

Original Article

Cite this article: Dahlstedt-Hassler L, Siegbahn A, Kojoj R, Schedin A, Lagerbäck C, and Lind P. (2026) Qualitative and quantitative evaluation of auto-segmented clinical target volumes and organs at risk in radiotherapy of rectal cancer. *Journal of Radiotherapy in Practice*. 25(e9), 1–7. doi: [10.1017/S1460396926100405](https://doi.org/10.1017/S1460396926100405)

Received: 6 October 2025

Revised: 29 January 2026

Accepted: 30 January 2026

Keywords:

AI; auto-segmentation; clinical target volume; radiotherapy planning; rectal cancer

Corresponding author:

Love Dahlstedt-Hassler;

Email: love.dahlstedt-hassler@regionstockholm.se

Qualitative and quantitative evaluation of auto-segmented clinical target volumes and organs at risk in radiotherapy of rectal cancer

Love Dahlstedt-Hassler^{1,2} , Albert Siegbahn^{1,2} , Rafat Kojoj¹ , Anna Schedin¹, Cecilia Lagerbäck¹ and Pehr Lind^{1,2} 

¹Department of Oncology, Södersjukhuset: Sodersjukhuset, Stockholm, Sweden and ²Department of Clinical Science and Education, Södersjukhuset, Karolinska Institutet, Stockholm, Sweden

Abstract

Introduction: Manual contouring (MC) is time-consuming work in radiotherapy planning for rectal cancer. Artificial intelligence (AI) can reduce the time required for clinical target volume (CTV) and organs-at-risk (OARs) delineation. In this study, we evaluated the quality of auto-segmented CTVs and OARs.

Methods: Dose-planning data were collected from ten patients who underwent preoperative radiotherapy for locally advanced rectal cancer in 2024. Auto-segmented structures from the AI-Rad and Contour+ software tools were added. Constructed AI-CTVs, based on Contour+ segmentations and AI-OARs, i.e., bladder, femoral heads and bowel bag, by both AI tools, were compared to their MC counterparts by use of quantitative metrics, volumetric/surface Dice similarity coefficients (vDSC/sDSC) and maximum/average Hausdorff distance (HD/aHD). The constructed AI-CTVs and MC counterparts were graded by two radiotherapists with two qualitative methods.

Results: The median vDSC, sDSC, HD and aHD values of our constructed AI-CTVs compared with the MC-CTVs were 0.86, 0.61, 23.19 and 0.62 mm, respectively. For both AI tools, the agreement in the OAR metrics was overall good but less similar for the bowel bag. The qualitative evaluations of the AI-CTVs, compared to the MC-CTVs, were in clear favour of the MC-CTVs. The cranial-anterior nodal levels were anatomical areas with poorer coverage, where the contouring guidelines differed.

Conclusion: The quality of our constructed AI-CTVs was inferior to the MC-CTVs. Thus, the auto-segmentation methods need further development on this aspect for use in the clinical setting. In contrast, the agreement of the quantitative metrics for the OARs was overall good, except for the bowel bag.

Introduction

The incidence of rectal cancer is increasing.¹ Treatment of locally advanced rectal cancer often involves neoadjuvant radiotherapy (RT). In cases of complete clinical response, some patients may not require surgery and can instead be closely monitored in a watch-and-wait program.²

The planning of rectal cancer irradiation involves the manual contouring (MC) of organs at risk (OARs), the gross tumour volume (GTV) and the clinical target volume (CTV). A time-consuming aspect of the CTV delineation is the contouring of the lymphatic nodal levels (LN). Contouring is often performed according to annotation guidelines, such as RTOG³ or the UK Royal College of Radiologists' (RCR) guidance.⁴ Delineation is a labour-intensive process that requires quality control to prevent excessive incidental irradiation and underdosed volumes. MC is also associated with inter-observer variation,⁵ which is sometimes compensated for by expanding the final treatment volume.⁶ To improve the efficiency of this process, the use of image-processing artificial intelligence (AI), auto-segmentation, has become a rapidly emerging tool in RT planning.

Currently, there are numerous available auto-segmentation tools on the market for clinical use, especially for OARs segmentation.⁷ Benefits include shorter lead times to treatment, more efficient use of limited professional resources and reduced inter-observer variation.⁸ These clinical advantages are all dependent on the quality of the delineations.⁷ Poorly auto-segmented structures could require substantial manual correction⁷ or may result in the inadvertent acceptance of suboptimal segmentations.⁸ As the field of auto-contouring continues to evolve, ongoing in-house quality control and assurance are essential to achieve superior therapeutic outcomes.

The methods used to evaluate the clinical benefit of auto-segmentation differ.⁸ While time measurements could be considered representative for clinical benefit, they are themselves time-consuming. Consequently, evaluations are often conducted using various surrogate measures,

© The Author(s), 2026. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.



frequently including quantitative geometric metrics or qualitative assessments of the segmentations.⁸ At present, there is no consensus on how these surrogate measures should be evaluated.⁸

While the clinical benefits of auto-segmentation of OARs have been studied and reported in earlier publications, the use of auto-segmented CTVs for specific tumours has not been explored to the same extent. A potential advantage of auto-segmenting the CTVs in rectal cancer RT planning has been suggested in only a few published works,^{9–14} although these employed varying evaluation methods and yielded different results. In 2024, our department had the opportunity to evaluate two deep learning-based software tools simultaneously in RT planning. The clinical application of these tools in our practice was limited to a few OARs.

In our study, in the setting of RT planning for the treatment of locally advanced rectal cancer, we aimed to investigate the potential use of a constructed CTV structure for rectal tumours by merging some of the auto-generated structures (AI-CTV) and estimate the clinical usefulness of these structures by comparing them to the manually contoured CTVs (MC-CTV) using a combination of quantitative and qualitative methods. We also quantitatively evaluated the auto-segmented OARs in comparison with their manually contoured counterparts.

Methods

Patients: Planning computed tomography images and contour data from ten consecutive rectal cancer patients who underwent preoperative radiotherapy at our department in 2024 were retrospectively collected and anonymised for the performed analyses. The indication for treatment was stage II–III, with no differing treatment protocol between the groups.

CT scanner and software: All patients underwent contrast-enhanced treatment-planning CT using a Siemens Somatom Definition Edge scanner (Siemens Healthineers AG, Forchheim, Germany), following our standard protocol with patient-size-adapted dose parameters and a typical slice thickness of 2 mm.

We used the treatment planning software Monaco 6.1.2.0 (Elekta, Stockholm, Sweden) for the MC; Hero 2024.2.0 (Hero Imaging, Umeå, Sweden) for quantitative data acquirement and SPSS v 29.0.2.0 (IBM, Armonk, USA) for statistical analysis. The auto-contouring software tools used in the study were AI-Rad Companion Organs RT VA50 (Siemens Healthineers AG, Forchheim, Germany) and Contour+ v1.2.7 (MVision AI Oy, Helsinki, Finland), henceforth referred to as AI-Rad and Contour+, respectively. Both these AI tools are trained on RTOG-annotated datasets for their rectal cancer structures. During the preparation of the article, we used an AI proofreading tool from Wordwise AI (Wordwise, Des Moines, USA) for language improvement. After using this tool, we reviewed and edited the content as needed.

Ethical approval was granted for the conducted retrospective analyses by the Swedish Ethical Review Authority (Etikprövningsmyndigheten), diary number 2025-04703-01.

Statistical analysis: Due to the low sample size, with presumed non-distribution, we used non-parametric tests, the Wilcoxon signed-rank test, for the difference between quantitative results of the different AI tools as well as the difference in qualitative scores between AI and MC. We used Kendall's rank correlation coefficient for correlation comparisons.

CTV segmentation: The manual contours, used in the clinical setting, served as ground truth in the comparisons. These contours were delineated and peer reviewed according to local protocol,

which followed the RCR guidance,⁴ in which the CTV is defined as a combination of relevant nodal volumes, such as the anterior obturator nodes and the clinical target volume of the tumour (CTVT), in turn, defined as the GTV structure with added margins of 2 cm cranially and caudally and 1.5 cm in all other directions.

Comparable AI-CTV and AI-GTV segments were not directly provided by the auto-segmentation tools. Therefore, a CTV was constructed using Contour+ structures in a stepwise manner (Figure 1): the structures labelled 'rectum' and 'recto-sigmoideum' were fused into a single structure, and slices cranial and caudal to the tumour borders were removed, creating the AI-GTV (Figure 1a–b). This AI-GTV was then expanded by adding margins (2 cm cranially and caudally and 1.5 cm in other directions) and fused with an auto-segmented rectal nodal volume (Figure 1c–d), resulting in the construction of the AI-CTV structure (Figure 1e). This volume could then be compared with the 'MC-CTV' (Figure 1f).

The intersection between our constructed AI-GTV and MC-CTVT was controlled to exclude any target misses (Figure 2a).

For AI-Rad, a similar approach for constructing an AI-CTV was not possible with the segments provided. However, the intersection between the MC-CTV and the internal iliac nodes suggested by AI-Rad was checked to measure the coverage of these lymph volumes (Figure 2b) and to explore a potential benefit of basing the contouring on this structure.

OAR segmentation: The bladder, bowel bag and femoral heads were manually contoured in the clinical setting and were available as ground truth in comparisons with auto-segmented structures. For the auto-segmented bowel bag, some revisions were made to enable a proper comparison with the MC bowel bag. AI-Rad provided the structure labelled 'Abdomino-peritoneal space', which included urogenital organs, which were cropped. For both AI software, all slices cranial and caudal to the MC-Bowel bag, as well as the CTV, were cropped. These revised structures were deemed acceptable for comparison with the MC-Bowel bag.

Metrics used to evaluate auto-segmentation: The quantitative metrics used in this study were the volumetric/surface Dice similarity coefficient (vDSC/sDSC), as well as the maximum/average Hausdorff distance (HD/aHD). The image structures were converted to binary masks prior to the calculation of these metrics. The tolerance value for the sDSC was, for all evaluated structures, set to 2 mm, a value considered acceptable for pelvic organs.¹⁵

Qualitative evaluation: Contour quality was reviewed by two experienced oncologists (co-authors A.S. and C.L). We employed two different methods to assess the cases:

Grading test: The structures MC-CTV and AI-CTV were first blinded and randomised, then reviewed separately within the treatment planning module, with one week's interval between the reviews of structures from the same case. A five-point Likert scale (Table 1) was developed and adapted based on similar scales used by Yu *et al.*¹⁶ and Almberg *et al.*¹⁷ In the evaluations, the MC-GTV was highlighted to indicate the tumour's location. No other information was given during the review. In cases graded 3 or lower, reviewers were also asked to specify the anatomical area requiring improvement and the nature of the needed adjustment, viz., expansion or reduction of contours. To facilitate the correlation analysis, the results from the five-point scale were converted symmetrically into a three-point scale, where scores of 1–2 corresponded to 1 (unacceptable), 3 to 2 (intermediate) and 4–5 to 3 (acceptable).

Preference test: As the grading test provided information about the entire segmentation, the reviewer could infer which structure

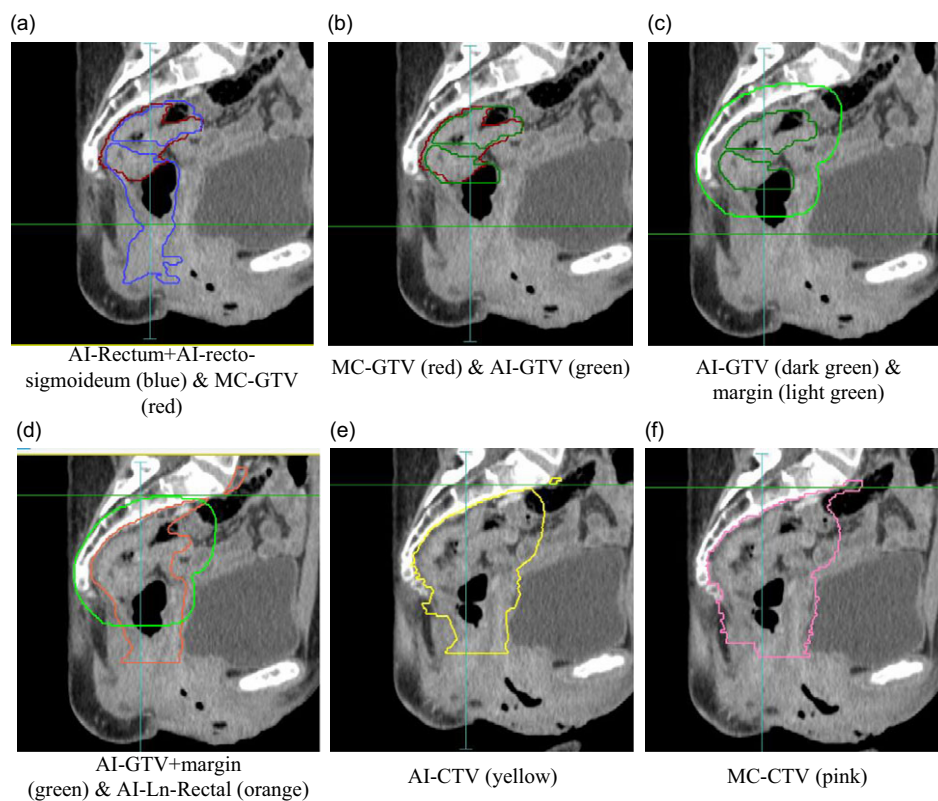


Figure 1. Images from a single case illustrating the steps involved in constructing the AI-CTV. (a) AI-rectum + AI-recto-sigmoidum (blue) & MC-GTV (red). (b) MC-GTV (red) & AI-GTV (green). (c) AI-GTV (dark green) & margin (light green). (d) AI-GTV + margin (green) & AI-Ln-Rectal (orange). (e) AI-CTV (yellow). (f) MC-CTV (pink). Abbreviations: AI, artificial intelligence; MC, manual contours; GTV, gross tumour volume; Ln, lymph nodal volume; CTV, clinical target volume.

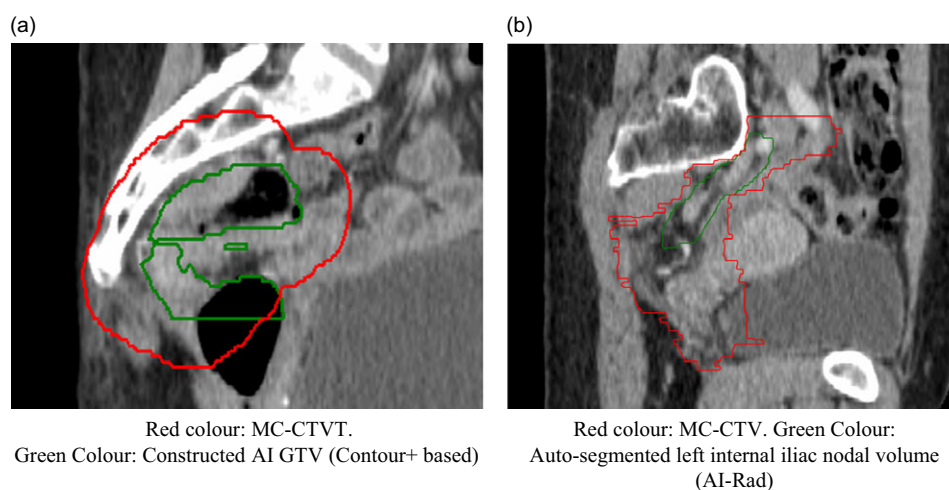


Figure 2. Sagittal view of the intersection between auto-segmentation-based structures and manual contoured target structures. (a) Red colour: MC-CTVT. Green Colour: Constructed AI GTV (Contour+ based). (b) Red colour: MC-CTV. Green Colour: Auto-segmented left internal iliac nodal volume (AI-Rad). Abbreviations: MC, manual contours; CTVT, clinical target volume of tumour; GTV, gross tumour volume; CTV, clinical target volume (including nodal fields); AI, artificial intelligence.

had been delineated by AI, potentially introducing bias. Therefore, a second test was devised based on the 'Turing test' also designed by Wu et al.¹⁴ In each case, 10 slices of the CTV were randomly selected, with contours of MC-CTV and AI-CTV randomly assigned the colours red or green in each individual slice (Figure 3). Each slice was then scored for preference of either colour using a bipolar 5-point Likert scale: 'strongly for red/green', 'for red/green' or 'neither for nor against red/green'. With ten cases, 100 images were thus reviewed.

Results

Quantitative metrics: In Table 2, summary statistics of the quantitative metrics of the constructed CTVs and CTVTs

(Contour+-based) as well as for OARs from both evaluated AI tools are presented. For the constructed CTVs, the median values of vDSC, sDSC, HD and aHD were 0.86, 0.61, 23.2 mm and 0.62 mm, respectively. For the OARs, the differences between Contour+ and AI-Rad contours were tested for zero median, with *p*-values listed in the footnotes. There was a significant difference between the AI tools analysed for most OARs and metrics, except for the bowel bag, where the Contour+ OARs had a greater similarity to their MC counterpart, with higher vDSC and sDSC values and lower HD and aHD values compared to the AI-Rad OARs.

Qualitative tests: Table 3 presents both qualitative and quantitative metrics of our constructed AI-CTV, based on Contour+ auto-segmentations. Median grading test scores are

Table 1. Grading test: questions assessing clinical benefit of CTV structures

Score	How would you rate the amount of corrections needed to approve the structure? How much time is estimated to be saved by correcting the structure versus manually contouring from scratch?
1	Contour is unusable and needs to be redrawn from scratch. No time saved.
2	Contour requires major corrections. Little to no time saved.
3	Contour requires minor corrections. Time saved.
4	Contour may need minor edits that are not essential (stylistic differences, but not clinically important). Significant time saved.
5	Contour requires no correction and can be used as-is. Significant time saved.

CTV, clinical target volume.

displayed on a case-by-case basis. The median score on the five-point grading scale was 2.0 (range 1–3) for the constructed AI-CTV, indicating overall poor quality and a need for ‘major corrections with little to no time saved’. All constructed AI-CTVs were deemed to require revision by both reviewers. The median score for the MC-CTV was 4.3 (range 3–5), indicating overall acceptance of the MC.

For the preference test, the median of the reviewers’ mean scores for each case’s ten reviewed slices is presented. A score of 1 indicates a strong preference for MC, 3 indicates neutrality (neither for nor against MC or AI), and 5 indicates a strong preference for AI. A total median score of 2.0 represents an average, but not strong, preference for MC over AI.

There was a consistent difference between AI and MC segmentations, favouring MC in both the grading and preference tests, *p*-value <0.01. Some differences were observed between the two reviewers, although the variation was mostly limited to single steps on the scale.

In the poorly graded CTV structures, a frequently noted omission was the cranial-anterior nodal levels, that is, the nodes around the superior rectal artery and the anterior internal iliac artery, either due to the entire intended level being excluded or an insufficient margin around these vessels.

The correlation values between the metrics and the grading test score for the constructed AI-CTV are presented in Table 4. We found no statistical significance in this analysis. The correlation coefficients trended as negative for vDSC/sDSC and positive for HD/aHD, which is contrary to the expected relationship between a good metric score and a high grading score. This pattern was more pronounced when using the three-point scale compared to the five-point scale, although it was still not statistically significant.

Intersection analysis: The volumes of the constructed AI-GTV were covered by the structure MC-GTV+margin to an average extent of 99.6% (Figure 2a). The AI-Rad-segmented external iliac nodal level was, on average, 97.2% (dx) and 96.2% (sin) covered by the MC-CTV (Figure 2b).

Discussion

Interpretation of CTV results: The median vDSC of 0.86 for our constructed AI-CTV was comparable to mean values reported in

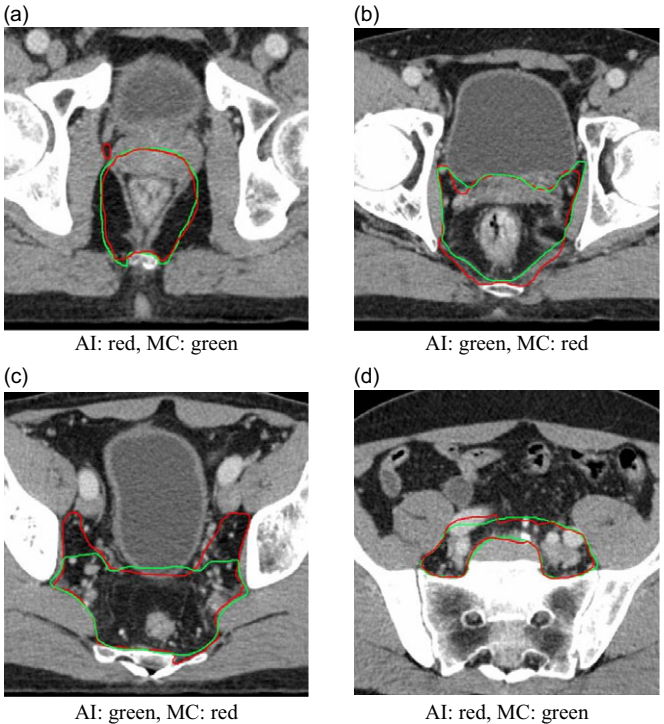


Figure 3. Examples of slices reviewed in the preference test. The slices were arranged from caudal to cranial direction. The colour given to structures was individually randomised in each slice. (a) AI: red, MC: green. (b) AI: green, MC: red. (c) AI: green, MC: red. (d) AI: red, MC: green. Abbreviations: AI, constructed clinical target volume (CTV); MC, manually contoured CTV.

other rectal cancer studies, e.g., Geng: 0.85⁹, Wu: 0.9¹⁴, Men: 0.87¹² and Song: 0.87¹¹. In contrast, the sDSC result, median 0.61, was lower than that reported by Song: 0.79¹¹; however, they used a higher tolerance value of 4 mm. Our result for the HD, with a median value of 23.19 mm, was considerably higher than the mean values in the aforementioned studies, viz., between 7 and 9 mm. The maximum HD is sensitive to outlier values,¹⁸ and the shape of the CTV and the process of its construction may have resulted in more outliers, suggesting that maximum HD might not be a suitable metric for this structure. We have not found comparable results of aHD. Based on data from other organ structures, our result, a median of 0.62, was interpreted as intermediate (0.4–0.6).¹⁹

While our vDSC results for the constructed AI-CTV were acceptable, the corresponding qualitative grading score was low. In contrast, Geng *et al.* report considerably higher grading scores for their AI-CTV. This discrepancy between the qualitative results and vDSC values provides further evidence that vDSC results should be interpreted with caution.

The correlation analysis did not align as expected, indicating a relationship between bad metric scores and higher qualitative grades. However, since none of the analysed AI-CTVs were given a good grade, the variation reflects only different degrees of bad gradings. Moreover, this result was not statistically significant. Therefore, no conclusion can be drawn regarding the correlation between quantitative metrics and qualitative assessments.

Interpretation of OARs results: Contour+ performed slightly better than AI-Rad across all OARs. However, both tools produced acceptable results, and our findings were comparable to those of other studies assessing OARs in the pelvic region.^{11,12} For the bowel

Table 2. Summary of statistics for quantitative metrics

Volume	AI-type	vDSC			sDSC [2 mm]			HD			aHD		
		Med.	(IQR)	Diff.	Med.	(IQR)	Diff.	Med.	(IQR)	Diff.	Med.	(IQR)	Diff.
CTV	C+	0.86	(0.04)	.	0.61	(0.06)	.	23.19	(6.75)	.	0.62	(0.19)	.
	AIR	.	.	.	.	.	.	.	.	.	.	.	.
CTVT	C+	0.94	(0.03)	.	0.85	(0.18)	.	9.49	(12.80)	.	0.13	(0.17)	.
	AIR	.	.	.	.	.	.	.	.	.	.	.	.
Bladder	C+	0.96	(0.02)	0.020*	0.97	(0.02)	0.094*	4.28	(3.16)	−1.34*	0.06	(0.04)	−0.051*
	AIR	0.94	(0.02)	.	0.87	(0.13)	.	6.50	(4.43)	.	0.10	(0.02)	.
Bowel bag	C+	0.86	(0.07)	0.041*	0.57	(0.22)	0.053*	32.11	(13.93)	−8.45*	0.59	(0.36)	−0.894*
	AIR	0.83	(0.09)	.	0.48	(0.25)	.	34.72	(21.46)	.	0.78	(0.47)	.
FH left	C+	0.96	(0.01)	0.021*	0.99	(0.01)	0.043*	2.96	(0.48)	−2.33*	0.05	(0.01)	−0.077*
	AIR	0.94	(0.01)	.	0.95	(0.04)	.	3.79	(1.19)	.	0.08	(0.02)	.
FH right	C+	0.96	(0.01)	0.021*	0.99	(0.02)	0.040*	3.25	(1.20)	−1.66*	0.05	(0.02)	−0.067*
	AIR	0.94	(0.02)	.	0.95	(0.06)	.	3.91	(0.35)	.	0.08	(0.04)	.

*p < 0.01.

vDSC, volumetric Dice similarity coefficient; sDSC, surface Dice similarity coefficient; HD, Hausdorff distance; aHD, average Hausdorff distance; Med, median; IQR, inter-quartile range; CTV, clinical target volume; CTVT, clinical target volume of the tumour; FH, femoral head.

Table 3. Case-by-case display of qualitative metrics and qualitative test scores for the constructed AI and MC CTV, including nodal fields

Case No (<i>n</i> = 10)	Quantitative metrics				Grading test, 5-graded ^a			Grading test, 3-graded			Preference test ^b
	vDSC	sDSC [2mm]	HD	aHD	AI	MC	Diff. MC-AI	AI	MC	Diff. MC-AI	MC vs AI
1	0.86	0.71	19.9	0.47	1	3	2	1	2	1	1.7
2	0.90	0.84	23.6	0.34	2	4	2	1	2.5	1.5	2.0
3	0.88	0.65	22.0	0.49	2.5	5	2.5	1.5	3	1.5	1.9
4	0.80	0.72	20.2	0.74	1.5	4.5	3	1	3	2	2.1
5	0.83	0.59	22.8	0.60	2.5	3.5	1	1.5	2.5	1	2.6
6	0.86	0.62	26.6	0.64	1.5	4.5	3	1	3	2	2.8
7	0.86	0.60	30.5	0.68	2	4.5	2.5	1.5	3	1.5	1.8
8	0.84	0.54	22.5	0.67	2.5	4	1.5	1.5	2.5	1	1.7
9	0.86	0.64	16.9	0.46	2	4	2	1	3	2	1.8
10	0.89	0.60	28.8	0.49	2	5	3	1	3	2	2.0
Median	0.86	0.60	23.2	0.62	2.0	4.3	2.3*	1.0	3.0	2.0*	1.9*

^aScored according to the question sheet in Table 1.^bA score of 1 indicates a strong preference for MC over AI contours, a score of 5 indicates the opposite, and a score of 3 indicates an equal preference.

*p < 0.01.

MC, manual contour; AI, constructed clinical target volume; vDSC, volumetric Dice similarity coefficient; sDSC, surface Dice similarity coefficient; HD, Hausdorff Distance; aHD, average Hausdorff distance.

bag structures, there was a greater degree of variation. The constructed bowel bag structures were not qualitatively reviewed in this study. However, similar results, with less degree of agreement between the auto-segmented bowel bag structures and their MC counterparts compared to other pelvic OARs, have been reported in previous reports as well,^{11,12} suggesting that this structure is more difficult to auto-segment.

Evaluation of AI-CTV construction: Although our constructed AI structures were graded poorly, there was almost complete coverage of the AI-GTV by the MC-GTV with margins

(Figure 2a), a test designed to safeguard against an overall miss of the target by the structure constructed by the auto-segmented rectum structure (Figure 1f).

Evaluation of the constructed CTV: In our trial, the MC-CTV was found to be significantly superior to our constructed AI-CTV in both quantitative methods and previously described quantitative metrics, strongly suggesting a low possibility for clinical use. This outcome may be attributed to the generally lower quality of the constructed AI structure compared to the MC; however, other factors could also have influenced the results:

Table 4. Correlation coefficient values of quantitative metrics versus grading test results of the constructed AI-CTV, 5-point and 3-point scales

Quantitative metric		AI-CTV Grading test, 5-point	AI-CTV Grading test, 3-point
vDSC	Correlation coefficient	−0.025	−0.304
	P-value	0.462	0.143
sDSC [2 mm]	Correlation coefficient	−0.126	−0.365
	P-value	0.318	0.100
HD	Correlation coefficient	0.076	0.183
	P-value	0.388	0.261
aHD	Correlation coefficient	0.025	0.365
	P-value	0.462	0.100

AI-CTV, constructed clinical target volume; vDSC, volumetric Dice similarity coefficient; sDSC, surface Dice similarity coefficient; HD, Hausdorff distance; aHD, average Hausdorff distance.

- **Contouring guidelines:** One possible factor that could explain differences observed is the use of different contouring guidelines. While our results did not favour our constructed AI-CTV, similar studies, e.g., Wu *et al.*¹⁴ and Geng *et al.*,⁹ reported no significant differences between their MC and AI-CTV structures. A key distinction is that the model used by Geng was based on a dataset from the study site itself, whereas the Contour+ AI employed in our study was trained on datasets using contouring guidelines different from those adopted by our department. The poor performance of the AI-CTV around the superior rectal artery exemplifies this, as this structure is not fully included in the RTOG guidelines,³ whereas a greater portion is included in the RCR guidance.⁴ Although Contour+ provided an additional structure covering the anterior nodal volume around the internal iliac artery, which could encompass this area, this structure instead encompassed a wider region anteriorly; therefore, it was not incorporated into our constructed AI-CTV.
- **The error resulting from the use of an AI tool based on contouring guidelines that differ from those of the local department is of importance.** In another study, Geng *et al.*²⁰ present a method of adapting commercial software to the clinic's own cases and contouring conventions, offering an alternative approach to reducing these discrepancies rather than rejecting an otherwise useful AI tool.
- **Intra-observer variation:** Although the MC-CTVs were peer-reviewed and approved for treatment in our clinical setting, a few cases still required minor correction before they were deemed perfect by the two reviewers for use in RT. This further highlights the existing interobserver variation in both manual contouring and target review. While this variation may have influenced the results, the AI-CTVs in these cases still performed worse than their MC-CTV counterparts.
- **Choice of qualitative methodology:** There is considerable variation in the designs of grading tests used in similar studies, so direct comparisons of results between different studies should be approached with caution. While Baroudi *et al.* have proposed the use of their five-point scale in future research,⁸ we found, during the design and testing of their scale, a lack of nuance between certain grade levels, i.e., levels

1 and 2, rendered similar conclusions. A five-point scale was employed for the review owing to its lower risk of bias compared to scales with fewer points.⁸ The results were condensed to a three-point scale primarily for reducing parameters for the association analysis. These three categories correspond to clinically relevant distinctions. Since the reduction was symmetrical, the risk of bias was considered low.

- **Bias:** The grading test employed is considered more clinically relevant than the preference test. Additionally, reviewing both AI-CTV and MC-CTV in a blinded manner allowed for the performance of statistical tests of independence, despite the low number of cases. However, the preference test is less susceptible to bias, as it is easier to blind and results in a greater total number of evaluated images. Given the consistent preference for the MC across both tests, we interpret that the influence of bias on our results was minor.

Limitations: Our study is limited by the small sample size, partly due to the time-restricted trial licence of Contour+. Furthermore, only two reviewers were involved. Likert scales are inherently subjective, disproportionate and susceptible to bias. To partially mitigate these limitations, two distinct qualitative methods were employed. Given the significant differences observed between our constructed AI-CTV and MC-CTV structures across both methods, some conclusions can still be drawn.

Conclusion: In conclusion, our study did not find clinical utility of our constructed AI-CTV structures, in contrast to the few previously published data on auto-segmentation of rectal cancer target volumes. However, we observed that the contouring errors in these structures were limited in magnitude and may partly reflect differences in guidelines used to set up the AI tools employed and our local protocol. We hope that the findings of our study will provide novel and valuable guidance for the further developments and applications of auto-segmentation software, as well as for study design and metric interpretation. The use of AI tools still holds considerable promise for reducing the time spent contouring OARs and CTVs in RT planning following future improvements, and we are currently pursuing this objective through further research and development.

Acknowledgements. We thank Mikael Andersson Franko, biostatistician at the Department of Clinical Science and Education, Södersjukhuset, Karolinska Institutet, for helpful input on our statistical analysis and representatives of MVision AI Oy (Helsinki, Finland) for the use of their software in this research report.

Financial support. This work was supported by grants from the Swedish Cancer Society (Cancerfonden), 22 2052 S, and Stockholm Gotland Regional Cancer Centre, project number 2025/44.

Competing interests. The authors declare none.

References

1. Barot S, Liljegren A, Nordenvall C, Blom J, Radkiewicz C. Incidence trends and long-term survival in early-onset colorectal cancer: a nationwide Swedish study. *Ann Oncol* 2025; 36 (11): 1400–1408. <https://doi.org/10.1016/j.annonc.2025.07.019>.
2. Rydbeck D, Azhar N, Blomqvist L, *et al.* Short-term outcomes from the 'Watch and Wait' (WoW) study: prospective cohort study. *BJS Open* 2024; 9 (1): zrae151. <https://doi.org/10.1093/bjsopen/zrae151>.
3. Myerson RJ, Garofalo MC, El Naqa I, *et al.* Elective clinical target volumes for conformal therapy in anorectal cancer: a radiation therapy oncology

- group consensus panel contouring atlas. *Int J Radiat Oncol Biol Phys* 2009; 74 (3): 824–30. <https://doi.org/10.1016/j.ijrobp.2008.08.070>.
4. National rectal cancer intensity-modulated radiotherapy (IMRT) guidance. The Royal College of Radiologists; 2021. Accessed June 9, 2025. <https://www.rcr.ac.uk/our-services/all-our-publications/clinical-oncology-publications/national-rectal-cancer-intensity-modulated-radiotherapy-imrt-guidance/>.
 5. Segedin B, Petric P. Uncertainties in target volume delineation in radiotherapy—are they relevant and what can we do about them? *Radiology and Oncology* 2016; 50 (3): 254–262. <https://doi.org/10.1515/raon-2016-0023>.
 6. van Herk M. Errors and margins in radiotherapy. *Semin Radiat Oncol* 2004; 14 (1): 52–64. <https://doi.org/10.1053/j.semradonc.2003.10.003>.
 7. Doolan PJ, Charalambous S, Roussakis Y, et al. A clinical evaluation of the performance of five commercial artificial intelligence contouring systems for radiotherapy. *Frontiers in Oncology* 2023; 13: 1213068. <https://doi.org/10.3389/fonc.2023.1213068>.
 8. Baroudi H, Brock KK, Cao W, et al. Automated contouring and planning in radiation therapy: what is ‘clinically acceptable’? *Diagnostics* 2023; 13 (4): 667. <https://doi.org/10.3390/diagnostics13040667>.
 9. Geng J, Zhu X, Liu Z, et al. Towards deep-learning (DL) based fully automated target delineation for rectal cancer neoadjuvant radiotherapy using a divide-and-conquer strategy: a study with multicenter blind and randomized validation. *Radiat Oncol* 2023; 18 (1): 164. <https://doi.org/10.1186/s13014-023-02350-0>.
 10. Matoska T, Patel M, Liu H, Beriwal S. Review of deep learning based autosegmentation for clinical target volume: current status and future directions. *Advances in Radiation Oncology* 2024; 9 (5): 101470. <https://doi.org/10.1016/j.adro.2024.101470>.
 11. Song Y, Hu J, Wu Q, et al. Automatic delineation of the clinical target volume and organs at risk by deep learning for rectal cancer postoperative radiotherapy. *Radiother Oncol* 2020; 145: 186–192. <https://doi.org/10.1016/j.radonc.2020.01.020>.
 12. Men K, Dai J, Li Y. Automatic segmentation of the clinical target volume and organs at risk in the planning CT for rectal cancer using deep dilated convolutional neural networks. *Med Phys* 2017; 44 (12): 6377–6389. <https://doi.org/10.1002/mp.12602>.
 13. Larsson R, Xiong JF, Song Y, et al. Automatic delineation of the clinical target volume in rectal cancer for radiation therapy using three-dimensional fully convolutional neural networks. *Annu Int Conf IEEE Eng Med Biol Soc* 2018; 2018: 5898–5901. <https://doi.org/10.1109/embc.2018.8513506>.
 14. Wu Y, Kang K, Han C, et al. A blind randomized validated convolutional neural network for auto-segmentation of clinical target volume in rectal cancer patients receiving neoadjuvant radiotherapy. *Cancer Medicine* 2022; 11 (1): 166–175. <https://doi.org/10.1002/cam4.4441>.
 15. Rhee DJ, Akinfenwa CPA, Rigaud B, et al. Automatic contouring QA method using a deep learning—based autocontouring system. *Journal of Applied Clinical Medical Physics* 2022; 23 (8): e13647. <https://doi.org/10.1002/acm2.13647>.
 16. Yu C, Anakwenze CP, Zhao Y, et al. Multi-organ segmentation of abdominal structures from non-contrast and contrast enhanced CT images. *Scientific Reports* 2022; 12 (1): 19093. <https://doi.org/10.1038/s41598-022-21206-3>.
 17. Almberg SS, Lervag C, Frengen J, et al. Training, validation, and clinical implementation of a deep-learning segmentation model for radiotherapy of loco-regional breast cancer. *Radiother Oncol* 2022; 173: 62–68. <https://doi.org/10.1016/j.radonc.2022.05.018>.
 18. Taha AA, Hanbury A. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. *BMC Med Imaging* 2015; 15: 29. <https://doi.org/10.1186/s12880-015-0068-x>.
 19. Van Dijk LV, Van Den Bosch L, Aljabar P, et al. Improving automatic delineation for head and neck organs at risk by Deep Learning Contouring. *Radiotherapy and Oncology* 2020; 142: 115–123. <https://doi.org/10.1016/j.radonc.2019.09.022>.
 20. Geng J, Sui X, Du R, et al. Localized fine-tuning and clinical evaluation of deep-learning based auto-segmentation (DLAS) model for clinical target volume (CTV) and organs-at-risk (OAR) in rectal cancer radiotherapy. *Radiat Oncol* 2024; 19 (1): 87. <https://doi.org/10.1186/s13014-024-02463-0>.