Advances in Radiation Oncology (2024) 9, 101483

advances
in radiation oncology

www.advancesradonc.org

Scientific Article

Evaluation of Deep Learning Clinical Target Volumes Auto-Contouring for Magnetic Resonance Imaging-Guided Online Adaptive Treatment of Rectal Cancer

Nicole Ferreira Silvério, MSc, Wouter van den Wollenberg, PhD, Anja Betgen, MSc, Lisa Wiersema, BSc, Corrie Marijnen, Prof. Dr., Femke Peters, MD, PhD, Uulke A. van der Heide, Prof. Dr., Rita Simões, PhD, and Tomas Janssen, PhD*

Department of Radiation Oncology, The Netherlands Cancer Institute, Amsterdam, The Netherlands

Received 5 December 2023; accepted 11 February 2024

Purpose: Segmentation of clinical target volumes (CTV) on medical images can be time-consuming and is prone to interobserver variation (IOV). This is a problem for online adaptive radiation therapy, where CTV segmentation must be performed every treatment fraction, leading to longer treatment times and logistic challenges. Deep learning (DL)-based auto-contouring has the potential to speed up CTV contouring, but its current clinical use is limited. One reason for this is that it can be time-consuming to verify the accuracy of CTV contours produced using auto-contouring, and there is a risk of bias being introduced. To be accepted by clinicians, auto-contouring must be trustworthy. Therefore, there is a need for a comprehensive commissioning framework when introducing DL-based auto-contouring in clinical practice. We present such a framework and apply it to an in-house developed DL model for auto-contouring of the CTV in rectal cancer patients treated with MRI-guided online adaptive radiation therapy.

Methods and Materials: The framework for evaluating DL-based auto-contouring consisted of 3 steps: (1) Quantitative evaluation of the model's performance and comparison with IOV; (2) Expert observations and corrections; and (3) Evaluation of the impact on expected volumetric target coverage. These steps were performed on independent data sets. The framework was applied to an in-house trained nnU-Net model, using the data of 44 rectal cancer patients treated at our institution.

Results: The framework established that the model's performance after expert corrections was comparable to IOV, and although the model introduced a bias, this had no relevant impact on clinical practice. Additionally, we found a substantial time gain without reducing quality as determined by volumetric target coverage.

Conclusions: Our framework provides a comprehensive evaluation of the performance and clinical usability of target auto-contouring models. Based on the results, we conclude that the model is eligible for clinical use.

Introduction

Radiation therapy is crucial in the multidisciplinary treatment of rectal cancer, focusing on reducing local recurrence rates and downstaging tumors before surgery.[1,2]

Sources of support: None.

Data sharing statement: All patient data were acquired and accessed via the prospective Momentum registry. Access to Momentum data is available only to members of the Momentum consortium.

*Corresponding author: Tomas Janssen, PhD; Email: t.janssen@nki.nl

https://doi.org/10.1016/j.adro.2024.101483

Magnetic resonance imaging-guided online adaptive radiation therapy (MRIgOART) allows for daily adaptation to anatomic changes through the use of high-resolution, high contrast images,[3,4] significantly reducing required margins for intrafraction target motion resulting in approximately 30% reduction of the high-dose volume,[3,5] which potentially lowers treatment-related toxicities.[6,7]

A crucial step in MRIgOART is the recontouring of clinical target volumes (CTVs). However, manual contouring of the target and organs at risk (OARs) for radiation therapy is a labor-intensive and time-consuming process that is prone to variability among clinicians. As a result, it is often considered the weakest link in the radiation therapy treatment chain.[8-11] MRIgOART presents additional challenges for CTV contouring. First, it introduces time pressure on the contouring process, as prolonged contouring may result in organ and patient motion, which can compromise treatment accuracy.[12] Second, it requires the presence of a physician or highly skilled technician during treatment, which affects staff efficiency and treatment costs. Third, CTV contouring must be performed for every treatment session, adding to clinical staff workload. Therefore, the development of accurate and efficient CTV contouring methods is essential for the widespread clinical adoption of MRIgOART.

Recent advancements in artificial intelligence, particularly deep learning (DL) with convolutional neural networks, show promise for automated medical image contouring.[13,14] DL models can enhance contouring consistency and speed.[15,16] Although automated contouring of OARs is becoming more common in radiation therapy,[17,18] automated target contouring is less frequently studied.[19] In particular for rectal cancer, although a few studies have focused on segmenting the primary tumor from multimodal MRI,[20-22] CTV segmentation remains less explored.[23,24]

Despite DL-based auto-contouring's potential, its clinical use is limited.[25] One of the reasons is that it can be difficult to assess the quality of the auto-contours. In particular, for CTV auto-contouring, validation of the DL model requires careful consideration before clinical introduction. To be eligible for clinical use, a DL model for contouring must produce robust and accurate delineations that are suitable for treatment. However, often used geometric segmentation metrics such as the Dice coefficient or mean surface distance fail to answer this question for at least 2 reasons. First, because of known interobserver variability (IOV), a certain amount of mismatch is expected.[26] Second, a CTV is typically not irradiated as such, but instead a planning target volume (PTV) margin is applied. Therefore, the question becomes whether any residual errors in the auto-contours require an increase of the PTV margin to ensure target coverage. This is especially relevant because the auto-contouring model might introduce a (small) systematic bias, even after manual correction. The potential clinical impact of this bias should be assessed. Additionally, geometric segmentation metrics have been shown to not correlate well with clinical time-gain after the introduction of automated contouring.[27] These factors make the assessment of the clinical eligibility of the auto-contouring model for clinicians highly complex, whereas for a model to be accepted by clinicians, auto-contouring must be trustworthy. Therefore, there is a need for a comprehensive commissioning framework when introducing DL-based auto-contouring in clinical practice.

Our work aims to develop a framework to bridge the gap between a CTV auto-contouring model that performs well on geometric metrics and a model that is eligible for clinical use. We have proposed and applied a comprehensive commissioning framework for DL-based CTV auto-contouring, addressing 3 critical questions:

1. How do the differences between (corrected) auto-contours and a predefined reference contour compare with IOV?
2. How do experts evaluate auto-contours, and what time savings do they offer?
3. How do (corrected) auto-contours affect PTV margins and target coverage?

This framework provides a clinically relevant analysis of CTV auto-contouring quality, facilitating its safe implementation. As a case study, we have applied this framework to our in-house mesorectum CTV auto-contouring model for MRIgOART in rectal cancer treatments.

Methods and Materials

Imaging data

In this study we retrospectively included 44 patients with intermediate risk or locally advanced rectal cancer treated who were treated in our institution on the Unity MR-Linac (Elekta AB, Stockholm). Thirty-six patients were treated with 5 × 5 Gy, and 8 patients were treated with 25 × 2 Gy; 33 patients (32 short course, 1 long course) were treated with an adapt to shape (ATS) procedure,[28] and the remaining patients (4 short course, 7 long course) were treated using an adapt to position (ATP) with a library of plans approach.[29] All patients in the study were enrolled in the Momentum prospective registration study (NCT04075305)[30] and as such gave written informed consent for the retrospective use of their data.

For the ATS treatments, 4 3D T2 weighted MRI acquisitions were acquired for each treatment fraction: an MRI used for adaptation ($\text{MRI}_{\text{adapt}}$), an MRI for verification made just before treatment (MRI_{ver}), an MRI post treatment (MRI_{post}), and an MRI during treatment ($\text{MRI}_{\text{during}}$). All acquisitions had a Field of View of 400 × 448 × 249 mm^3, repetition time (TR) of 1300 ms, and echo time (TE) of 128 ms. Voxel size was (1.2 mm)^3 for MRI_{adapt} and 1.2 × 1.2 × 2.4 mm^3 for MRI_{ver}, MRI_{during}, and MRI_{post}, resulting in acquisition times of approximately 6 and 3 minutes respectively. For the ATP patients, MRI_{during} was not acquired. For the long course treatments, only the first fraction of each week was included.

For each patient a planning CT (MRI based pseudo-CT[31]) acquired a week before treatment was also available.

Contouring data

Figure 1a and b provide a graphical overview of the different contours we distinguished in this work. On the planning CT the target volumes and OARs were contoured according to national guidelines.[32] In the ATS workflow in clinical practice, a deformable or rigid registration from the planning CT to daily acquired MRI_{adapt} was performed, and the structures were propagated accordingly to MRI_{adapt}. Subsequently, the contours were manually verified and corrected before treatment. These contours we designate as the *clinical standard.*

The *clinical standard* data were acquired under time pressure by a variety of clinicians during real patient treatment. To obtain a highly self-consistent data set, the mesorectum CTV was also contoured retrospectively by 2 experienced radiation technology therapists (RTT). For each fraction, the clinical contours on MRI_{adapt} were checked and carefully adjusted. Subsequently, the corrected contours were rigidly copied to the MRI_{ver}, MRI_{post} and, if present, MRI_{during}, and manually adjusted. All scans of one patient were contoured by the same RTT. Contours were verified and, if needed, corrected by a radiation oncologist with over 10 years of experience. These contours we designate as the *gold standard.*

Contours produced by the auto-contouring model are designated as *predicted.* Contours that were manually verified and corrected are designated as *corrected.*

Data sets

The data were split randomly and on the patient level into subsets. In this study we used the following data sets:

- Training data: the auto-contouring model was trained on 20 patients (332 scans) using the *gold standard* contours.
- Validation data: 5 patients (74 scans) using the *gold standard* contours. During model development this set was used to choose the most appropriate training strategy.
- Test data: 5 patients (77 scans) using the *gold standard* contours.
- Commissioning data: 19 patients (including the 5 test data patients) with in addition *prediction* and *correction* contours on MRI_{adapt} and *gold standard* contours on MRI_{post} (93 scans). For the patients who were treated with an ATS procedure, there were *clinical standard* contours present as well (79 scans).
- IOV data: 10 patients (all included in the training data) were contoured on MRI_{adapt} of the first treatment fraction by 5 experts (1 resident radiation oncologist and 4 RTTs, trained and certified by a radiation oncologist to perform daily CTV contour adaptation in an online adaptive workflow); for each

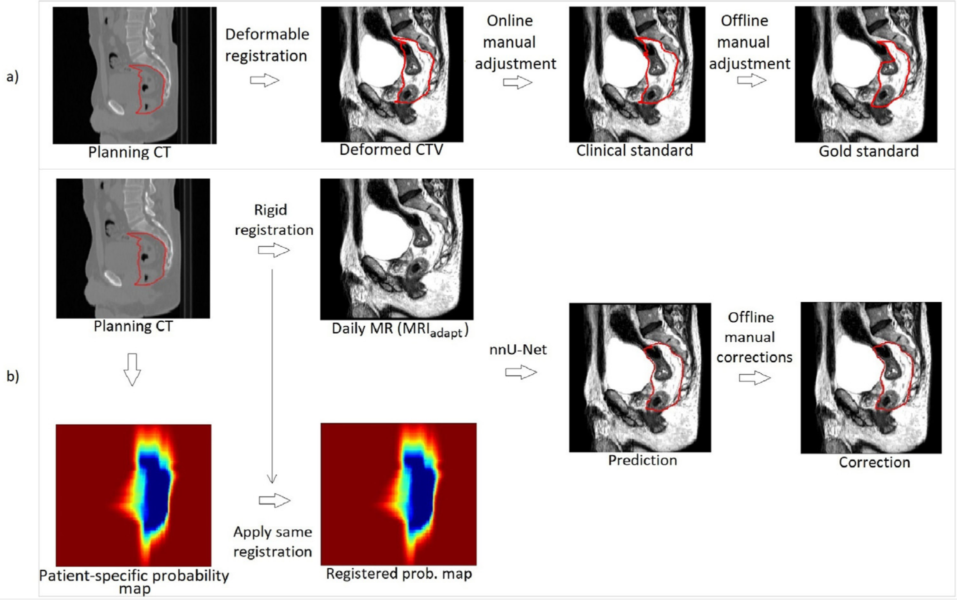

Figure 1 Overview of how the contours used in this study have been acquired. Panel a shows the clinical and gold standard contours. Panel b shows the prediction and correction contours.

MRI$_{adapt}$ there were 5 mesorectum CTV contours present.

Model architecture and training

The nnU-Net framework[33] was used to train an auto-contouring model. nnU-Net is a state-of-the-art segmentation framework that automatically configures model parameters for the data and task at hand. The default settings of the framework were used for training; the loss function was a combination of the Dice and cross-entropy. Stochastic gradient descent with Nesterov momentum (μ = 0.99) was used as an optimizer with a poly learning rate schedule and an initial learning rate of 0.1. For inference we deviated from the default settings by making use of the best performing weight configuration instead of the resulting weights at the end of training as a form of early stopping. The model was trained to segment the mesorectum CTV on MRI scans. To improve model performance, we provided the model with a patient-specific probability map as a second input channel, indicating the prior voxelwise probability of the mesorectum being present. This probability map was based on a population standard deviation map, generated from mesorectum contours on daily CBCT images from prior work.[29] A patient specific probability map was made by combining the population map with the patient specific mesorectum CTV as contoured on the planning CT. Using a rigid match between the planning CT and the MRI$_{adapt}$, the probability map was subsequently matched to the MRI$_{adapt}$ (see Fig. 1b). Because the patient specific contour implicitly contains the physician intent for each patient, this probability map effectively propagates the expected physician intent to each online adaptive fraction.

Evaluation and commissioning framework

To assess the quality of CTV auto-contouring we proposed the following framework:

Quantitative evaluation of model performance and comparison with IOV

Quantitative evaluation was done using commonly used geometric metrics: the Sørensen-Dice coefficient (Dice), 95% Hausdorff distance (HD$_{95}$), Surface Dice at 3 mm (SD$_3$), and bidirectional mean surface distance (MSD). These metrics were calculated using DeepMind's Python package (https://github.com/deepmind/surface-distance). Model performance was quantified by comparing the *predicted* mesorectum contour with the *gold standard* contour on the test data set. To understand local variation in performance, this analysis was repeated while taking the most cranial 20%, middle 60%, and most caudal 20% of *gold standard* delineated slices into account. The local analysis was done using HD$_{95}$ and MSD only, because the Dice and SD$_3$ are correlated with volume, which makes the interpretation of these metrics evaluated on different volumes prone to confusion.

To benchmark the quantitative results, the same geometric metrics both globally and per subregion were determined on the IOV data set. Each metric was calculated for each observer pair and per patient the average over all observer pairs was determined. By quantifying the IOV in this way, we explicitly chose not to compare the different observers with a *gold standard.* The assumption here was that each observer in the IOV set produced a delineation that is of sufficient quality for clinical use and therefore the comparison against this set provided an indication of clinically equivalent contours.

Differences between model performance and IOV were tested for significance using a 2-sided Mann-Whitney *U* test using α = 0.05.

Expert observations and corrections

To evaluate the use of auto-contouring in practice, a single experienced RTT corrected the *predicted* contour on 93 scans (commissioning data). A subjective assessment of the quality of the prediction was made using a scale of 0 to 4 (0 = no corrections needed; 1 = small adjustments, low impact; 2 = mix of small and large adjustments; 3 = large, high impact adjustments needed; 4 = unusable). The categories were discussed with the RTT beforehand. Time spent correcting was measured manually. The time required for mesorectum CTV correction in the clinical ATS workflow was measured for 13 consecutive patients (51 fractions) not further included in this study.

The corrections made were quantified using the same metrics as in step 1, both globally and per subregion, by comparing the *predicted* CTV with the *corrected* CTV. To quantify the quality of the corrections, the *corrected* CTV was compared with the *clinical standard* CTV and benchmarked against the IOV results, similar to as in step 1.

Evaluate the impact on expected target coverage

The expected target coverage for a certain PTV margin was determined by assuming all residual uncertainty was intrafraction motion, similar as in Kensen et al.[5] The mesorectum CTV delineated on MRI$_{adapt}$ was anisotropically (with the anterior expansion being 1.6 times the expansion in the other directions) expanded to a PTV, using a range of margins of 0.5 to 10 mm (stepsize of 0.5 mm). Next, the volumetric overlap of the PTV with the *gold standard* CTV on MRI$_{post}$ was determined. For every patient the volumetric overlap was averaged over all treatment fractions. Coverage reached by at least 90% of the population was reported. Expected target coverage was determined using the *predicted*, *corrected*, and *clinical standard* CTV from the commissioning data set.

Table 1 Overview of evaluation steps and data used for each step

		Quantitative evaluation				Expert evaluation	Expected target coverage		
Test data	Predictions		▨						
	Gold standard		▨						
Commissioning data	Predictions			▨		▨	▨		
	Corrections			▨	▨			▨	
	Clinical standard				▨				▨
	Gold standard (MRI_{post} only)						▨	▨	▨
IOV data	Manual contours (x 5)	▨							

Note: The gray boxes show which contouring data sets are compared for each evaluation step.

Table 1 shows an overview of which data set was used for each evaluation step. Figure 1a and b show an overview of how every type of CTV contour was acquired.

Results

In the training data MRI_{during} was not always available. Three patients missed MRI_{during} for one fraction, 2 patients missed 2 fractions, and one patient missed 4 fractions. Also, for one patient a single fraction was missing completely. For one patient in both the test and commissioning data, none of the scans were available for one single fraction. Additionally, one patient in the commissioning data (but not included in the test data) also only had data present for 4 fractions.

Quantitative evaluation of model performance and comparison with IOV

The geometric comparison of the predicted contours with the *gold standard* contours and the IOV is shown in Table 2. Model performance was slightly worse compared with IOV, with significant differences found for Dice, SD_3, and MSD, but not for HD_{95}.

The local differences in performance for the cranial 20%, middle 60%, and caudal 20% of the contours is presented in Table 3. These results show that both the model performance and the IOV was worst in the caudal region of the contours, whereas the best performance was in the middle region.

Expert observations and corrections

Of the 93 mesorectum CTV contours presented to the experts, 9 needed no corrections (score 0), and 2 were deemed unusable (score 4). The majority of contours (56 contours) required only minor corrections (score 1, see Table 4). Average time spent verifying and correcting the predicted mesorectum CTV contours was 3m08s, (range, 0m35s-8m35s), compared with an average correction time of 7m34s (range, 2m40s-14m58s), in the current

Table 2 Geometric metrics evaluating model performance and interobserver variation using the test data set and IOV data set (section 2.3)

	Prediction vs Gold standard	IOV analysis	*P* value (Predicted vs Gold standard) vs IOV
Dice	0.89 (0.87-0.90)	0.91 (0.90-0.92)	0.01
SD_3	0.81 (0.78-0.84)	0.86 (0.85-0.91)	<0.01
HD_{95} (mm)	11 (7-14)	10 (6-13)	0.62
MSD (mm)	2 (2-2)	2 (1-2)	0.03

Abbreviations: HD_{95} = 95% Hausdorff distance; IOV = interobserver variation; MSD = mean surface distance; SD_3 = surface dice at 3 mm.
Notes: Values are reported as median (interquartile range). *P* values are calculated using the Mann-Whitney *U* test.

Table 3 Local geometric metrics evaluating model performance and interobserver variation (IOV) using the test data set and IOV data set (section 2.3)

	Prediction vs. Gold Standard			IOV analysis		
	Cranial 20%	Middle 60%	Caudal 20%	Cranial 20%	Middle 60%	Caudal 20%
HD_{95} (mm)	9 (6-14)	3 (3-6)	16 (9-22)	6 (4-8)	3 (2-6)	10 (7-12)
MSD (mm)	2 (1-3)	1 (1-1)	3 (2-5)	1 (1-1)	1 (1-1)	2 (2-2)

Abbreviations: HD_{95} = 95% Hausdorff distance; IOV = interobserver variation; MSD = mean surface distance.
Note: Values are reported as median (interquartile range).

Table 4 Qualitative scores assigned to the predicted contours by the expert correcting them

Score	Occurrence
0: no corrections needed	9
1: small adjustments, low impact	56
2: mix of small and large adjustments	26
3: large, high impact adjustments needed	0
4: unusable contour	2

clinical ATS workflow. See the supplemental material for examples of predicted contours and their corrections.

The *predicted* contours were compared with their respective *corrections*, and the *corrected* contours were compared with the *clinical standard.* The results for the global comparison are shown in Table 5. With a Dice score of 0.98, these results showed a high degree of similarity between the *predicted* and *corrected* contours. When comparing the similarity between the *corrected* and *clinical standard* contours with the IOV from step 1, only a significant difference was found for the HD_{95}, with the IOV having the larger HD_{95}.

The local differences between the *predictions* and their *corrections* as well as the local differences between the *corrections* and the *clinical standard* contours are shown in Table 6. Similar to results shown in Table 3, it was found that the best performance was in the middle region of the contour.

Evaluate the impact on expected target coverage

In the last stage of the evaluation framework, the volumetric target coverages of the *predicted, corrected*, and *clinical standard* contours were compared (Fig. 2). After

Table 5 Geometric metrics evaluating manual corrections to model predictions using the commissioning data set and interobserver variation data set

	Prediction vs. Correction	Correction vs. Clinical	IOV analysis	*P* value (Correction vs Clinical) vs IOV
Dice	0.98 (0.95-0.99)	0.92 (0.90-0.93)	0.91 (0.90-0.92)	0.35
SD_3	0.95 (0.91-0.98)	0.88 (0.85-0.91)	0.86 (0.85-0.91)	0.57
HD_{95} (mm)	4 (1-9)	6 (4-8)	10 (6-13)	0.01
MSD (mm)	0 (0-1)	1 (1-2)	2 (1-2)	0.28

Abbreviations: HD_{95} = 95% Hausdorff distance; IOV = interobserver variation; MSD = mean surface distance.
Notes: Values are reported as median (interquartile range). Results for IOV are a repetition from Table 2. *P* values are calculated using the Mann-Whitney *U* test.

Table 6 Local geometric metrics evaluating manual corrections to model predictions using the commissioning data set and interobserver variation data set

	Prediction vs. Correction			Correction vs. Clinical		
	Cranial 20%	Middle 60%	Caudal 20%	Cranial 20%	Middle 60%	Caudal 20%
HD_{95} (mm)	3 (1-8)	1 (0-3)	3 (1-12)	5 (4-7)	3 (3-4)	5 (4-9)
MSD (mm)	0 (0-1)	0 (0-0)	0 (0-1)	1 (1-2)	1 (1-1)	1 (1-2)

Abbreviations: HD_{95} = 95% Hausdorff distance; MSD = mean surface distance.
Notes: Values are reported as median (interquartile range).

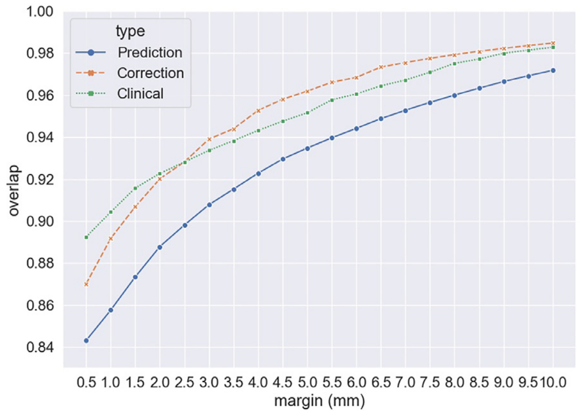

Figure 2 Posttreatment overlap averaged over the fractions reached by 90% of the population. Planning target volume margins were anisotropic, with the anterior margin 1.6 times the reported margin. Clinically, a margin of 5 mm (8 mm anterior) is used. In blue are results for the predicted contours; in orange, results for the corrected contours; in green, results for the clinical standard contours.

correcting, the target coverage increased compared with the uncorrected predictions, and the coverage for the *corrected* contours was very similar to the coverage obtained by the *clinical standard* contours. With our institution's clinically used PTV margin of 5 mm (8 mm anterior), a relative volumetric coverage was obtained of 0.93, 0.96, and 0.95 when using the *predicted, corrected*, and *clinical* CTV respectively.

Discussion

In this work, we developed a framework to evaluate the performance and clinical eligibility of a deep learning (DL) contouring model for clinical target volume (CTV) auto-contouring. The framework aimed to address 3 main questions: (1) How do the differences between (corrected) auto-contours and a predefined reference contour compare with IOV? (2) How do experts evaluate the auto-contours, and what is the time saved by using auto-contouring? (3) What is the impact of using the (corrected) auto-contours in terms of the relation between the PTV margins used and the expected target coverage?

To address these questions, we compared the model's performance to a predefined ground truth and to the IOV, both globally and locally. We also had an expert review and correct the auto-contours and tracked the timing to establish a potential time gain in clinical practice. We compared the corrected contours to a predefined reference and to the predicted contour and evaluated the clinical relevance of using the different CTV contours based on geometric target coverage. The approach required the addition of 2 data sets to the standard used train/test/validate data,[34] namely a data set showing IOV and a *commissioning* data set that included model predictions, corrections, and an independent reference delineation (in our case the *clinical standard* contours).

Literature evaluating the performance of (auto-)contouring models focuses primarily on using geometric measures. However, these metrics may not always reflect clinical relevance, such as dosimetric parameters[35] or editing time.[36] Additionally, it can be challenging to evaluate the impact of IOV and the potential bias introduced by the model in clinical practice. To supplement these analyses, various methods have been proposed in the literature including: a Turing test,[37] expert evaluations through scoring, investigation of time saved in recontouring,[36] evaluation of the dosimetric impact on treatment planning or plan evaluation,[38] and IOV studies.[8] These approaches aim to provide an understanding of the performance and clinical usability of auto-contouring models.

However, to establish clinical usability, a comprehensive framework, combining manual model corrections, IOV, and clinical impact is advisable. Frederick et al.[38] proposed such a framework to evaluate contour propagation, in which an automated treatment plan based on propagated contours was compared with treatment plans based on contours delineated by multiple observers using dosimetric measures. However, this approach requires multiple observers to delineate *every* scan in the evaluation data set, which may not always be feasible and may limit the available evaluation data. Our approach separates the IOV and commissioning data to evaluate a large data set and include IOV data simultaneously. Additionally, our analysis supplements this framework by also considering the corrections made to the contours (which would be standard in clinical practice). Although dosimetric impact is important, for CTV contouring it is also necessary to consider PTV margins and expected coverage to assess the clinical impact. Our framework provides a more comprehensive evaluation of the performance and clinical usability of the model.

As a case study, we applied the framework to the task of mesorectum CTV delineation for use in MRIgOART. Our results are consistent with existing literature: in Savjani et al.[19] the best performing models had a Dice coefficient of 0.73-0.94, and others[23,24] report a mean Dice of 0.90 for rectal cancer CTV delineation, making our performance (Dice: 0.89) comparable. A median Dice of 0.93 was reported by White et al.[39] for mesorectum delineation IOV, similar to our finding of 0.91. Our finding that variation was largest in the cranial/caudal region was consistent with Nijkamp et al.[26]

In this work, we did not consider dosimetric metrics but evaluated expected target coverage based on geometry only by calculating the volumetric overlap between the mesorectum PTV on MRI$_{\text{adapt}}$ and the *gold standard*

CTV on MRI_{post}. We performed this calculation for different PTV margins without explicitly choosing a required coverage criterion. In rectal cancer treatment planning, the 95% prescription isodose line is typically tightly conformal to the PTV, which makes the calculated volumetric overlap equivalent to the volume receiving 95% of the prescribed dose.[5] However, when applying our framework to other treatment sites, this choice should be carefully reconsidered, and it might be necessary to evaluate expected target coverage by dose accumulation on the posttreatment image.

One notable finding was that the HD_{95} before correction was the only metric not significantly different compared with IOV, whereas the opposite was true after correction. The HD_{95} is sensitive to the largest errors, suggesting each observer made different choices despite clinical consensus guidelines, with an IOV upper quartile HD_{95} of up to 13 mm, mainly in the caudal extent of the CTV. This uncertainty propagates to the model prediction, and we hypothesize that the large spread in the HD_{95} makes the difference insignificant. However, after corrections the HD_{95} between the corrected and clinical contour improved with respect to the IOV. This suggests that the model's bias can be effectively corrected through manual review.

In summary, we presented a framework for the evaluation and commissioning of DL auto-contouring models for clinical targets. We applied this framework to a DL model for mesorectum CTV auto-contouring. The framework established that the model's performance was comparable to IOV, and although the model introduced a bias, this had no relevant impact on clinical practice. Additionally, we found a substantial time gain without diminishing quality, determined by volumetric target coverage. Based on these results, we conclude that the model is eligible for clinical use.

Disclosures

The Netherlands Cancer Institute receives institutional funding and research support from Elekta AB (Stockholm) and research support from Philips Healthcare. Femke Peters participated in a Data Safety Monitoring Board for a study into BioXmark fiducials for rectum boost.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the author(s) used ChatGPT (https://chat.openai.com) to enforce the word count limit. After using this tool/service the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

Supplementary materials

Supplementary material associated with this article can be found in the online version at doi:10.1016/j.adro.2024.101483.

References

1. Sauer R, Liersch T, Merkel S, et al. Preoperative versus postoperative chemoradiotherapy for locally advanced rectal cancer: Results of the German CAO/ARO/AIO-94 randomized phase III trial after a median follow-up of 11 years. *J Clinical Oncol*. 2012;30:1926-1933.
2. Kapiteijn E, Marijnen CAM, Nagtegaal ID, et al. Preoperative radiotherapy combined with total mesorectal excision for resectable rectal cancer. *N Engl J Med*. 2001;345:638-646.
3. Intven MPW, de Mol van Otterloo SR, et al. Online adaptive MR-guided radiotherapy for rectal cancer; feasibility of the workflow on a 1.5T MR-linac: Clinical implementation and initial experience. *Radiother Oncol*. 2021;154:172-178.
4. Kerkmeijer LGW, Valentini V, Fuller CDD, Slotman BJ. Editorial: Online adaptive MR-guided radiotherapy. *Front Oncol*. 2021;11: 748685.
5. Kensen C, Janssen T, Betgen A, Wiersema L, Peters F, Remeijer P, et al. Effect of intrafraction adaptation on PTV margins for MRI guided online adaptive radiotherapy for rectal cancer. *Radiat Oncol*. 2022;17(110).
6. Holyoake DLP, Partridge M, Hawkins MA. Systematic review and meta-analysis of small bowel dose—volume and acute toxicity in conventionally-fractionated rectal cancer radiotherapy. *Radiother Oncol*. 2019;138:38-44.
7. Appelt AL, Bentzen SM, Jakobsen A, Vogelius IR. Dose-response of acute urinary toxicity of long-course preoperative chemoradiotherapy for rectal cancer. *Acta Oncol*. 2015;54:179-186.
8. Brouwer CL, Steenbakkers RJHM, van den Heuvel E, et al. 3D variation in delineation of head and neck organs at risk. *Radiat Oncol*. 2012;7:32.
9. Steenbakkers R, Brouwer C, Bourhis J, et al. Improvement of delineation quality of organs at risk in head and neck using the consensus guidelines. *Radiother Oncol*. 2016;119:S137-S138.
10. Nijkamp J, de Jong R, Sonke JJ, Remeijer P, van Vliet C, Marijnen C. Target volume shape variation during hypo-fractionated preoperative irradiation of rectal cancer patients. *Radiat Oncol*. 2009;92 (2):202-209.
11. Nijkamp J, de Haas-Kock DFM, Beukema JC, Neelis KJ, Woutersen D, Ceha H, et al. Target volume delineation variation in radiotherapy for early stage rectal cancer in the Netherlands. *Radiat Oncol*. 2012;102(1):14-21.
12. Kleijnen JPJE, Van Asselen B, Burbach JPM, et al. Evolution of motion uncertainty in rectal cancer: Implications for adaptive radiotherapy. *Phys Med Biol*. 2015;61:1-11.
13. Litjens G, Kooi T, Bejnordi BE, et al. A survey on deep learning in medical image analysis. *Med Image Anal*. 2017;42:60-88.
14. Fu Y, Lei Y, Wang T, Curran WJ, Liu T, Yang X. A review of deep learning based methods for medical image multi-organ segmentation. *Phys Med*. 2021;85:107-122.
15. 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. *Radiother Oncol*. 2020;142:115-123.

16. van der Veen J, Willems S, Robben D, Crijns W, Maes F, Nuyts S. Benefits of deep learning for delineation of organs at risk in head and neck cancer. *Radiother Oncol*. 2019;133:68-74.
17. Sharp G, Fritscher KD, Pekar V, et al. Vision 20/20: Perspectives on automated image segmentation for radiotherapy. *Med Phys*. 2014;41: 050902.
18. Cardenas CE, Yang J, Anderson BM, Court LE, Brock KB. Advances in auto-segmentation. *Semin Radiat Oncol*. 2019;29:185-197.
19. Savjani RR, Lauria M, Bose S, Deng J, Yuan Y, Andrearczyk V. Automated tumor segmentation in radiotherapy. *Semin Radiat Oncol*. 2022;32:319-329.
20. Trebeschi S, van Griethuysen JJM, Lambregts DMJ, et al. Deep learning for fully-automated localization and segmentation of rectal cancer on multiparametric MR. *Sci Rep*. 2017;7:5301.
21. Wang J, Lu J, Qin G, et al. Technical note: A deep learning-based autosegmentation of rectal tumors in MR images. *Med Phys*. 2018;45:2560-2564.
22. Wang M, Xie P, Ran Z, et al. Full convolutional network based multiple side-output fusion architecture for the segmentation of rectal tumors in magnetic resonance images: A multi-vendor study. *Med Phys*. 2019;46:2659-2668.
23. 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;44:6377-6389.
24. 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 Med*. 2022;11:166-175.
25. Brouwer CL, Dinkla AM, Vandewinckele L, et al. Machine learning applications in radiation oncology: Current use and needs to support clinical implementation. *Phys Imaging Radiat Oncol*. 2020;16:144-148.
26. Nijkamp J, de Haas-Kock DFM, Beukema JC, Neelis KJ, Woutersen D, Ceha H, et al. Target volume delineation variation in radiotherapy for early stage rectal cancer in the Netherlands. *Radiat. Oncol*. 2012;102(1):14-21.
27. Vaassen F, Hazelaar C, Vaniqui A, 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.
28. Winkel D, Bol GH, Kroon PS, et al. Adaptive radiotherapy: The Elekta Unity MR-linac concept. *Clin Transl Radiat Oncol*. 2019;18:54-59.
29. Beekman C, van Triest B, van Beek S, Sonke JJ, Remeijer P. Margin and PTV volume reduction using a population based library of plans strategy for rectal cancer radiotherapy. *Med Phys*. 2018;45(10):4345-4354.
30. de Mol van Otterloo SR, Christodouleas JP, Blezer ELA, Akhiat H, Brown K, Choudhury A, et al. The MOMENTUM Study: An International Registry for the Evidence-Based Introduction of MR-Guided Adaptive Therapy. *Front Oncol*. 2020;10.
31. Köhler M, Vaara T, Grootel MV, Hoogeveen R, Kemppainen R, Renisch S. MR-only simulation for radiotherapy planning. *Philips White Paper*. 2015.
32. Valentini V, Gambacorta MA, Barbaro B, et al. International consensus guidelines on clinical target volume delineation in rectal cancer. *Radiother Oncol*. 2016;120:195-201.
33. Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. *Nat Methods*. 2021;18:203-211.
34. Vandewinckele L, Claessens M, Dinkla AM, et al. Overview of artificial intelligence-based applications in radiotherapy: Recommendations for implementation and quality assurance. *Radiother Oncol*. 2020;153:55-66.
35. Kaderka R, Gillespie EF, Mundt RC, et al. Geometric and dosimetric evaluation of atlas based auto-segmentation of cardiac structures in breast cancer patients. *Radiother Oncol*. 2019;131:215-220.
36. Vaassen F, Hazelaar C, Vaniqui A, et al. Evaluation of measures for assessing time-saving of automatic organ-at-risk segmentation in radiotherapy. *Phys Imaging Radiat Oncol*. 2019;13:1-6.
37. Gooding MJ, Smith AJ, Tariq M, et al. Comparative evaluation of autocontouring in clinical practice: A practical method using the Turing test. *Med Phys*. 2018;45:5105-5115.
38. Frederick A, Roumeliotis M, Grendarova P, et al. A framework for clinical validation of automatic contour propagation: Standardizing geometric and dosimetric evaluation. *Pract Radiat Oncol*. 2019;9:448-455.
39. White I, Hunt A, Bird T, et al. Interobserver variability in target volume delineation for CT/MRI simulation and MRI-guided adaptive radiotherapy in rectal cancer. *Br J Radiol*. 2021;94: 20210350.