pelvic nodal autosegmentation algorithm of Siemens healthineers: comparison of male versus female pelvis. Adv Radiat Oncol 2023;000:101326. https://doi.org/10.1016/j.adro.2023.101326.

[21] Kim N, Chang JS, Kim YB, Kim JS. Atlas-based auto-segmentation for postoperative radiotherapy planning in endometrial and cervical cancers. Radiat Oncol 2020;15: 106. https://doi.org/10.1186/s13014-020-01562-y.

[22] Casati M, Piffer S, Calusi S, Marrazzo L, Simontacchi G, Di Cataldo V, et al. Methodological approach to create an atlas using a commercial auto-contouring software. J Appl Clin Med Phys 2020;21:219–30. https://doi.org/10.1002/acm2.13093.

[23] Casati M, Piffer S, Calusi S, Marrazzo L, Simontacchi G, Di Cataldo V, et al. Clinical validation of an automatic atlas-based segmentation tool for male pelvis CT images. J Appl Clin Med Phys 2022;23:e13507. https://doi.org/10.1002/acm2.13507.

[24] Greenham S, Dean J, Fu CKK, Goman J, Mulligan J, Tune D, et al. Evaluation of atlas-based auto-segmentation software in prostate cancer patients. J Med Radiat Sci 2014;61:151–8. https://doi.org/10.1002/jmrs.64.

[25] Duan J, Bernard M, Downes L, Willows B, Feng X, Mourad WF, et al. Evaluating the clinical acceptability of deep learning contours of prostate and organs-at-risk in an automated prostate treatment planning process. Med Phys 2022;49:2570–81. https://doi.org/10.1002/mp.15525.

[26] Wang Y, Boyd G, Zieminski S, Kamran SC, Zietman AL, Miyamoto DT, et al. A pair of deep learning auto-contouring models for prostate cancer patients injected with a radio-transparent versus radiopaque hydrogel spacer. Med Phys 2023;50: 3324–37. https://doi.org/10.1002/mp.16375.

[27] Palazzo G, Mangili P, Deantoni C, Fodor A, Broggi S, Castriconi R, et al. Real-world validation of Artificial Intelligence-based Computed Tomography auto-contouring for prostate cancer radiotherapy planning. Phys Imaging Radiat Oncol 2023;28: 100501. https://doi.org/10.1016/j.phro.2023.100501.

[28] De Kerf G, Claessens M, Raouassi F, Mercier C, Stas D, Ost P, et al. A geometry and dose-volume based performance monitoring of artificial intelligence models in radiotherapy treatment planning for prostate cancer. Phys Imaging Radiat Oncol 2023;28:100494. https://doi.org/10.1016/j.phro.2023.100494.

[29] Wong WKH, Leung LHT, Kwong DLW. Evaluation and optimization of the parameters used in multiple-atlas-based segmentation of prostate cancers in radiation therapy. Br J Radiol 2015;89:20140732. https://doi.org/10.1259/bjr.20140732.

[30] Li Y, Wu W, Sun Y, Yu D, Zhang Y, Wang L, et al. The clinical evaluation of atlas-based auto-segmentation for automatic contouring during cervical cancer radiotherapy. Front Oncol 2022;12:945053. https://doi.org/10.3389/fonc.2022.945053.

[31] Francis S, Jayaraj PB, Pournami PN, Puzhakkal N, et al. ContourGAN: Auto-contouring of organs at risk in abdomen computed tomography images using generative adversarial network. Wiley Periodicals LLC. Int J Imaging Syst Technol 2023;33:1494–504. https://doi.org/10.1002/ima.22901.

[32] Suresh R, Niemela J, Akram S, Valdman A, Olsson CE. A comparative study between AI-generated, real-life clinical as well as reference rectal volumes defined in accordance with the Swedish National STRONG Guidelines in Prostate Cancer Radiotherapy. Int J Radiat Oncol 2021;111:e138. https://doi.org/10.1016/j.ijrobp.2021.07.579.

[33] Malhotra P, Gupta S, Koundal D, Zaguia A, Enbeyle W. Deep neural networks for medical image segmentation. Hindawi J Healthc Eng 2022. https://doi.org/10.1155/2022/9580991.

[34] Minaee S, Boykov Y, Porikli F, Plaza A, Kehtarnavaz N, Terzopoulos D. Image segmentation using deep learning: A survey. IEEE Trans Pattern Anal Mach Intell 2022;44:3523–42. https://doi.org/10.1109/TPAMI.2021.3059968.

[35] Hesamian MH, Jia W, He X, Kennedy P. Deep learning techniques for medical image segmentation: achievements and challenges. J Digit Imaging 2019;32: 582–96. https://doi.org/10.1007/s10278-019-00227-x.

[36] Zhang Y, Paulson E, Lim S, Hall WA, Ahunbay E, Mickevicius NJ, et al. A patient-specific autosegmentation strategy using multi-input deformable image registration for magnetic resonance imaging - guided online adaptive radiation therapy: A feasibility study. Adv Radiat Oncol 2020;5:1350–8. https://doi.org/10.1016/j.adro.2020.04.027.

[37] Vassen F, Hazelaar C, Vaniqui A, Gooding M, van der Heyden B, Canters R, et al. Evaluation of measures for assessing time-saving of automatic organ-at-risk segmentation in radiotherapy. Phys Imaging Radiat Oncol 2020;13:1–6. https://doi.org/10.1016/j.phro.2019.12.001.

[38] Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159–74. https://doi.org/10.2307/2529310.

[39] Gambacorta MA, Valentini C, Dinapoli N, Boldrini L, Caria N, Barba MC, et al. Clinical validation of atlas-based auto-segmentation of pelvic volumes and normal tissue in rectal tumors using auto-segmentation computed system. Acta Oncol 2013;52:1676–81. https://doi.org/10.3109/0284186x.2012.754989.

[40] Macomber MW, Phillips M, Tarapov I, Jena R, Nori A, Carter D, et al. Autosegmentation of prostate anatomy for radiation treatment planning using deep decision forests of radiomic features. Phys Med Biol 2018;63:235002. https://doi.org/10.1088/1361-6560/aaeaa4.

[41] Urago Y, Okamoto H, Kaneda T, Murakami N, Kashihara T, Takemori M, et al. Evaluation of auto-segmentation accuracy of cloud-based artificial intelligence and atlas-based models. Radiat Oncol 2021;16:175. https://doi.org/10.1186/s13014-021-01896-1.

[42] Tian Z, Liu L, Zhang Z, Fei B. PSNet: prostate segmentation on MRI based on a convolutional neural network. J Med Imaging 2018;55(2):021208. https://doi.org/10.1117/1.JMI.5.2.021208.

[43] Vagni M, Tran HE, Romano A, Chiloiro G, Boldrini L, Zormpas-Petridis K, et al. Auto-segmentation of pelvic organs at risk on 0.35T MRI using 2D and 3D Generative Adversarial Network models. Phys Med 2024;119:103297. https://doi.org/10.1016/j.ejmp.2024.103297.

[44] Kawula M, Hadi I, Nierer L, Vagni M, Cusumano D, Boldrini L, et al. Patient-specific transfer learning for auto-segmentation in adaptive 0.35 T MRgRT of prostate cancer: a bi-centric evaluation. Med Phys 2023;50(1573–1585). https://doi.org/10.1002/mp.16056.

[45] Balagopal A, Kazemifar S, Nguyen D, Lin M, Hannan R, Owrangi A, et al. Fully automated organ segmentation in male pelvic CT images. Phys Med Biol 2018;63 (24):245015. https://doi.org/10.1088/1361-6560/aaf11c.

[46] Liu Z, Liu X, Guan H, Zhen H, Sun Y, Chen Q, et al. Development and validation of a deep learning algorithm for auto-delineation of clinical target volume and organs at risk in cervical cancer radiotherapy. Radiother Oncol 2020;153:172–9. https://doi.org/10.1016/j.radonc.2020.09.060.

[47] Liu Z, Liu X, Xiao B, Miao Z, Sun Y, Zhang F, et al. Segmentation of organs-at-risk in cervical cancer CT images with a convolutional neural network. Phys Med 2020; 69:184–91. https://doi.org/10.1016/j.ejmp.2019.12.008.

[48] Wang Z, Chang Y, Peng Z, Lv Y, Shi W, Wang F, et al. Evaluation of deep learning-based autosegmentation algorithms for delineating clinical target volume and organs at risk involving data for 125 cervical cancer patients. J Appl Clin Med Phys 2020;21:272–9. https://doi.org/10.1002/acm2.13097.

[49] Sartor H, Minarik D, Enqvist O, Ulén J, Wittrup A, Bjurberg M, et al. Auto-segmentations by convolutional neural network in cervical and anorectal cancer with clinical structure sets as the ground truth. Clin Transl Radiat Oncol 2020;25: 37–45. https://doi.org/10.1016/j.ctro.2020.09.004.

[50] Hoque SMH, Pirrone G, Matrone F, Donofrio A, Fanetti G, Caroli A, et al. Clinical use of a commercial artificial intelligence-based software for autocontouring in

radiation therapy: geometric performance and dosimetric impact. Cancers 2023; 15:5735. https://doi.org/10.3390/cancers15245735.

[51] Huyskens DP, Maingon P, Vanuytsel L, Remouchamps V, Roques T, Dubray B, et al. A qualitative and a quantitative analysis of an auto-segmentation module for prostate cancer. Radiother Oncol 2009;90:337–45. https://doi.org/10.1016/j.radonc.2008.08.007.

[52] Duan J, Vargas CE, Yu NY, Laughlin BS, Toesca DS, Keole S, et al. Incremental retraining, clinical implementation, and acceptance rate of deep learning auto-segmentation for male pelvis in a multiuser environment. Med Phys 2023;50: 4079–91. https://doi.org/10.1002/mp.16537.