Comparison between artificial intelligence-based and manual organ delineations in pretreatment computed tomography scans of prostate cancer

Radiation Protection Dosimetry, 2026, **202(3–4)**, 204–213
https://doi.org/10.1093/rpd/ncaf184
Paper

Comparison between artificial intelligence-based and manual organ delineations in pretreatment computed tomography scans of prostate cancer patients: a visual grading study

Eirini Polymeri[1,2,*], Åse A. Johnsson[1,2], Olof Enqvist[3,4], Johannes Ulén[4], Jon Kindblom[5], Karin Braide[6], Hans-Jurgen Wiltz[7], Margareta Tanyasiová[5], Elin Trägårdh[8], Lars Edenbrandt[9], Henrik Kjölhede[10,11], Angelica Svalkvist[12,13]

[1]Department of Radiology, Institute of Clinical Sciences, Sahlgrenska Academy, University of Gothenburg, Medicinaregatan 3, Medicinareberget, 413 90, Gothenburg, Region Västra Götaland, Sweden
[2]Department of Radiology, Sahlgrenska University Hospital, Blå Stråket 5, Region Västra Götaland, 413 45, Gothenburg, Sweden
[3]Department of Electrical Engineering, Chalmers University of Technology, Hörsalsvägen 9-11, Region Västra Götaland, 412 96, Gothenburg, Sweden
[4]Eigenvision AB, Bredgatan 4, 211 30, Region Skåne, Malmö, Sweden
[5]Department of Oncology, Sahlgrenska University Hospital, Blå Stråket 2, Region Västra Götaland, 413 45, Gothenburg, Sweden
[6]Department of Oncology, Institute of Clinical Sciences, University of Gothenburg, Medicinaregatan 3, Medicinareberget, 413 90, Gothenburg, Region Västra Götaland, Sweden
[7]Department of Oncology and Radiotherapy, Region Kronoberg, Central lasarett Växjö, Strandvägen 8, 351 85, Växjö, Sweden
[8]Clinical Physiology and Nuclear Medicine, Lund University and Skåne University Hospital, 221 85, Malmö, Region Skåne, Sweden
[9]Department of Molecular and Clinical Medicine, Institute of Medicine, Sahlgrenska Academy, University of Gothenburg, Blå Stråket 5, 413 45, Region Västra Götaland, Gothenburg, Sweden
[10]Department of Urology, Institute of Clinical Sciences, Sahlgrenska Academy, University of Gothenburg, Medicinaregatan 5, Medicinareberget, 413 90, Region Västra Götaland, Gothenburg, Sweden
[11]Department of Urology, Sahlgrenska University Hospital, Blå Stråket 5, 413 45, Region Västra Götaland, Gothenburg, Sweden
[12]Department of Medical Radiation Sciences, Institute of Clinical Sciences, Sahlgrenska Academy, University of Gothenburg, Medicinaregatan 5, Medicinareberget, 413 90, Region Västra Götaland, Gothenburg, Sweden
[13]Department of Biomedical Engineering and Medical Physics, Sahlgrenska University Hospital, Blå Stråket 5, 413 45, Region Västra Götaland, Gothenburg, Sweden

*Corresponding author. Institute of Clinical Sciences at Sahlgrenska Academy, Department of Radiology, Sahlgrenska University Hospital, Blå Stråket 5, SE-413 45 Gothenburg, Sweden. E-mail: eirini.polymeri@gu.se

Abstract

This study aimed to evaluate the clinical acceptability of artificial intelligence (AI)-based organ segmentations on pretreatment CT images of prostate cancer patients using manual organ delineations as a reference. Paired AI-based segmentations and manual delineations of the prostate, urinary bladder, and rectum were evaluated by three observers, according to a 4-grade Likert-scale, based on quality criteria, developed through a Delphi process. Visual grading characteristics (VGC) analysis was performed. When comparing the ratings of AI-based ($n = 360$) and manual delineations ($n = 360$), the area under the VGC-curve (AUC_{VGC}) was 0.36 (95% CI 0.27–0.44), 0.35 (95% CI 0.28–0.41), and 0.3 (95% CI 0.22–0.40) for the prostate, urinary bladder, and rectum, respectively, indicating inferior ratings for the algorithm. Few AI segmentations (8%) were considered clinically unacceptable, while in 67% no or minor changes were needed. Despite superior ratings for manual delineations, most AI-segmentations needed no or minor changes, indicating clinical acceptability.

Received: June 19, 2025. **Revised:** December 2, 2025. **Accepted:** December 10, 2025

Introduction

Subjective image interpretation is one of the main sources of variability in the visual assessment of medical images [1, 2]. Organ delineation on images from computed tomography (CT) or magnetic resonance imaging (MRI) is a necessary procedure for radiation treatment planning of tumours. Several studies have shown that inaccurate organ delineation can lead to poorer clinical outcomes [3–5]. However, this task is highly observer-dependent and subject to inter- and intraobserver variability [6–8]. Inter- and intraobserver variability in organ delineation for radiation treatment planning are multifactorial in origin. Several possible solutions have been introduced, including standardized protocols, guidelines, and multimodal imaging [9]. However, inter- and intraobserver variability remain a challenge.

The development of artificial intelligence (AI) and its increasing use in oncological imaging and radiation oncology have had a dramatic impact in the past decade [10]. Several studies have shown the applicability of AI as a tool in organ delineations for radiation treatment planning of various forms of cancer, including prostate cancer (PCa) [11–15]. However, machine learning has been applied almost exclusively for prostate delineation on MRI for PCa diagnosis and staging, while its application to CT has been limited [16]. Moreover, the focus has been largely placed on the AI-based segmentation of target organs as well as organs at risk (OAR), which has reduced both manual delineation time and interobserver variability [12, 17]. Nevertheless, the quality and clinical acceptability of automated AI-based organ segmentation have not been adequately explored in PCa patients.

Most studies on AI-based organ segmentations have focused on geometric measurements, such as Hausdorff distance (HD), mean surface distance (MSD), and Dice–Sørensen Coefficient (DSC), showing good agreement between AI and the ground truth [18–24]. Yet, these metrics can lead to ambiguity regarding the clinical utility of the AI-based segmentations as they are strongly dependent on the ground truth. In addition, agreement between geometric measurements does not reflect clinical acceptability.

In 2007, Båth and Månsson [25] described the use of visual grading characteristics (VGC) analysis as a method to compare image quality using ordinal data. To date, this methodology has primarily been used for image quality assessment [26–29]. To the best of our knowledge, visual grading studies evaluating the quality and potential clinical use of AI-based organ segmentations for the radiotherapy planning of PCa patients are limited [22, 30–32].

Consequently, this study aimed to determine whether an AI algorithm can achieve clinically acceptable segmentations of the prostate and surrounding risk organs (OAR) on pretreatment CT scans using visual grading.

Materials and methods

Patient and image material

The study was approved by the Swedish Ethical Review Authority (2019-03205).

The CT scans of the patients included in the study were derived from a larger study cohort, which has been described in detail previously, along with the acquisition parameters used for the CT examinations [24]. The large dataset of CT scans comprised 1530 PCa patients at the Department of Oncology, Sahlgrenska University Hospital, Gothenburg, Sweden. Imaging was performed before radiation treatment planning between 2006 and 2018 [24].

The AI algorithm used in the present study is a U-NET-based model and was developed by mathematicians for research purposes only. The method is not used in the clinical routine, and it was used for organ segmentation in CT images within a research framework described in detail in a previous study [24]. The algorithm was trained and validated on manual delineations of the image material described above, in order to create a test set of automated AI-based organ segmentations (prostate $n = 329$, urinary bladder $n = 335$, rectum $n = 175$) (Fig. 1) [24].

All manual delineations underwent quality control prior to inclusion in the study. The AI-based segmentations of the test set were compared with the corresponding manual delineations to calculate Dice Sørensen Coefficient (DSC) for each organ, which has been reported previously [24].

AI-based segmentations were divided into four categories according to DSC values: <0.7, 0.7–0.79, 0.8–0.89, and >0.9, describing the range of algorithm's performance across organs. From this overall distribution, 120 organ segmentations (40 of each organ) were then randomly collected and included in the present study, to mimic the DSC distribution in the larger cohort, ensuring that both well-performing and lower-performing segmentations were represented. There were no rectum and urinary bladder segmentations with DSC < 0.7, while there were few prostate segmentations with DSC < 0.7 or > 0.9 (Fig. 1). For the visual grading, the assessments of the AI-based segmentations and their corresponding manual delineations were obtained for each organ (Fig. 1).

Presentation of study data

The research platform "Recomia" (www.recomia.org) [33], a noncommercial and nonprofit program, was

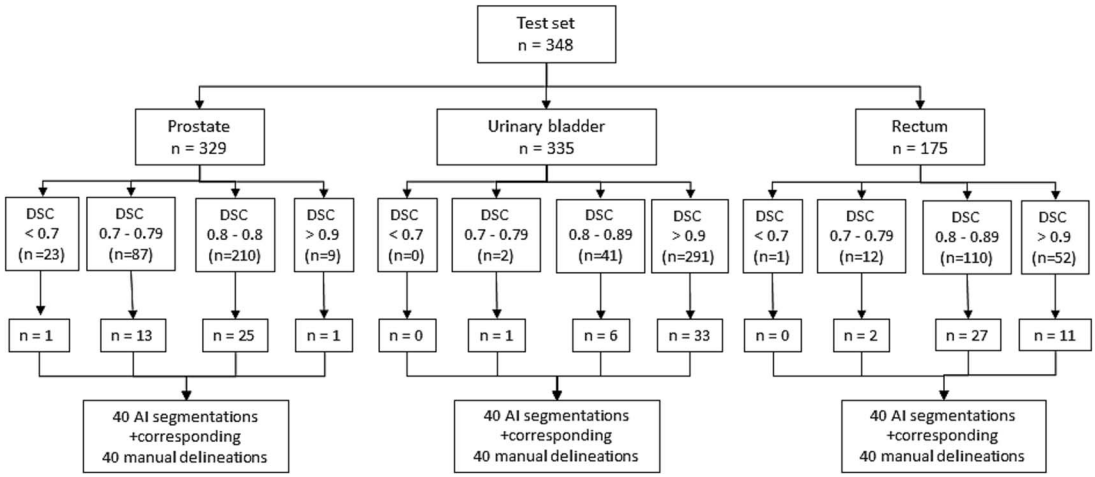

Figure 1. Flow chart of the collection of cases for visual grading, where the test set was a part of a previous study [24].

used for the visual evaluation of all organ segmentations. The AI algorithm segments the prostate and seminal vesicles with two separate labels and presenting these two labels to the observers would be a trivial way for them to determine if a delineation was AI-based and could lead to an unwanted bias. To ensure that manual and AI-based segmentations were presented in the same way to the observers for visual grading, all segmentations included the prostate gland and excluded the seminal vesicles, in accordance with the clinical delineation protocol used for the test dataset. If the manual prostate delineation for a case included the seminal vesicles, the AI-based segmentations of the prostate and seminal vesicles were merged into one label. If the manual prostate delineation for a case did not include the seminal vesicles, the AI-based segmentation of the seminal vesicles was not made visible to the observers. Consequently, the exclusion or the inclusion of seminal vesicles was consistent across AI and manual contours, minimizing potential bias in visual grading. The AI-based segmentations of the urinary bladder and rectum were presented without intervention.

Visual assessment of AI-segmentation's clinical acceptability

All information related to the organ segmentations was blinded. The AI segmentations and manual delineations were presented separately and in a randomized order to each observer. The observers rated the quality of each segmentation independently using a 4-grade Likert scale, as shown in Table 1. The rating scale was developed through a Delphi procedure, which is a well-established method for reaching consensus [34]. The group who contributed to this procedure consisted of eight specialists: two radiologists, one urologist, three radiation oncologists, one nuclear medicine physician, and one medical physicist. For the rectum, the criteria of acceptance included delineations up to the rectosigmoid junction. For the prostate, cases with and without delineations of the vesicles were included.

Table 1. Rating scale of clinical acceptance of organ segmentations for each organ.

Grading	Can the organ segmentation be clinically accepted/used for radiation treatment planning?
1.	Segmentation can be accepted/used for treatment planning
2.	No, would perform minor changes
3.	No, would perform major changes
4.	No, would delete and perform a new delineation

Observers

Three radiation oncologists who work at different hospitals in the country participated in the study. Observer 1 (O1) is a specialist in urology and radiation oncology with 12 years of experience (K.B.), and Observer 2 (O2) and Observer 3 (O3) are radiation oncologists with 20 and 12 years of experience, respectively (M.T. and H.-J.W.). Detailed information on how to use the research platform was provided to the observers before the study. Three consensus meetings were also held prior to the study, to obtain a clear understanding of radiation treatment planning guidelines and interpretation of ratings, e.g. acceptance criteria regarding the delineation limits of the rectum. Before evaluating the segmentations included in the study cohort, the observers practised on a separate training material consisting of 10 cases per organ, analogous to those

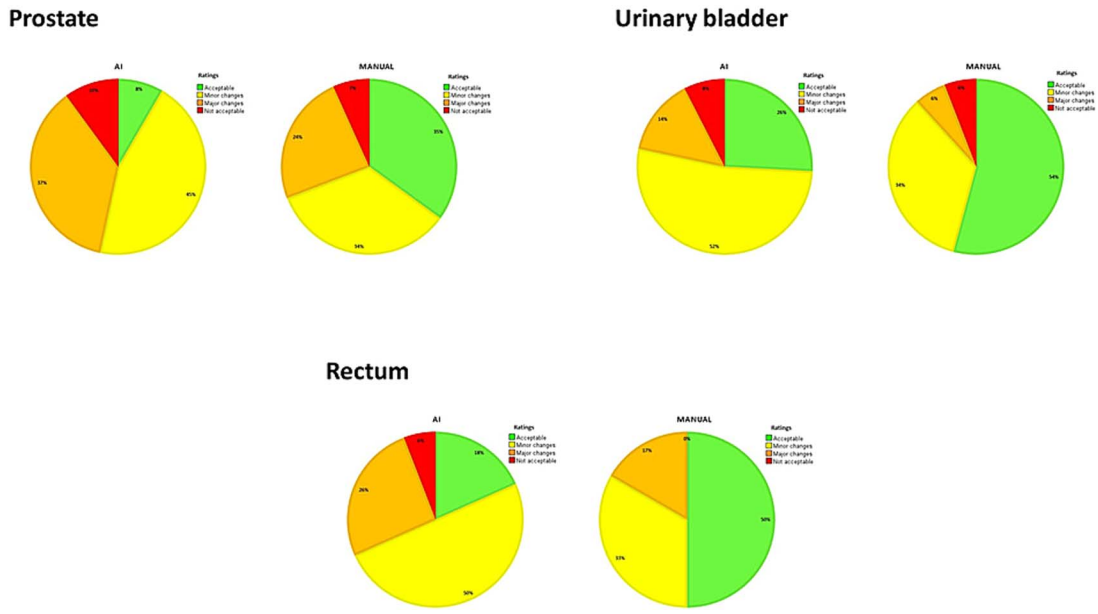

Figure 2. Pie charts showing the distribution of the visual grading of the observers for the 40 AI-based segmentations and their corresponding 40 manual delineations of each organ.

included in the actual study, and assessed according to the same procedure described above.

Statistical analysis

Statistical analysis was performed using VGC analysis, a nonparametric, rank-invariant statistical method for analysing visual grading data [25]. In VGC analysis, paired ratings of organ delineation quality are compared by plotting the AI-based segmentation ratings against the manual delineation ratings to create a VGC curve. The separation between the ratings of manual and automatic segmentations is determined by the area under the VGC curve (AUC_{VGC}), where an AUC_{VCG} of 0.5 represents no difference between the quality of the delineations, an AUC_{VGC} <0.5 represents higher ratings for the manual delineations, and an AUC_{VGC} > 0.5 represents higher ratings for the AI-based segmentations. The software VGC Analyzer [35] was used to determine AUC_{VGC} and statistically analyse the results. Using VGC Analyzer, the AUC_{VGC} can be determined using both the trapezoidal rule and binormal curve fitting. The statistical analysis can be performed for both paired and nonpaired data, and the results are presented for both the fixed reader situation (results applicable to the observers in the study) and the random reader situation (results applicable to a general population of observers).

To study the correlation between observer ratings for the AI-based segmentations of each organ and DSC, Spearman's correlation coefficient was analysed using SPSS Statistics 29 (IBM). The level of significance was set to values outside the range of the 95% confidence interval ($P \leq .05$).

Results

In total, 120 AI-based organ segmentations and their corresponding manual delineations were evaluated by the observers. In the VGC analysis, the trapezoidal rule for curve fitting was used. Due to the small number of observers, the analysis was based on the fixed reader situation.

The ratings by the observers showed that a large proportion of all AI-based organ segmentations were acceptable for clinical use with no or minor modifications (67%), which corresponded to 53%, 78%, and 68% of the prostate, urinary bladder, and rectum segmentations, respectively. The corresponding results for the manual delineations were 69%, 88%, and 83%, respectively. Further, 37%, 14%, and 26% of the AI-based segmentations of the prostate, urinary bladder, and rectum, respectively, were assessed as acceptable with major changes, compared with 24%, 6%, and 17% of the manual delineations of the corresponding organs. A small number of delineations were evaluated as not acceptable, which corresponded to 10%, 8%, and 6% of the AI-based segmentations of the prostate, urinary bladder, and rectum, respectively, and 7%, 6%, and 0% of the manual delineations of the corresponding organs

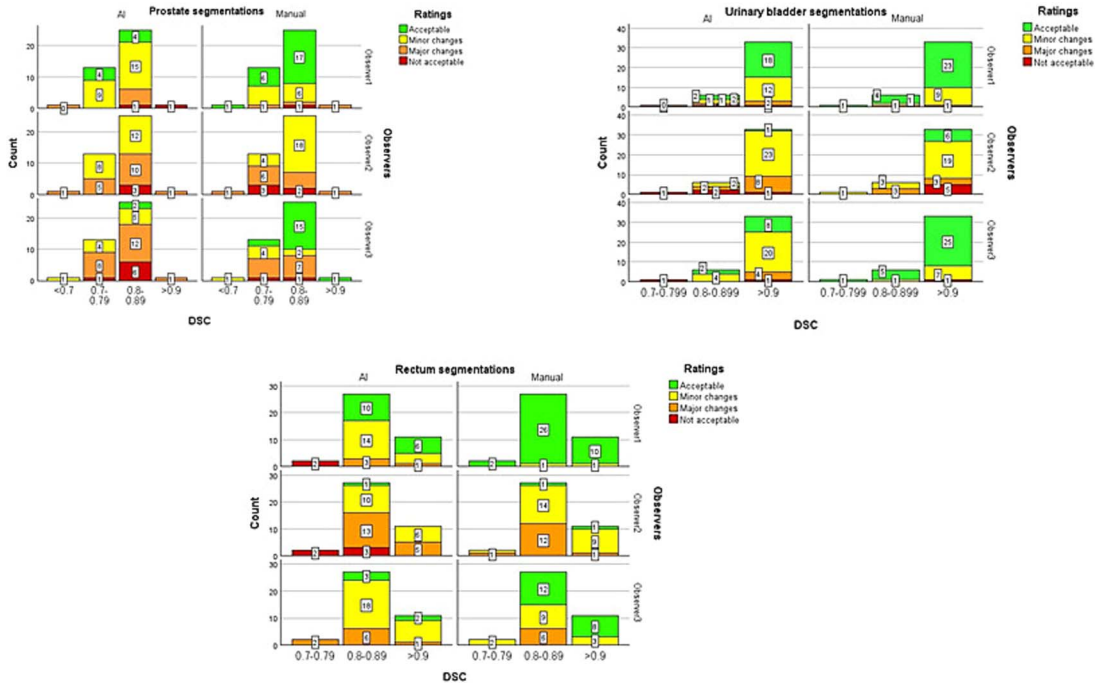

Figure 3. Distribution of the observer ratings in relation to DSC.

Table 2. The area under the curve of the VGC analysis for each organ.

Organ of interest	AUC_{VGC}[a]	95% CI[b]	*P*-value
Prostate	0.36	0.27–0.44	.002
Urinary bladder	0.35	0.28–0.41	< .001
Rectum	0.30	0.22–0.40	< .001

[a]Area under the VGC curve. [b]95% Confidence interval.

(Fig. 2). There was substantial variation in absolute ratings between the observers (Fig. 3). However, as VGC Analyzer [35] compare ratings individually for each observer, the variation between observers in absolute ratings will not influence the results of the statistical analysis.

For all AI-based organ segmentations, AUC_{VGC} values were significantly <0.5. The AUC_{VGC} for the prostate segmentations was 0.36 (95% CI 0.27–0.44), while the urinary bladder and rectum segmentations had AUC_{VGC} of 0.35 (95% CI 0.28–0.41) and 0.3 (95% CI 0.22–0.40), respectively (Table 2). The low AUC_{VGC} of the AI-based segmentations of all organs indicates that manual delineations were rated significantly better (Fig. 4).

Spearman's coefficient analysis showed no correlation between the observer ratings and DSC for any organ AI segmentation (Fig. 5).

Input from the observers

The observers provided feedback on the quality of segmentations and their ability to identify the AI-generated segmentations. Overall, observers noted certain cues that could potentially differentiate the segmentations, including adherence to anatomical organ boundaries, as AI-based segmentations occasionally contained segmentation errors that a human would not make. Other features of note were the asymmetry and angularity of the AI-generated segmentations, as well as the presence of extra contour lines into the neighbouring organs. Further, minor strikes were sometimes evident outside the organ of interest in the AI-based segmentations.

Discussion

Visual grading by the observers showed that a large proportion of the AI-based segmentations was clinically acceptable with no or minor modifications, although manual delineations generally received higher ratings. The variability in absolute ratings for both manual delineations and AI-based segmentations, despite the consensus reached among observers on how to interpret the different rating alternatives [6–8, 36], highlights the inherent subjectivity and difficulty of pelvic organ delineation, rather than poor quality of the training data or inadequate reader assessment, that may exist in clinical reality.

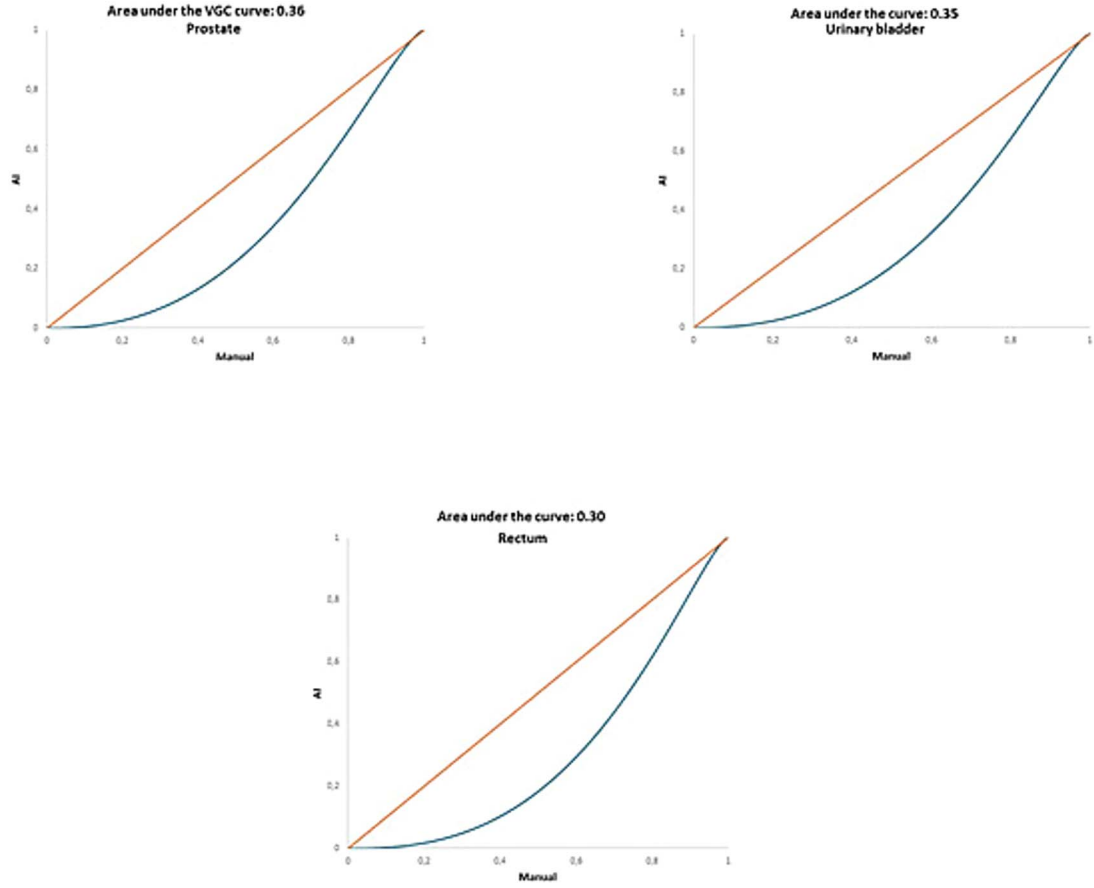

Figure 4. The area under the VGC curve (AUC_{VGC}) for the prostate and the OARs.

An important consideration is the use of manual delineations as ground truth for the training of the AI algorithm, despite the absence of an absolute reference standard in pelvic organ delineation. Although all manual delineations underwent quality control before inclusion, they remain expert-dependent interpretations. While manual delineations remain the clinical standard, AI-based segmentations are still being perceived as inferior when deviating from conventional manual styles, despite inherently subjective expert evaluations [6–8]. Consequently, the observer's ratings in this study reflected both the quality of the training delineations and the AI-generated segmentations.

Since the study material represented segmentations with different DSC, there was a large variation in the delineation qualities of the different organs. Yet, observer ratings for the AI segmentations showed no correlation with DSC, probably reflecting inter-observer variability. The visual grading by the observers suggested the need for some corrections of the AI-based segmentations, although the algorithm achieved accurate organ segmentations across various imaging datasets, despite the need for manual intervention. Low soft tissue contrast in CT contributes to poor differentiation between the bladder, the surrounding small intestine and the adjacent prostate [37]. Further, anatomical variations of the urinary bladder or the presence of nearby enlarged lymph nodes may cause difficulties in training the algorithm and lead to suboptimal AI-based segmentations, as reflected by the observers' ratings.

Over the years, various upper limits for the delineation of the rectum have been applied in manual delineations, although the rectosigmoid junction has recently been established as the upper limit in international guidelines [38]. The 12-year span of the included examinations, during which radiotherapy guidelines were revised, together with anatomic and motility variations of the bowel contributed to lower ratings for the AI-based segmentations, of which only 18% were considered acceptable without any corrections, compared with 50% of the corresponding manual delineations.

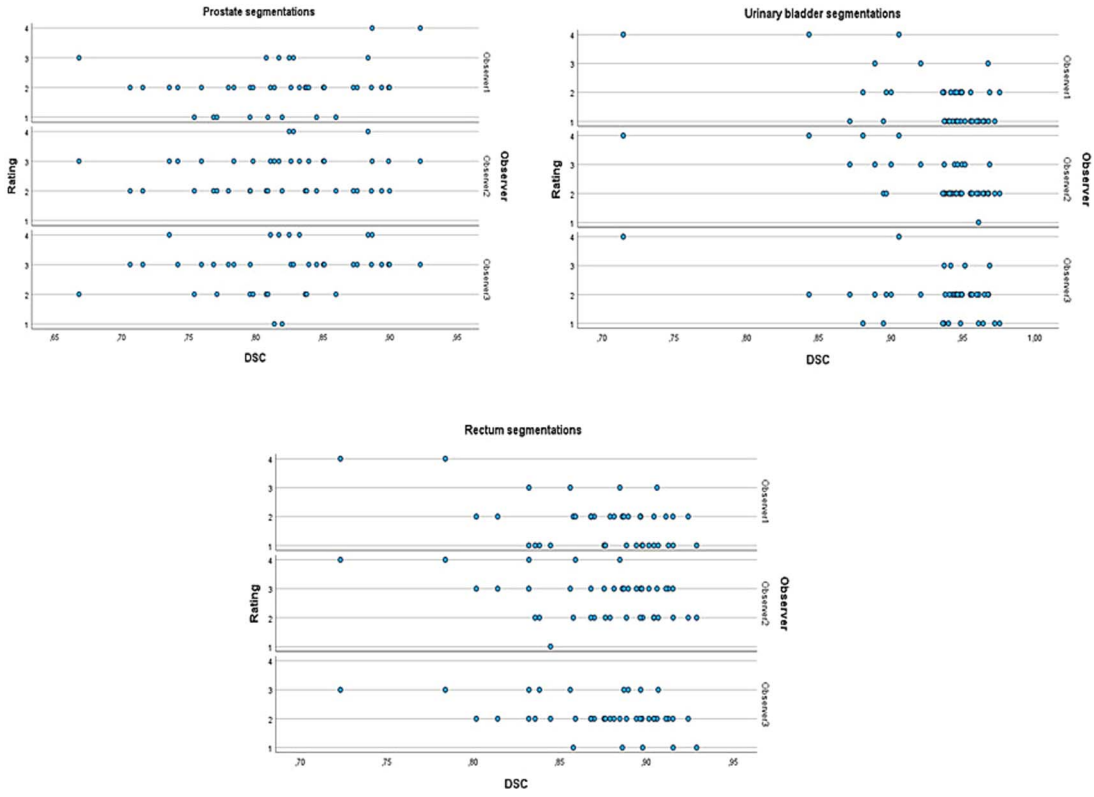

Figure 5. Scatter plots showing the correlation between the ratings by the observers and DSC.

Nevertheless, the study material was deliberately chosen to include representative images from each DSC category and to reflect the overall heterogeneity within the source cohort of over 1500 patients [24].

Prostate ratings also varied greatly, even in images with high DSC values, mainly because of varying manual delineations used as training material. Delineation is problematic at the apex of the prostate and near the seminal vesicles as it is difficult to differentiate between soft tissues in this area [7, 39, 40]. In a recent study, this issue affected the delineations of the prostate, resulting in differing AI- and manual segmentations [23]. Gardner *et al*. have also encountered this issue when comparing human consensus organ delineations and Deformable Image Registration-generated (DIR) delineations [41]. The results of this study showed variation between the delineations, suggesting that manual correction is necessary for their clinical use.

Several recent studies have focused on the automated segmentation of target volumes and PCa tumours, including the seminal vesicles and the prostatic bed [42–44]. Still, studies evaluating the clinical usefulness of AI-based organ segmentations are limited. A recent study demonstrated successful AI-based segmentation of both the target organ and OAR in images of PCa patients, the results of which showed good agreement with the corresponding manual delineations on pretreatment CT scans [24]. The present study further explored the visual grading of AI-based organ segmentations in PCa patients with respect to their future clinical utility and—despite successful AI-based segmentations—highlighted the high variability between visual assessments. Schreier *et al*. [45] have also evaluated the segmentation quality of a deep-learning tool that was applied to the prostate and OAR before radiation treatment and showed comparable DSC values to those of the aforementioned study [24]. Still, Schreier *et al*. included fewer patients and mostly cone-beam CT (CBCT) images, which provides lower tissue contrast compared with CT [46]. This may have contributed to poorer recognition of the algorithm's segmentation errors compared to CT. Radici *et al*. have also reached the same conclusion regarding CT and CBCT segmentations of the prostate, urinary bladder, and rectum (10 per organ) and also showed no significant differences in geometric measurements between AI-based segmentations and consensus manual organ delineations [23]. The present study included 40 AI-based CT segmentations of each organ and visual grading was performed.

In another study, Urago *et al*. [47] compared AI- and atlas-based organ segmentations of CT images and, after estimating them visually, showed that the AI-derived segmentations needed fewer manual corrections. In the present study, the ground truth consisted of manually delineated clinical CT scans, some of which were also combined with MRI, making the delineations more accurate.

A recent study by Duan *et al*. [22] also investigated the clinical acceptability of AI-segmentations based on visual assessments. However, the algorithm used in the present study was trained on a larger dataset, and DSC values were also taken into account for the analysis, which allowed for the comparison between geometric measurements and visual assessments. Moreover, observers from different clinics with varying experience levels participated in the present study, which reflects the existing variability in clinical practice.

A challenge regarding the use of AI is observer vigilance and the risk of potential identification of AI as the source of organ segmentation. In the present study, subtle segmentation errors made by the algorithm were noted by the observers, particularly at the rectosigmoid junction or near organ boundaries. This could potentially lead to the association of these cues with AI-based segmentations, leading to unwanted bias. Ivarsson and Lindwall [48] described the effect of human attitude towards AI, highlighting a tendency of individuals to align their personal beliefs about an AI product's origin and accuracy with their perceptions. However, this issue was outside the scope of the present study. A structured study methodology is needed to evaluate whether observers can identify AI-generated segmentations in this clinical context.

The current study had some limitations. The cohort consisted of a heterogeneous set of CT images obtained between 2006 and 2018, during which both image quality and the delineation guidelines varied. Moreover, MRI was not yet routinely used in all patients, although multimodal imaging improves pelvic organ visualization and segmentation accuracy as well as irradiation doses [38, 49–52]. However, variations in the manual delineation of prostate tumours have also been observed in MRI [36, 53, 54].

Although a limited number of images was included from the initial cohort, representative cases of all DSC values were included for visual grading. Only 40 AI-based segmentations from each organ were presented to the observers, yet, this number exceeded the size of the patient cohort (n = 10) in the study by Radici *et al*. [23]. However, future studies with more cases and observers would further evaluate the clinical usefulness of AI-based segmentations.

In the present study, the different organs were presented to the observers separately rather than simultaneously as in clinical practice. This allowed for an independent assessment of each organ without possible bias from the neighbouring structures.

Although the Likert scale used in this study primarily functioned as a binary measure of clinical acceptability, this approach was chosen to reflect a clinically relevant decision-making process. Additionally, the fact that the obtained ratings could be treated as ordinal data enabled the use of VGC analysis as a complementary evaluation method.

Conclusion

In conclusion, the majority of AI-generated organ segmentations was subjectively rated as clinically acceptable for treatment planning of PCa patients, where no or minor modifications were considered necessary. Only a small number of the AI-based segmentations were not considered clinically acceptable. The comparative visual analysis underscored the potential of combining AI technology and human expertise, yet, highlighted the issue of subjective variability between observers regardless of segmentation method used. For clinical implementation, future studies could define objective criteria for acceptable AI segmentations and specify the level of manual adjustment while still retaining clinical benefit of an AI-assisted workflow.

Abbreviations

CT, computed tomography; MRI, magnetic resonance imaging; OAR, organs at risk; AI, Artificial Intelligence; PCa, prostate cancer; HD, Hausdorff distance; MSD, Mean surface distance; DSC, Sørensen–Dice similarity coefficient; VGC, visual grading characteristics; AUC_{VGC}, Area under the VGC curve.

Author contributions

E.P. identified the research objective, developed the methodology, analysed the data, project administration, writing the manuscript—original draft, editing the manuscript. Å.A.J. developed the methodology, supervised the research study, reviewed the manuscript. O.E. and J.U. developed the software, reviewed the manuscript. J.K. supervision, reviewed the manuscript. K.B., H.-J.W., and M.T. evaluated and rated the segmentations. E.T. reviewed the manuscript. L.E. and H.K. developed the methodology, supervised the research study, reviewed the manuscript. A.S. analysed the data, reviewed the manuscript.

Conflict of interest: None declared.

Funding

This work was supported by Göteborgs Läkarsällskap and the Department of Radiology at the University Hospital of Sahlgrenska.

References

1. Alpert HR, Hillman BJ. Quality and variability in diagnostic radiology. *J Am Coll Radiol* 2004;**1**:127–32. https://doi.org/10.1016/j.jacr.2003.11.001
2. Benchoufi M, Matzner-Lober E, Molinari N. *et al.* Interobserver agreement issues in radiology. *Diagn Interv Imaging* 2020;**101**:639–41. https://doi.org/10.1016/j.diii.2020.09.001
3. Weber DC, Tomsej M, Melidis C. *et al.* QA makes a clinical trial stronger: evidence-based medicine in radiation therapy. *Radiother Oncol* 2012;**105**:4–8. https://doi.org/10.1016/j.radonc.2012.08.008
4. Peters LJ, O'Sullivan B, Giralt J. *et al.* Critical impact of radiotherapy protocol compliance and quality in the treatment of advanced head and neck cancer: results from TROG 02.02. *J Clin Oncol* 2010;**28**:2996–3001. https://doi.org/10.1200/JCO.2009.27.4498
5. Ohri N, Shen X, Dicker AP. *et al.* Radiotherapy protocol deviations and clinical outcomes: a meta-analysis of cooperative group clinical trials. *J Natl Cancer Inst* 2013;**105**:387–93. https://doi.org/10.1093/jnci/djt001
6. Guzene L, Beddok A, Nioche C. *et al.* Assessing Interobserver variability in the delineation of structures in radiation oncology: a systematic review. *Int J Radiat Oncol Biol Phys* 2023;**115**:1047–60. https://doi.org/10.1016/j.ijrobp.2022.11.021
7. Fiorino C, Reni M, Bolognesi A. *et al.* Intra- and inter-observer variability in contouring prostate and seminal vesicles: implications for conformal treatment planning. *Radiother Oncol* 1998;**47**:285–92. https://doi.org/10.1016/S0167-8140(98)00021-8
8. Njeh CF. Tumor delineation: the weakest link in the search for accuracy in radiotherapy. *J Med Phys* 2008;**33**:136–40. https://doi.org/10.4103/0971-6203.44472
9. Vinod SK, Min M, Jameson MG. *et al.* A review of interventions to reduce inter-observer variability in volume delineation in radiation oncology. *J Med Imaging Radiat Oncol* 2016;**60**:393–406. https://doi.org/10.1111/1754-9485.12462
10. Vandewinckele L, Claessens M, Dinkla A. *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
11. Chung SY, Chang JS, Choi MS. *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**:44. https://doi.org/10.1186/s13014-021-01771-z
12. Kiljunen T, Akram S, Niemelä J. *et al.* A deep learning-based automated CT segmentation of prostate cancer anatomy for radiation therapy planning-a retrospective Multicenter study. *Diagnostics (Basel)* 2020;**10**:959. https://doi.org/10.3390/diagnostics10110959
13. Jamtheim Gustafsson C, Lempart M, Swärd J. *et al.* Deep learning-based classification and structure name standardization for organ at risk and target delineations in prostate cancer radiotherapy. *J Appl Clin Med Phys* 2021;**22**:51–63. https://doi.org/10.1002/acm2.13446
14. Samarasinghe G, Jameson M, Vinod S. *et al.* Deep learning for segmentation in radiation therapy planning: a review. *J Med Imaging Radiat Oncol* 2021;**65**:578–95. https://doi.org/10.1111/1754-9485.13286
15. Isaksson LJ, Summers P, Mastroleo F. *et al.* Automatic segmentation with deep learning in radiotherapy. *Cancers (Basel)* 2023;**15**. https://doi.org/10.3390/cancers15174389
16. Chen X, Liu X, Wu Y. *et al.* Research related to the diagnosis of prostate cancer based on machine learning medical images: a review. *Int J Med Inform* 2024;**181**:105279. https://doi.org/10.1016/j.ijmedinf.2023.105279
17. Palazzo G, Mangili P, Deantoni C. *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
18. Kazemifar S, Balagopal A, Nguyen D. *et al.* Segmentation of the prostate and organs at risk in male pelvic CT images using deep learning. *Biomed Phys Eng Express* 2018;**4**:055003. https://doi.org/10.1088/2057-1976/aad100
19. Astaraki M, Severgnini M, Milan V. *et al.* Evaluation of localized region-based segmentation algorithms for CT-based delineation of organs at risk in radiotherapy. *Phys Imaging Radiat Oncol* 2018;**5**:52–7. https://doi.org/10.1016/j.phro.2018.02.003
20. Chen X, Sun S, Bai N. *et al.* A deep learning-based auto-segmentation system for organs-at-risk on whole-body computed tomography images for radiation therapy. *Radiother Oncol* 2021;**160**:175–84. https://doi.org/10.1016/j.radonc.2021.04.019
21. Sartor H, Minarik D, Enqvist O. *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
22. Duan J, Bernard M, Downes L. *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
23. Radici L, Piva C, Casanova Borca V. *et al.* Clinical evaluation of a deep learning CBCT auto-segmentation software for prostate adaptive radiation therapy. *Clin Transl Radiat Oncol* 2024;**47**:100796
24. Polymeri E, Johnsson ÅA, Enqvist O. *et al.* Artificial intelligence-based organ delineation for radiation treatment planning of prostate cancer on computed tomography. Advances. *Radiat Oncol* 2024;**9**:101383. https://doi.org/10.1016/j.adro.2023.101383
25. Båth M, Månsson LG. Visual grading characteristics (VGC) analysis: a non-parametric rank-invariant statistical method for image quality evaluation. *Br J Radiol* 2007;**80**:169–76. https://doi.org/10.1259/bjr/35012658
26. Svalkvist A, Fagman E, Vikgren J. *et al.* Evaluation of deep-learning image reconstruction for chest CT examinations at two different dose levels. *J Appl Clin Med Phys* 2023;**24**:e13871. https://doi.org/10.1002/acm2.13871
27. Beer L, Polanec SH, Baltzer PAT. *et al.* 4D perfusion CT of prostate cancer for image-guided radiotherapy planning:

a proof of concept study. *PLoS One* 2019;**14**:e0225673. https://doi.org/10.1371/journal.pone.0225673
28. Polanec SH, Lazar M, Wengert GJ. *et al.* 3D T2-weighted imaging to shorten multiparametric prostate MRI protocols. *Eur Radiol* 2018;**28**:1634–41. https://doi.org/10.1007/s00330-017-5120-5
29. Arnoldner MA, Polanec SH, Lazar M. *et al.* Rectal preparation significantly improves prostate imaging quality: assessment of the PI-QUAL score with visual grading characteristics. *Eur J Radiol* 2022;**147**:110145. https://doi.org/10.1016/j.ejrad.2021.110145
30. Wong J, Fong A, McVicar N. *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. https://doi.org/10.1016/j.radonc.2019.10.019
31. Cha E, Elguindi S, Onochie I. *et al.* Clinical implementation of deep learning contour autosegmentation for prostate radiotherapy. *Radiother Oncol* 2021;**159**:1–7. https://doi.org/10.1016/j.radonc.2021.02.040
32. Mackay K, Bernstein D, Glocker B. *et al.* A review of the metrics used to assess auto-contouring systems in radiotherapy. *Clin Oncol (R Coll Radiol)* 2023;**35**:354–69. https://doi.org/10.1016/j.clon.2023.01.016
33. Trägårdh E, Borrelli P, Kaboteh R. *et al.* RECOMIA—a cloud-based platform for artificial intelligence research in nuclear medicine and radiology. *EJNMMI Phys* 2020;**7**:51. https://doi.org/10.1186/s40658-020-00316-9
34. Nasa P, Jain R, Juneja D. Delphi methodology in healthcare research: how to decide its appropriateness. *World J Methodol* 2021;**11**:116–29. https://doi.org/10.5662/wjm.v11.i4.116
35. Båth M, Hansson J. VGC ANALYZER: a software for statistical analysis of fully crossed multiple-reader multiple-case visual grading characteristics studies. *Radiat Prot Dosim* 2016;**169**:46–53. https://doi.org/10.1093/rpd/ncv542
36. Chen MY, Woodruff MA, Dasgupta P. *et al.* Variability in accuracy of prostate cancer segmentation among radiologists, urologists, and scientists. *Cancer Med* 2020;**9**:7172–82. https://doi.org/10.1002/cam4.3386
37. Dirix P, Haustermans K, Vandecaveye V. The value of magnetic resonance imaging for radiotherapy planning. *Semin Radiat Oncol* 2014;**24**:151–9. https://doi.org/10.1016/j.semradonc.2014.02.003
38. Salembier C, Villeirs G, De Bari B. *et al.* ESTRO ACROP consensus guideline on CT- and MRI-based target volume delineation for primary radiation therapy of localized prostate cancer. *Radiother Oncol* 2018;**127**:49–61. https://doi.org/10.1016/j.radonc.2018.01.014
39. Wachter S, Wachter-Gerstner N, Bock T. *et al.* Interobserver comparison of CT and MRI-based prostate apex definition. Clinical relevance for conformal radiotherapy treatment planning. *Strahlenther Onkol* 2002;**178**:263–8. https://doi.org/10.1007/s00066-002-0907-x
40. Rasch C, Steenbakkers R, van Herk M. Target definition in prostate, head, and neck. *Semin Radiat Oncol* 2005;**15**:136–45. https://doi.org/10.1016/j.semradonc.2005.01.005
41. Gardner SJ, Wen N, Kim J. *et al.* Contouring variability of human- and deformable-generated contours in radiotherapy for prostate cancer. *Phys Med Biol* 2015;**60**:4429–47. https://doi.org/10.1088/0031-9155/60/11/4429
42. Matoska T, Patel M, Liu H. *et al.* Review of deep learning based autosegmentation for clinical target volume: current status and future directions. *Adv Radiat Oncol* 2024;**9**:101470. https://doi.org/10.1016/j.adro.2024.101470
43. Wen F, Chen Z, Wang X. *et al.* Deep learning based clinical target volumes contouring for prostate cancer: easy and efficient application. *J Appl Clin Med Phys* 2024;**25**:e14482. https://doi.org/10.1002/acm2.14482
44. Hou Z, Gao S, Liu J. *et al.* Clinical evaluation of deep learning-based automatic clinical target volume segmentation: a single-institution multi-site tumor experience. *Radiol Med* 2023;**128**:1250–61. https://doi.org/10.1007/s11547-023-01690-x
45. Schreier J, Genghi A, Laaksonen H. *et al.* Clinical evaluation of a full-image deep segmentation algorithm for the male pelvis on cone-beam CT and CT. *Radiother Oncol* 2020;**145**:1–6. https://doi.org/10.1016/j.radonc.2019.11.021
46. Lechuga L, Weidlich GA, Cone Beam CT. *et al.* Fan Beam CT: a comparison of image quality and dose delivered between two differing CT imaging modalities. *Cureus.* 2016;**8**:e778. https://doi.org/10.7759/cureus.778
47. Urago Y, Okamoto H, Kaneda T. *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
48. Ivarsson J, Lindwall O. Suspicious minds: the problem of trust and conversational agents. *Comput Support Coop Work* 2023;**32**:545–71. https://doi.org/10.1007/s10606-023-09465-8
49. Evans PM. Anatomical imaging for radiotherapy. *Phys Med Biol* 2008;**53**:R151–91. https://doi.org/10.1088/0031-9155/53/12/R01
50. Daryanani A, Turkbey B. Recent advancements in CT and MR imaging of prostate cancer. *Semin Nucl Med* 2022;**52**:365–73. https://doi.org/10.1053/j.semnuclmed.2021.11.013
51. Ali AN, Rossi PJ, Godette KD. *et al.* Impact of magnetic resonance imaging on computed tomography-based treatment planning and acute toxicity for prostate cancer patients treated with intensity modulated radiation therapy. *Pract Radiat Oncol* 2013;**3**:e1–9. https://doi.org/10.1016/j.prro.2012.04.005
52. Steenbakkers RJHM, Deurloo KEI, Nowak PJCM. *et al.* Reduction of dose delivered to the rectum and bulb of the penis using MRI delineation for radiotherapy of the prostate. *Int J Radiat Oncol Biol Phys* 2003;**57**:1269–79. https://doi.org/10.1016/S0360-3016(03)01446-9
53. Borofsky S, George AK, Gaur S. *et al.* What are we missing? False-negative cancers at multiparametric MR imaging of the prostate. *Radiology* 2018;**286**:186–95. https://doi.org/10.1148/radiol.2017152877
54. Montagne S, Hamzaoui D, Allera A. *et al.* Challenge of prostate MRI segmentation on T2-weighted images: inter-observer variability and impact of prostate morphology. *Insights Imaging* 2021;**12**:71. https://doi.org/10.1186/s13244-021-01010-9

Radiation Protection Dosimetry, 2026, 202(3–4), 204–213
https://doi.org/10.1093/rpd/ncaf184
Paper