Advances in Radiation Oncology (2024) 9, 101394

advances in radiation oncology

www.advancesradonc.org

Scientific Article

Artificial Intelligence–Based Autosegmentation: Advantages in Delineation, Absorbed Dose-Distribution, and Logistics

Gustavo R. Sarria, MD,[a,*] Fabian Kugel, PhD,[a] Fred Roehner, PhD, MD,[a] Julian Layer, MD,[a,b] Cas Dejonckheere, MD,[a] Davide Scafa, MD,[a] Muemtaz Koeksal, MD,[a] Christina Leitzen, MD,[a] and Leonard Christopher Schmeel, MD[a]

[a]*Department of Radiation Oncology; and* [b]*Institute of Experimental Oncology, University Hospital Bonn, University of Bonn, Bonn, Germany*

Received 20 June 2023; accepted 9 October 2023

Purpose: The study's purpose was to compare the performance of artificial intelligence (AI) in auto-contouring compared with a human practitioner in terms of precision, differences in dose distribution, and time consumption.

Methods and Materials: Datasets of previously irradiated patients in 3 different segments (head and neck, breast, and prostate cancer) were retrospectively collected. An experienced radiation oncologist (MD) performed organs-at-risk (OARs) and standard clinical target volume delineations as baseline structures for comparison. AI-based autocontours were generated in 2 additional CT copies; therefore, 3 groups were assessed: MD alone, AI alone, and AI plus MD corrections (AI+C). Differences in Dice similarity coefficient (DSC) and person-hour burden were assessed. Furthermore, changes in clinically relevant dose-volume parameters were evaluated and compared.

Results: Seventy-five previously treated cases were collected (25 per segment) for the analysis. Compared with MD contours, the mean DSC scores were higher than 0.7 for 74% and 80% of AI and AI+C, respectively. After corrections, 17.1% structures presented DSC score deviations higher than 0.1 and 10.4% dose-volume parameters significantly changed in AI-contoured structures. The time consumption assessment yielded mean person-hour reductions of 68%, 51%, and 71% for breast, prostate, and head and neck cancer, respectively.

Conclusions: In great extent, AI yielded clinically acceptable OARs and certain clinical target volumes in the explored anatomic segments. Sparse correction and assessment requirements place AI+C as a standard workflow. Minimal clinically relevant differences in OAR exposure were identified. A substantial amount of person-hours could be repurposed with this technology.

Introduction

Artificial intelligence (AI) is being increasingly related to current advances in medical care worldwide. Automation of processes and usage optimization of both material and human resources are 2 aspects to highlight among its advantages. The extent of its capabilities is a subject of research in different medical areas, such as oncology. In this regard, many opportunities arise in the field of radiation therapy (RT).[1]

Closely linked to technological development, workflows in RT depend in great manner on soft- and hardware availability. Quality assurance, treatment precision, and patient

Sources of support: This work had no specific funding.
Data will be shared upon reasonable request.
Corresponding author: Gustavo R. Sarria, MD; E-mail: gustavo.sarria@ukbonn.de

https://doi.org/10.1016/j.adro.2023.101394

safety are important items related to the latter, which have been historically performed by human practitioners.[2] As a part of these, a correct segmentation of organs at risk (OARs) and treatment target volumes defines a starting point for the subsequent involvement of other technological components.[3] This duty commonly lies on the professional team of an RT department and is particularly time-consuming, which could widely vary according to expertise and case difficulty. The process involves interpreting complex medical images and making critical decisions about where to target the radiation delivery. Additionally, manual contouring is subject to interobserver variability (ie, different clinicians may contour the same image differently), leading to inconsistencies and potential errors.[4,5] Nonetheless, AI has partly replaced human-performed tasks, enhancing the overall efficacy at all workflow levels.[1]

Modern AI solutions include deep learning algorithms, such as convolutional neural networks (CNNs), which are trained on large datasets of medical images. CNNs can identify patterns in medical images and learn to distinguish between different anatomic structures, such as tumors and healthy tissue.[6] Despite these recent advances, AI still faces certain challenges. There is a need for validation studies to ensure that the AI algorithms are accurate and reliable. These should focus on comparing AI-powered autocontouring with manual contouring by experts to ensure that the AI algorithm produces comparable or better results.[7] On the other hand, there are also concerns about a potential replacement of human expertise with AI-powered autocontouring. Although AI can enhance the accuracy and consistency of contouring, it cannot replace the knowledge and clinical judgment of radiation oncologists, physicists, or radiation therapists. Therefore, it is essential to ensure that AI is used as a tool to support and optimize clinical decision-making rather than as a replacement for human expertise.

Given the abovementioned facts, we deemed relevant to address these concerns and hypothesized that AI-powered autocontouring has the potential to produce human-like contours, while avoiding major deviations in dose distribution and saving significant person-hours.

Methods and Materials

Target and OAR delineation

An internal assessment at our department among 5 radiation oncologists was performed to determine which practitioner could achieve the most precise and swift delineations based on template cases defined according to the European Society for Radiotherapy and Oncology, the American Society for Radiation Oncology, or the Global Harmonization Group contouring guidelines.[8-11] This expert would assume thereafter all manual contours, which were to be considered the baseline standard for comparison.

Cases previously treated at our department encompassing prostate (PC), breast (BC), and head and neck cancer (HNC) were screened for DICOM datasets. These cases encompassed adjuvant treatment cases for BC, primary for PC, and both treatment modalities for HNC, and were selected randomly until reaching the target number, regardless of anatomic variants. Only cases with nonanatomic elements (bolus, prostheses other than oral, pacemakers, etc) were excluded. After selection, they were anonymized and DICOM sets were tripled for intergroup comparison. The medical doctor contours (MD) were defined on the first DICOM set, in a controlled environment assuring no distractions and dedicated time, to emulate an ideal scenario. To this end, using built-in autodelineation and interpolation tools was allowed. For comparison purposes, the total contouring time was measured and registered. Baseline structures included OARs in all cases and certain clinical target volumes (CTVs) in PC and BC segments, according to recommendations from international guidelines.[8-10] These structures are listed in Table 1. Once the MD-baseline contours were created, deep learning–based autocontours were generated on the remaining 2 DICOM set copies with Limbus Contour v.1.6.0 (Limbus AI Inc), yielding an AI-only and AI plus MD corrections (AI+C) sets. The former remained unmodified, while the same MD manually corrected the latter, in order to achieve structures better fitting the actual clinical practice, if necessary. Differences between all contouring groups were assessed by means of the Dice similarity coefficient (DSC). A 0.7 or lower value was deemed as a geometric major difference. Corrections from AI with a $\geq$0.1 difference were considered major as well. All delineations were created with Eclipse version 15.5 (Varian Medical Systems).

Planning and dose delivery comparison

The original dose distribution was retrieved and placed unmodified on the MD sets for baseline assessment. Furthermore, these were also rigidly copied on the AI and AI +C contours and raw differences in OAR exposure were compared with the baseline, even if the constraints were not met, as the purpose was to identify the magnitude of deviations in organ exposure. Several constraints corresponding to common practice were adopted for each segment (Table 1), according to different published trials or guidelines (Appendix E1).

Time-consumption assessment

Times from AI processing, MD, and AI+C were collected for descriptive purposes and person-hours were calculated for comparison between groups.

Table 1 Anatomic segments, structures, and dose constraints

Segment	Structures	Constraint
Breast	Lungs	D_{50} (b)
		D_{35} (b)
		D_{15} (b)
	Breasts*	D_{5} (c)
		$D_{0.1\ cc}$ (c)
	Heart	D_{mean}
		D_{10}
		D_{5}
	LAD	D_{mean}
		V_{30}
	Spinal cord	$D_{0.035\ cc}$
		$D_{2\ cc}$
	Thyroid	D_{mean}
	Esophagus	D_{mean}
Prostate	Bowel	V_{45}
	Bladder	V_{70}
		V_{45}
	Anal canal	D_{mean}
	Femurs	V_{50} (b)
		$D_{0.035\ cc}$ (b)
	Penile bulb	V_{50}
	Rectum	V_{72}
		V_{70}
		V_{45}
	Cauda equina	$D_{0.035\ cc}$
	Lymphatic pathways CTV	-
	Seminal vesicles	-
	Prostate	-
Head and neck	Plexi	$D_{3\ cc}$ (b)
		$D_{0.035\ cc}$ (b)
	Oral cavity	D_{mean}
	Esophagus	D_{mean}
		V_{35}
	Submandibular	D_{mean} (b)
	Thyroid	D_{mean}
	Larynx	D_{mean}
		$D_{0.035\ cc}$
	Lips	D_{mean}
		$D_{0.035\ cc}$
	Pharyngeal constrictor	D_{mean}
	Parotids	D_{mean} (b)
		$D_{7\ cc}$ (b)
	Spinal cord	$D_{0.035\ cc}$
	Mandible	-

(*continued on next page*)

Table 1 (*Continued*)

Abbreviations: b = bilateral; c = contralateral; CTV = clinical target volume; D = dose delivered at relative (%) volume (or absolute if indicated); LAD = left anterior descending artery; V = volume receiving certain dose (Gy).

* Breast contours were used as both organs at risk and CTV.

Structure and constraint list per anatomic segment. The constraint references are listed in the Appendix E1.

Statistical analysis

Mean DSC values plus standard deviation (SD) were registered for each studied anatomic segment in AI and AI+C and matched to those MD-generated for descriptive purposes. Differences in the resulting mean OAR-doses for each constraint were compared with the unpaired *t* test for independent variables. A $P < .05$ value determined a significant difference. The analysis was performed with MATLAB R2020b (The MathWorks Inc).

Ethics statement

Institutional review board release was obtained before initiation, due to the retrospective nature of this study. All datasets were anonymized before inclusion. This investigation was conducted according to the principles of the Declaration of Helsinki.

Results

Seventy-five cases were collected (25 for each anatomic segment) for the analysis. For BC structures, the overall mean DSCs were 0.82 (SD ± 0.19) for AI+C and 0.79 (SD ± 0.18) for AI. All structures showed DSC scores higher than 0.7, excepting for the left anterior descending artery (LAD) with 0.37 (SD ± 0.11) and 0.36 (SD ± 0.11) for AI +C and AI, respectively (Fig. 1). For PC, the overall mean DSC were 0.84 (SD ± 0.09) and 0.75 (SD ± 0.18). Major differences were detected for anal canal (0.68, SD ± 0.14, and 0.27, SD ± 0.1) and penile bulb (0.69, SD ± 0.13, and 0.60, SD ± 0.17). The cauda equina and seminal vesicles showed major differences only in the AI group (0.64, SD ± 0.09, and 0.62, SD ± 0.15; Fig. 2). Regarding HNC, the overall mean DSC were 0.72 (SD ± 0.15) and 0.70 (SD ± 0.16). Major differences were found in brachial plexi right (0.42, SD ± 0.07, and 0.37, SD ± 0.11) and left (0.43, SD

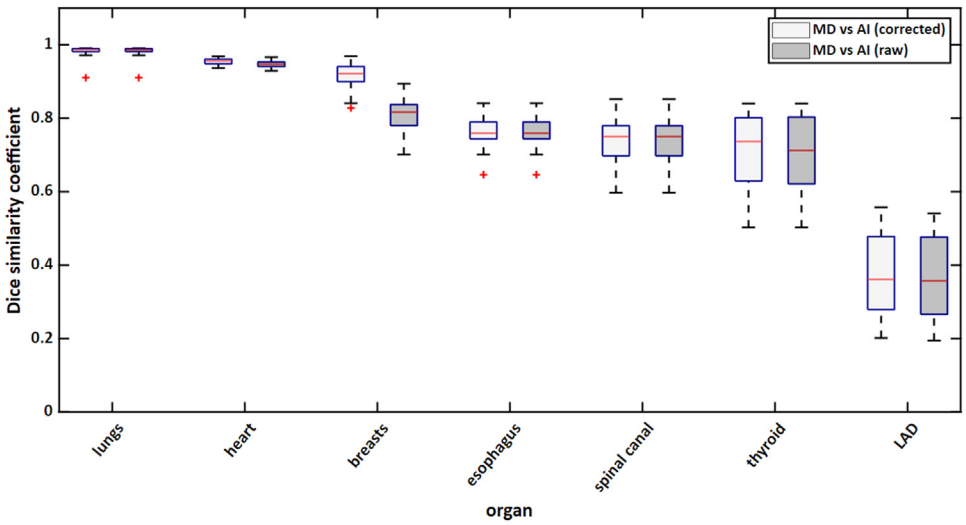

Figure 1 Dice similarity coefficient scores in breast cancer. MD baseline contours were compared with AI alone and AI plus MD corrections. *Abbreviations*: AI = artificial intelligence; MD = experienced radiation oncologist.

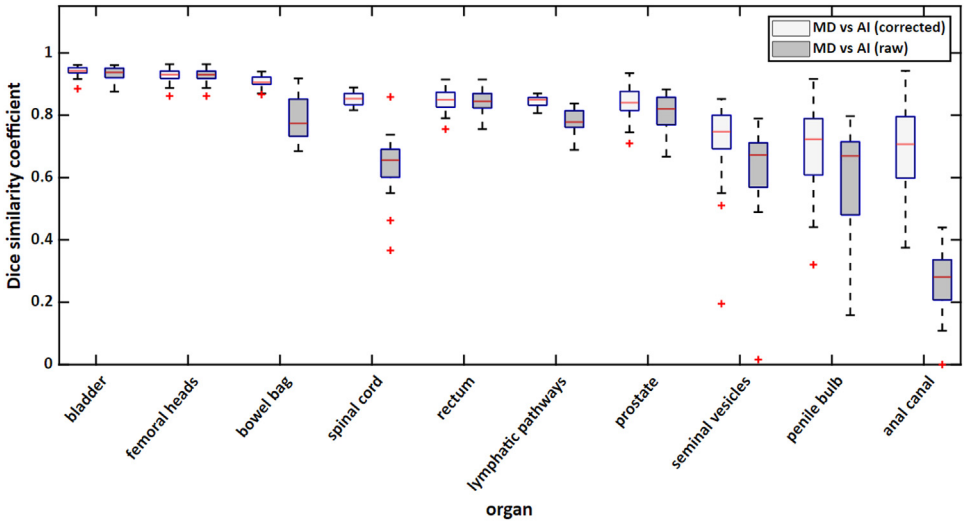

Figure 2 Dice similarity coefficient scores in prostate cancer. MD baseline contours were compared with AI alone and AI plus MD corrections. *Abbreviations*: AI = artificial intelligence; MD = experienced radiation oncologist.

± 0.06, and 0.41, SD ± 0.09), lips (0.57, SD ± 0.09, and 0.45, SD ± 0.09), and pharyngeal constrictor (0.62, SD ± 0.06, and 0.61, SD ±0.07; Fig. 3). Major corrections were required for anal canal, bowel bag, cauda equina, breasts, and lips, which represent 17.1% of all structures.

In terms of OAR exposure, significant dose distribution differences were identified for BC AI in contralateral breast D0.1 cc (P = .041). For PC AI+C and AI in mean bowel V45 (P = .0007 and P < .0001), anal canal D_{mean} (P = .0051 and P = .0013) and cauda equina $D_{0.035\ cc}$ (AI only, P < .0001). In HNC, only mean lips $D_{0.035\ cc}$ (P = .02 and P = .014) were significantly different. These translate into 10.4% of all structures presenting with dose-absorption deviations. The absolute dose differences per each constraint and box plots can be observed in Appendix E1 and E2, respectively.

The time consumption analysis yielded mean MD segmentation times (seconds) of 475 (SD ± 70), 752 (SD ± 80), and 644 (SD ± 45), for BC, PC, and HNC, respectively. Similarly, the mean correction times of AI-generated contours were 153 (SD ± 40), 367 (SD ± 91), and 187 (SD ± 15), and the AI processing times were 354 (SD ± 25), 201 (SD ± 18), and 140 (SD ± 48). Proportionally, these represent mean person-hour reductions of 68%, 51% and 71% for each of the explored anatomic segments, respectively (Fig. 4).

Discussion

CNNs have shown great promise in the field of automation in medicine. They have proven quite robust at analyzing

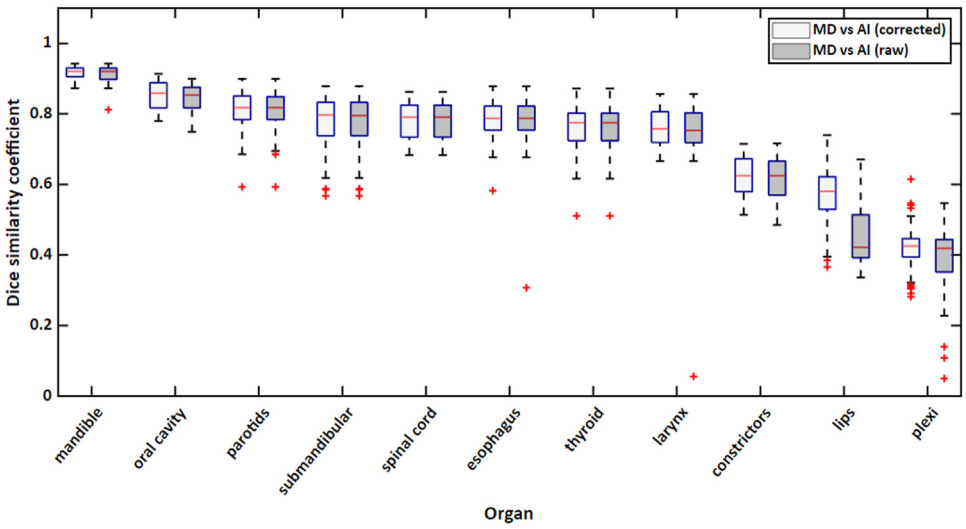

Figure 3 Dice similarity coefficient scores in head and neck cancer. MD baseline contours were compared with AI alone and AI plus MD corrections. *Abbreviations*: AI = artificial intelligence; MD = experienced radiation oncologist.

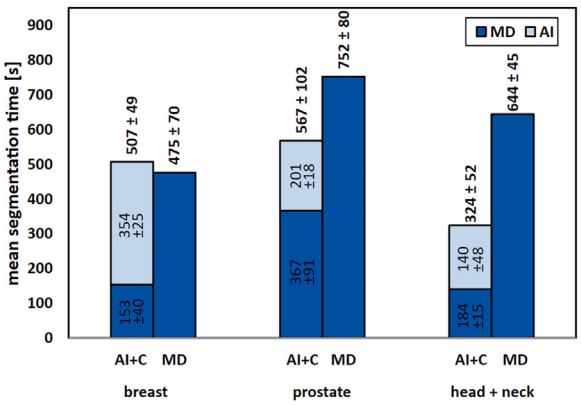

Figure 4 Mean (± standard deviation) segmentation times in seconds for MD, AI, and AI+C contours in all 3 investigated anatomic segments. *Abbreviations*: AI = artificial intelligence; C = corrections; MD = experienced radiation oncologist.

images and identifying patterns, making them optimal for the task of automated target volume delineation.[12] Advances in the past few years have led to a migration from atlas-based autocontouring to deep learning–based autocontours, which might adjust better to variabilities in daily practice and outperform the former, even suggesting corrections to clinically validated models.[13] In our study, we challenged the capacities of the Limbus Contour software against a highly experienced practitioner and evaluated the reach this could have on clinical routine. The 3 evaluated segments were selected as representative indications for patients treated at any RT department, according to their frequency.

The DSC is a useful statistical tool for contour comparison, by taking 2 different segmentations and measuring their overlap ratio. These values range from 0 to 1, where 1 represents full accordance between both structures. Although occasionally this might be subject to bias, depending on factors such as structure volume or number of voxels, it still represents one of the strongest available methods for performing these measurements.[14] In our analysis, the DSC showed overall acceptable correlations between the MD, AI, and AI+C delineations. Most significant differences were identified in small or long organs, albeit without major clinical meaning. In addition, it should be considered that all corrections (AI+C) were done until achieving clinically acceptable standards, as per the MD's judgment. The low correction rates of the AI-generated contours (17.1% of organs had DSC change >0.1) demonstrates a quite impressive accuracy, requiring minor to no correction at all. Moreover, the dosimetric analysis showed that, even after modifications, these almost never led to significant changes in the clinically relevant DVH parameters. In BC cases, the most remarkable differences were identified in LAD contours. This is relatively common, as pointed out in different studies assessing this structure, and highly dependent on imaging quality and anatomic configuration.[15-17] Nevertheless, the MD deemed these differences of little clinical relevance, as most of them were related to discrepancies at width, upper end, or lower end. Further minor differences were identified in esophagus, thyroid, and spinal cord, also linked to the previously mentioned. Noteworthy, the AI +C breast contours were converted into CTVs with great similarity to the standard MD (DSC 0.92). Our results are consistent with other previous studies addressing this same topic.[6,18,19] We must mention that no lymphatic pathway contours were considered in this segment, due to technical reasons. The software version available at the moment of the study counted with the American Society for Radiation Oncology templates, while our contours are routinely performed according to the European Society for Radiotherapy and Oncology recommendations. In

regard to PC, the most notable differences were observed in the anal canal, lymphatic pathways, penile bulb, and seminal vesicles. The first showed substantial deviations between MD and AI contours. A reason for this might be related to the template used by the software, which evidently differs from the Global Harmonization Group consensus guidelines,[8] creating smaller structures not reaching the outer sphincter. A similar situation was observed for the seminal vesicles, perhaps due to their variable anatomy. Likely, differences in total volume and cranial or caudal length led to lower DSC scores in the remaining OARs, mostly not clinically relevant (eg, cauda equina). Worth highlighting, the lymphatic pathways reached a satisfactory configuration with minor editing (DSC 0.85). These outcomes resemble those of Cha et al, although we included more OARs in our analysis, providing additional robustness.[20] For the HNC segment, no CTV analysis was considered, due to a higher inhomogeneity in volume prescription. Several differences in DSC scores were found in more organs than in the other explored segments, possibly due to the smaller nature of these structures. Yet, most of them were slightly below a 0.8 DSC score and above our 0.7 lower-limit score. Certain specific cases, such as the brachial plexus, presented with higher deviations. For practical reasons, our plexus contour includes the bony structures between foramina. This is to ensure a homogeneous dose distribution along the plexus area and minimize possible hotspots nearby. In this sense, the AI algorithm provides a more anatomic delineation, respecting the vertebral bodies, and in accordance with international guidelines.[8] Moreover, the plexus portions between both scalene muscles, which are the ones potentially closer to the treated volumes, were rather similar in all 3 groups. Clinically judged and despite the DSC outcomes, these differences were considered of minor relevance. Regarding metrics, we deem relevant to stress some points. The arbitrary cut-off point was adopted according to previous publications,[21-23] although no strong consensus has been reached to date. A reason for this might lie on a difference between metrics and clinical applicability, as the former does not necessarily guarantee the latter.[24,25] Consequently, these values must be taken as reference but should not directly affect clinical decision-making. Further metrics were not considered for the analysis (eg, Hausdorff distance) because single linear metrics do not provide a proper oversight of volume overlap. Instead, we deemed relevant to measure the differences in dose distribution among the 3 investigational groups, to achieve a better appreciation of a potential clinical effect. By placing the actual irradiated plan on the MD contours and assuming this as the baseline standard, we found that only 5 of 48 dose-volume parameters varied according to the segmentation method. These changes were located mostly in organs with large size or challenging anatomy. Furthermore, as we mainly report statistical differences, these could mean either lower or higher exposure rates (details in Appendix E1). Regardless of this, most differences were minimal and could be deemed clinically not relevant, supporting our findings based on the DSC coefficient.

The time-sparing analysis yielded interesting differences along all 3 groups. For instance, the AI output was evidently faster in the PC and HNC contours. Although it was also faster in the BC sets, this advantage appeared to be minor. This could be related to the number of structures per segment and their complexity. In addition, interpolation and integrated autodelineation tools perform better usually in larger organs, such as lungs or breasts. As expected, the AI+C times were considerably shorter than the MD's, resulting in a significant reduction of person-hours. It must be remarked that the MD contours were defined in a controlled, favorable environment, which is practically impossible to reproduce in daily practice. Moreover, our expert MD has vast experience in the field and has practiced for several years as contouring trainer. We established this ideal scenario in order to challenge the AI as much as possible. In a real-life setting, the benefit could be even more noticeable if other distracting factors are to be accounted for.

However, it is important to note that AI algorithms are not perfect, and there are some limitations to their accuracy. For example, they may struggle with low-quality images, artifacts, anatomic variants, or unusual positioning settings. Additionally, there is a considerable risk of overreliance on autocontouring algorithms, which can potentially lead to errors or skipping thorough assessments. To mitigate these risks, it is mandatory to incorporate manual review of autocontours into the treatment planning process. Therefore, we established AI+C as our standard workflow. Another important consideration is the risk of de-skilling among radiation oncologists and physicists, as manual contouring may be less frequently used with the increased use of AI solutions. To address these issues, it is compulsory to ensure training in the use of AI and acquire the skills necessary to review and adjust autocontours as needed. This can help ensure that the AI is used as a tool to improve clinical workflow, rather than a replacement for human expertise.

Our study carries certain shortcomings. Its single-center and single-practitioner character could yield variable outcomes, when trying to reproduce this methodology. The software and hardware version available at different centers could also contribute to the latter. Besides the abovementioned, our outcomes might not be comparable to those of MRI-guided RT, as these could result in even more accurate segmentations.[26] Notwithstanding, AI-based contouring has the potential to save person-hours, and to improve contouring consistency and workflows in comparison to human-performed manual contouring. As this technology continues to evolve, it has the potential to revolutionize the field of RT.

Conclusion

In great extent, AI yielded acceptable OARs and certain CTVs in the explored anatomic segments. Sparse correction and assessment requirements place AI+C as a standard workflow. Minimal clinically relevant differences in OAR-exposure were identified. A substantial number of person-hours could be repurposed with this technology.

Disclosures

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

Supplementary materials

Supplementary material associated with this article can be found in the online version at doi:10.1016/j.adro.2023.101394.

References

1. 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.
2. Munbodh R, Roth TM, Leonard KL, et al. Real-time analysis and display of quantitative measures to track and improve clinical workflow. *J Appl Clin Med Phys*. 2022;23:e13610.
3. Hernandez V, Hansen CR, Widesott L, et al. What is plan quality in radiotherapy? The importance of evaluating dose metrics, complexity, and robustness of treatment plans. *Radiother Oncol*. 2020;153:26-33.
4. Patrick H M, Souhami L, Kildea J. Reduction of inter-observer contouring variability in daily clinical practice through a retrospective, evidence-based intervention. *Acta Oncol*. 2021;60:229-236.
5. van der Veen J, Gulyban A, Nuyts S. Interobserver variability in delineation of target volumes in head and neck cancer. *Radiother Oncol*. 2019;137:9-15.
6. 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;144:152-158.
7. Savjani RR, Lauria M, Bose S, Deng J, Yuan Y, Andrearczyk V. Automated tumor segmentation in radiotherapy. *Semin Radiat Oncol*. 2022;32:319-329.
8. Mir R, Kelly SM, Xiao Y, et al. Organ at risk delineation for radiation therapy clinical trials: Global Harmonization Group consensus guidelines. *Radiother Oncol*. 2020;150:30-39.
9. Offersen BV, Boersma LJ, Kirkove C, et al. ESTRO consensus guideline on target volume delineation for elective radiation therapy of early stage breast cancer, version 1.1. *Radiother Oncol*. 2016;118:205-208.
10. Hall W A, Paulson E, Davis BJ, et al. NRG Oncology updated International Consensus Atlas on pelvic lymph node volumes for intact and postoperative prostate cancer. *Int J Radiat Oncol Biol Phys*. 2021;109:174-185.
11. Brouwer C L, Steenbakkers RJ, Bourhis J, et al. CT-based delineation of organs at risk in the head and neck region: DAHANCA, EORTC, GORTEC, HKNPCSG, NCIC CTG, NCRI, NRG Oncology and TROG consensus guidelines. *Radiother Oncol*. 2015;117:83-90.
12. Poortmans PMP, Takanen S, Marta GN, Meattini I, Kaidar-Person O. Winter is over: The use of artificial intelligence to individualise radiation therapy for breast cancer. *Breast*. 2020;49:194-200.
13. Rhee DJ, Cardenas CE, Elhalawani H, et al. Automatic detection of contouring errors using convolutional neural networks. *Med Phys*. 2019;46:5086-5097.
14. Zou KH, Warfield SK, Bharatha A, et al. Statistical validation of image segmentation quality based on a spatial overlap index. *Acad Radiol*. 2004;11:178-189.
15. Chin V, Finnegan RN, Chlap P, et al. Validation of a fully automated hybrid deep learning cardiac substructure segmentation tool for contouring and dose evaluation in lung cancer radiotherapy. *Clin Oncol (R Coll Radiol)*. 2023;35:370-381.
16. Harms J, Lei Y, Tian S, et al. Automatic delineation of cardiac substructures using a region-based fully convolutional network. *Med Phys*. 2021;48:2867-2876.
17. Spoor DS, Sijtsema NM, van den Bogaard VAB, et al. Validation of separate multi-atlases for auto segmentation of cardiac substructures in CT-scans acquired in deep inspiration breath hold and free breathing. *Radiother Oncol*. 2021;163:46-54.
18. Choi MS, Choi BS, Chung SY, et al. Clinical evaluation of atlas- and deep learning-based automatic segmentation of multiple organs and clinical target volumes for breast cancer. *Radiother Oncol*. 2020;153:139-145.
19. Buelens P, Willems S, Vandewinckele L, et al. Clinical evaluation of a deep learning model for segmentation of target volumes in breast cancer radiotherapy. *Radiother Oncol*. 2022;171:84-90.
20. Cha E, Elguindi S, Onochie I, et al. Clinical implementation of deep learning contour autosegmentation for prostate radiotherapy. *Radiother Oncol*. 2021;159:1-7.
21. Aoyama T, Shimizu H, Kitagawa T, et al. Comparison of atlas-based auto-segmentation accuracy for radiotherapy in prostate cancer. *Phys Imaging Radiat Oncol*. 2021;19:126-130.
22. Pera O, Martinez A, Mohler C, et al. Clinical validation of Siemens' syngo.via automatic contouring system. *Adv Radiat Oncol*. 2023;8:101177.
23. Dolz J, Kirisli H A, Fechter T, et al. Interactive contour delineation of organs at risk in radiotherapy: Clinical evaluation on NSCLC patients. *Med Phys*. 2016;43:2569.
24. Sherer MV, Lin D, Elguindi S, et al. Metrics to evaluate the performance of auto-segmentation for radiation treatment planning: A critical review. *Radiother Oncol*. 2021;160:185-191.
25. Rhee DJ, Akinfenwa CPA, Rigaud B, et al. Automatic contouring QA method using a deep learning-based autocontouring system. *J Appl Clin Med Phys*. 2022;23:e13647.
26. Sritharan K, Dunlop A, Mohajer J, et al. Dosimetric comparison of automatically propagated prostate contours with manually drawn contours in MRI-guided radiotherapy: A step towards a contouring free workflow? *Clin Transl Radiat Oncol*. 2022;37:25-32.