BJR|Artificial Intelligence, 2024, **1(1)**, ubae015
https://doi.org/10.1093/bjrai/ubae015
Advance access publication: 25 October 2024
Review

Clinical adoption of deep learning target auto-segmentation for radiation therapy: challenges, clinical risks, and mitigation strategies

Alessia De Biase, MSc[1,2,*], Nanna Maria Sijtsema, PhD[1], Tomas Janssen, PhD[3], Coen Hurkmans, PhD[4,5], Charlotte Brouwer, PhD[1], Peter van Ooijen, PhD[1,2]

[1]Department of Radiation Oncology, University Medical Centre Groningen (UMCG), Groningen, 9700 RB, The Netherlands
[2]Data Science Centre in Health (DASH), University Medical Centre Groningen (UMCG), Groningen, 9700 RB, The Netherlands
[3]Department of Radiation Oncology, The Netherlands Cancer Institute, Amsterdam, 1006 BE, The Netherlands
[4]Department of Radiation Oncology, Catharina Hospital Eindhoven, Eindhoven, 5602 ZA, The Netherlands
[5]Department of Electrical Engineering, Technical University Eindhoven, Eindhoven, 5600 MB, The Netherlands

*Corresponding author: Alessia De Biase, MSc, Department of Radiation Oncology, University Medical Centre Groningen (UMCG), Hanzeplein 1, 9713 GZ Groningen, The Netherlands (a.de.biase@umcg.nl)

Abstract

Radiation therapy is a localized cancer treatment that relies on precise delineation of the target to be treated and healthy tissues to guarantee optimal treatment effect. This step, known as contouring or segmentation, involves identifying both target volumes and organs at risk on imaging modalities like CT, PET, and MRI to guide radiation delivery. Manual segmentation, however, is time-consuming and highly subjective, despite the presence of contouring guidelines. In recent years, automated segmentation methods, particularly deep learning models, have shown promise in addressing this task. However, challenges persist in their clinical use, including the need for robust quality assurance (QA) processes and addressing clinical risks associated with the use of the models. This review examines the challenges and considerations of the clinical adoption of deep learning target auto-segmentation in radiotherapy, focused on the target volume. We discuss potential clinical risks (eg, over- and under-segmentation, automation bias, and appropriate trust), mitigation strategies (eg, human oversight, uncertainty quantification, and education of clinical professionals), and we highlight the importance of expanding QA to include geometric, dose-volume, and outcome-based performance monitoring. While deep learning target auto-segmentation offers significant potential benefits, careful attention to clinical risks and rigorous QA measures are essential for its successful integration in clinical practice.

Keywords: target auto-segmentation; deep learning; risk analysis; monitoring; quality assurance; clinical adoption; radiation oncology.

Setting the stage: segmentation in radiation therapy

Radiation therapy is a localized cancer treatment that employs high-energy radiation directed at a specific volume of the body to kill cancer cells. Achieving optimal treatment effect requires the exact identification of the tumour location and extension. Often, multiple 3D medical imaging modalities such as CT, PET, and MRI are combined in this process, since tumours are not always visible on single imaging modalities. These images, in combination with additional diagnostic or clinical information (tumour staging, physical examination, etc) are used by the radiation oncologist to identify the radiation target volume. The delineation of this volume is subsequently used to optimize the radiation beam fluence, aiming at delivering sufficient dose in the tumour and preserving the surrounding healthy tissues. Delineation of the target, also known as contouring or segmentation, is indispensable for achieving the therapeutic effect of the treatment (Figure 1).

In the segmentation task, two subgroups of structures can be distinguished: organs at risk (OARs), concerning all the healthy tissues/organs located in the radiation field, and the target volumes, which should receive sufficient dose to achieve local tumour control. The delineated OARs are used during treatment plan optimization to minimize radiation damage to these critical volumes, which could result in toxicity and complications. Identifying the boundaries of OARs is generally considered a less complex task compared to delineating the target volume. One significant reason is that OARs are typically anatomically defined structures or organs and therefore their location, shape, volume, and appearance generally exhibit consistent traits across patients, in contrast to tumours, where these factors can vary significantly[1].

In radiotherapy, the word 'target' may be used to indicate different structures, as illustrated in the example in Figure 2. The definitions provided below are in accordance with the International Commission on Radiation Units and Measurements (ICRU) Report 50[2]. Gross tumour volumes (GTV) are the visible tumours in imaging and in clinical examinations, representing the demonstrable extent of the malignant growth. Clinical target volumes (CTV) are targets which account for the presence of subclinical microscopic tumour spread beyond the visible GTV. The planning target volume (PTV) is derived by adding a margin to the CTV to

Received: 19 April 2024; **Revised:** 11 September 2024; **Accepted:** 22 October 2024
© The Author(s) 2024. Published by Oxford University Press on behalf of the British Institute of Radiology.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

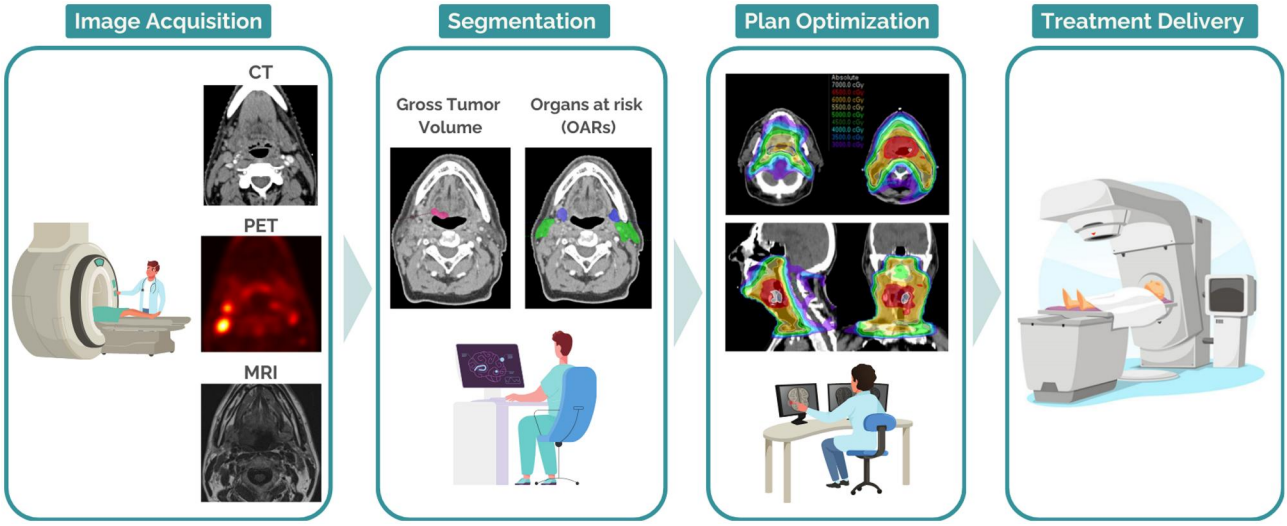


Figure 1. Illustration of the different steps of the radiation therapy treatment journey, showing an example of a head and neck cancer patient. In the image acquisition step, CT, PET, and MRI scans are acquired in treatment position. Subsequently, segmentations are performed for the tumour that is visible on imaging (gross tumour volume) and in clinical examinations, and the adjacent organs (organs at risk). In the third step, the radiation plan is optimized based on the patient's CT scan and on the segmentations performed in the previous step. Finally, as the last step, the treatment is delivered.

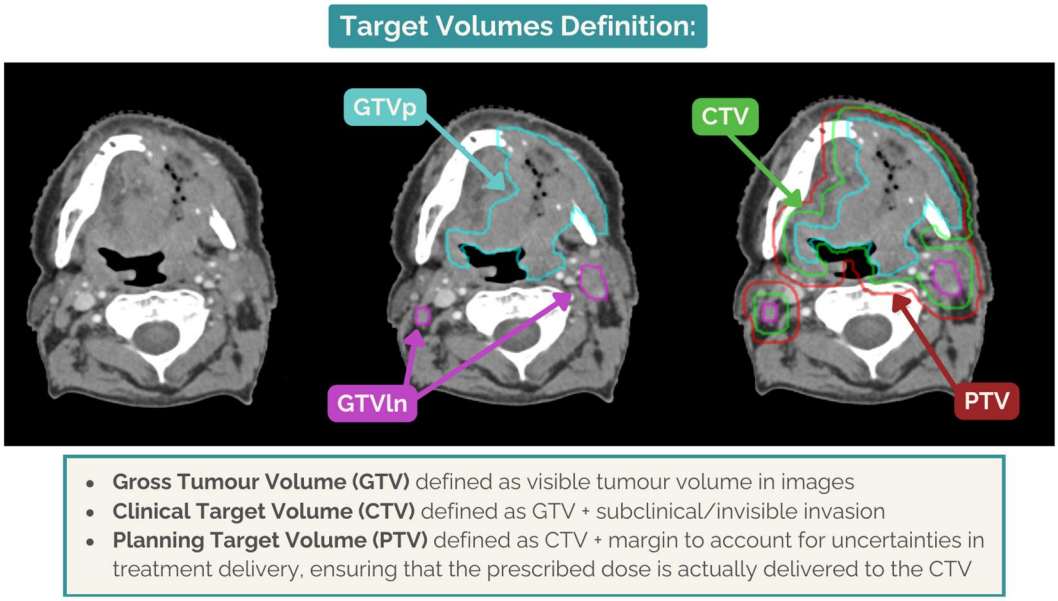


Figure 2. Illustration of an example of the target volumes contours on the axial slice of a CT scan used for radiotherapy planning of a head and neck cancer patient. The first image on the left shows the Planning CT only. In the middle image, the gross tumour volume of the head and neck primary tumour (GTVp) and the gross tumour volume of the pathologic lymph nodes (GTVln) are displayed. The image on the right side illustrates how the clinical target volume and planning target volume align with the GTV segmentations, in this case, offering a clear visual depiction of their interrelation with the preceding volumes. Abbreviation: GTV = gross tumour volume.

compensate for uncertainties in delineation, treatment delivery (eg, beam alignment, patient positioning) and patient anatomy (eg, organ motion and organ deformation). To precisely delineate the GTVs, a variety of imaging techniques is often needed. Conversely, defining CTVs involves not only imaging but also clinical information and anatomical considerations (eg, adjustments may be needed to exclude bone structures if the margin extends beyond them). The CTV is either defined based on GTV plus margin (referred to as the tumour CTV), anatomical regions (eg, prostate gland, lymph nodes levels), or a combination of both. CTV delineation guidelines vary based on tumour sites and staging[3-7].

Over the last two decades, CT imaging has stood as the primary modality in radiotherapy treatment, providing essential information for dose calculation, since CT Hounsfield units are easily related to radiation beam attenuation. Nonetheless, CT does not always offer sufficient soft-tissue or functional contrast crucial for visualizing tumour boundaries. To overcome this limitation, pre-treatment PET and/or MRI images are often co-registered with the planning CT or utilized independently in more advanced treatment techniques (eg, MR-only radiotherapy). Alongside imaging, clinical examinations of the patient provide valuable information for exact tumour localization. The role of the radiation oncologist is to finally

integrate all this complementary data to manually segment the target on the available modalities.

Interpretation of imaging by radiation oncologists is highly subjective[8] and dependent on clinical judgement, leading to inter- and intra-observer variability in target segmentation[9-11]. These differences in how the same structure is segmented by different or even the same observer, under similar circumstances, lead to inconsistencies in the dose delivery to the actual tumour volume[12]. Since the introduction of intensity-modulated radiotherapy, which results in steeper dose gradients, the need for more standardized target volume definitions increased[13,14]. As a consequence, an increase of tumour site-specific consensus delineation guidelines was observed[15]. However, despite their implementation, significant inter-observer variability persists[16], which may depend on the difficulty in their interpretation[17].

In summary, target segmentation is a time-consuming and subjective task, hence prone to error and inter-observer variability, highlighting the need for solutions that can enhance efficiency, accuracy, and consistency. Moreover, with the shortage of staff and the introduction of innovative techniques like online adaptive radiotherapy, rapid target segmentation becomes crucial in ensuring timely and accurate treatment delivery. In the era of automation and artificial intelligence (AI), deep learning (DL) models, trained on large amounts of annotated data, showed promises in learning complex patterns in different data sources to automate the segmentation process[18]. However, challenges remain in the widespread adoption of automated target segmentation in clinical settings[19].

The NRG Oncology recently published an assessment of AI-based organ auto-segmentation models for radiation therapy, evaluating their clinical utilization, accuracy, and challenges with a focus on their integration, potential benefits, and limitations[20]. However, as previously mentioned, significant differences exist between organ and target segmentation tasks. Building on this, our review aims to offer a comprehensive understanding of the unique challenges and considerations associated with the clinical adoption of target auto-segmentation technologies. Additionally, we draw attention to the crucial aspects of clinical risk analysis and monitoring to ensure the safe and effective utilization of these automated systems in radiotherapy treatment preparation. More specifically, the manuscript is organized as follows:

- The second section reviews the state-of-the-art DL models for target auto-segmentation;
- The third section examines the current landscape, challenges, and steps needed for clinical implementation and adoption of AI-based target auto-segmentation;
- The fourth section addresses the sources of failure in DL models, their clinical implications, and strategies for risk mitigation;
- The fifth section highlights the critical role of QA in radiotherapy, with a focus on DL target auto-segmentation models, and discusses challenges and strategies for optimization;
- The sixth section looks at future directions for DL target auto-segmentation in radiotherapy;
- The seventh section summarizes the key findings and discussion points of the review.

Deep learning target auto-segmentation

Auto-segmentation techniques have a long history of usage in radiotherapy. Rong et al[20] provided a complete overview, highlighting the advantages and disadvantages of those techniques in the clinical setting. When compared to the previously adopted methods, DL-based approaches stand out for increased robustness, and higher efficiency and performance. In fact, a significant drawback of prior methodologies lies in their heavy reliance on a single image characteristic including shape, appearance, and intensity (eg, thresholding or edge-based approaches) of structures. Moreover, commonly employed atlas-based approaches assume sufficient similarities between the atlas cases (used as a reference to guide the labelling of new data) and the new cases, enabling deformable transformations to achieve accurate new segmentations. While this is often the case for OARs, it could pose considerable challenges for target delineation, where significant variability in tumour shape, size, and location can occur.

Savjani et al[19] and Matoska et al[21] provided an overview of site-specific advances in GTV and CTV auto-segmentation in radiotherapy, respectively, using DL techniques. Building upon their contributions, we provide an updated version in Table 1, including the latest and best-performing target auto-segmentation studies published from 2022 to March 2024. A literature search was conducted in PubMed and in Google Scholar using the following keywords: ('tumor' OR 'tumour' OR 'cancer') AND ('segmentation' OR 'automatic segmentation' OR 'automated segmentation' OR 'auto-segmentation' OR 'contouring' OR 'delineation') AND ('deep learning'). Performance was assessed using segmentation accuracy metrics, with the dice similarity coefficient (DSC) being the most commonly employed. While DSC is frequently used, it is important to note that it has limitations, especially due to its heavy reliance on volume. Other metrics were also considered when available. Multi-centre studies were preferred; however, in the absence of such studies, we included single-centre ones. Studies with the largest patient cohorts were selected. In some tumour sites (glioblastoma, oropharynx, head and neck pathologic lymph nodes, kidney, and liver), this involved including the latest and best-performing results from DL programming challenges. The previously published review was used as reference to identify studies that achieved the highest or similar performance in comparison.

Compared to the findings of the previously cited review paper[19], the improvement in performance metrics is evident. For the tumour sites appearing in both reviews, the DSC score is of 0.90 vs. 0.89 for glioblastoma, 0.90 vs. 0.75 for brain metastases, 0.80 vs. 0.76 for oropharynx, 0.82 vs. 0.82 for lung, 0.87 vs. 0.79 for oesophagus, 0.84 vs. 0.85 for kidney, 0.90 vs. 0.73 for pancreas, 0.86 vs. 0.84 for liver, 0.88 vs. 0.94 for prostate, 0.87 vs. 0.86 for cervix. Despite not seeming significantly higher for every tumour site, this improvement is particularly noteworthy given the larger and more heterogeneous datasets used in the cited studies (glioblastoma, brain metastases, and lung are above 1000 cases each), indicating promising advancements in the field. GTV segmentation predominates over tumour CTV segmentation, largely due to (1) the fact that GTV definition corresponds directly to visible imaging data, simplifying model training; (2) the higher variety in tumour CTV delineations, which relies on guidelines, clinical factors and margins that are often

Table 1. Overview of best performing DL target auto-contouring methods by tumour site (January 2022-March 2024).

Tumour site	Target type	DL architecture	Additional Info	Dataset (*n*: size)	Imaging modalities	Performance
Brain						
Glioblastoma[23]	GTV	nn U-Net, Swin UNETR	Data Augmentation via registration and GANs. Ensemble of CNN and transformer-based networks.	BraTS 2023 ($n = 2040$)	MRI sequences (T1, T1Gd, T2 and FLAIR)	0.90 DSC 14.94 mm 95th HD
Brain metastases[24]	GTV	GHR-CNN: 2D gated high-resolution CNN	2D architecture. Gated channel attention mechanism	Multi-centre dataset 2012–2022 ($n = 1592$)	MRI (T1c)	0.90 DSC
Head and neck						
Oropharynx[25]	GTV	Auto3DSeg	Model ensemble. Semantic segmentation network with deep supervision.	HECKTOR 2022 ($n = 883$)	CT, FDG PET	0.80 DSC
Lymph nodes[25]	GTV	Auto3DSeg	Model ensemble. Semantic segmentation network with deep supervision.	HECKTOR 2022 ($n = 883$)	CT, FDG PET	0.77 DSC
Nasopharynx[26]	GTV	nn U-Net	Augmentation-invariant framework using intensity transformation strategies.	Multi-centre dataset ($n = 1057$)	MRI (T1)	0.88 DSC 4.99 mm 95th HD
Thorax						
Breast[27]	CTVp	3D U-Net	Multiple U-Net submodels are trained on one or a subgroup of ROIs.	Multi-centre dataset ($n = 190/200$)	CT	0.92-0.94 DSC 8.01-9.38 mm 95th HD
Lung[28]	GTV	2D U-Net with Exponential Linear Unit activations	Multi steps approach: (1) CT data harmonization, (2) lung isolation, (3) tumour detection, (4) tumour segmentation in the detected area	Multi-centre dataset ($n = 1328$)	CT	0.82 DSC 9.43 mm 95th HD
Oesophagus[29]	GTV	3D nn U-Net	Training performed on patches. Sliding window during inference.	Multi-centre dataset ($n = 580$)	CT with contrast	0.87 DSC 2.88-4.33 mm 95th HD
Abdomen						
Kidney[30]	GTV	Auto3DSeg	Bounding box detection for kidneys. Model ensemble from SegResNet and DiNTS algorithms.	KiTS 2023 ($n = 599$)	CT	0.84 DSC
Pancreas[31]	GTV	TD-Net: Trans-Deformer	2D Unet for coarse segmentation and Vision Transformer for fine segmentation.	Multi-centre dataset ($n = 363$)	CT with contrast	0.90-0.91 DSC
Liver[32]	GTV	Three paths encoder-decoder	Multi-scale selective feature fusion, multi-channel feature fusion, edge-inspiring, and edge-guiding modules.	LiTs2017, 3Dircadb ($n = 151$)	CT	0.86 DSC 11.21 mm 95th HD
Pelvis						
Prostate[33]	Prostate gland	2D U-Net	Deep supervision stage in the decoder blocks. Cyclical learning rate.	Multi-centre dataset ($n = 243$)	MRI (T2W)	0.88 DSC
Rectum[34]	GTV CTV	DpnUnet	Two models were trained: one on MRI, GTV as ground truth; one on CT, CTV as ground truth. Model: a U-shape encoder-decoder, locally integrated with dual-path-network modules.	Single-centre dataset ($n = 141$)	MRI CT	0.87 DSC 4.07 mm 95th HD 0.85 DSC 7.75 mm 95th HD
Cervix[35]	CTV	VB-Net	A bottleneck structure replaced the convolution, normalization, and activation layers in V-Net to increase model efficiency.	Single-centre dataset ($n = 535$)	CT	0.86-0.88 DSC

Abbreviations: DL = deep learning; GTV = gross tumour volume; DSC = dice similarity coefficient; 95th HD = Hausdorff distance 95th percentile; CTV = clinical target volume; UNETR = UNet Transformer; GHR-CNN = gated high-resolution convolutional neural network; ROIs = regions of interest; FLAIR = Fluid-attenuated inversion recovery; FDG = Fluorodeoxyglucose.

not directly discernible in imaging, posing additional challenges to DL models; (3) the greater accessibility of GTV data in openly available datasets. While architectures based on convolutional neural networks (CNN) maintain their dominance over transformer-based ones in most scenarios, significant emphasis is placed on refining pre-processing techniques, employing data augmentation strategies, and leveraging model ensembling, whether with identical or diverse architectures. CT images continue to be the preferred modality across various tumour sites, closely followed by MRI. Unlike CT, MRI protocols tend to exhibit considerable variation across different centres[22], making the generalization of MRI-based models more challenging. Notably, in this overview, head and neck cancer stands as the sole tumour site reliant on multimodality PET/CT imaging for accurate GTV auto-segmentation, likely due to the availability of easily accessible open multimodality datasets for this specific tumour site.

The images used as input for the models in Table 1 are all high-quality diagnostic scans that facilitate the tumour contouring process. None of the studies included imaging from online adaptive radiotherapy, such as cone-beam CT, likely due to the lower quality of these modalities. Given the limited number of studies focused on these specific imaging types and the fact that only a few tumour sites are treated with online adaptive radiotherapy, we considered them outside the scope of this review. Future research could address the unique challenges associated with using DL models for these modalities and tasks.

Clinical implementation

Integrating a DL model for target auto-segmentation into clinical practice defines its clinical implementation. Although the literature often distinguishes between in-house and commercially available solutions, differentiating between them can be challenging due to factors such as the data used for training and the models themselves, both of which can be either in-house or out-house. Figure 3 provides an overview of the various possibilities. In this context, an in-house solution refers to the implementation of a model developed internally and trained on institutional data. In-house software for target auto-segmentation offers several advantages. For instance, the ground truth data used for training aligns with local delineation guidelines, likely resulting in fewer adjustments to the outputs and a better understanding of when the model will work well and when it might not. Moreover, in-house solutions are typically developed within hospitals, where data availability is less constrained compared to external companies. However, there are also notable challenges with in-house solutions, particularly related to legal regulations. Both commercial and in-house auto-segmentation tools must comply with medical device regulations (MDR), but while regulatory compliance is handled by vendors for commercial tools, this responsibility falls on the hospital for in-house solutions. This requirement introduces an additional layer of complexity, necessitating careful navigation to ensure compliance and in-house expertise for maintenance and support.

To address some of these challenges, recent studies have utilized internal data to train models provided by external companies. For example, a study by Bakx et al[36] illustrates this approach in the context of CTV delineation for breast cancer. In their study, the authors trained a DL segmentation model provided by a commercial software company using their patient data. The generated CTV contours were then checked and corrected by radiation oncologists. The entire process of loading scans, checking, and correcting was timed and compared against the time needed for manual segmentation. The study found that radiation oncologists deemed 83% of the CTVs as clinically acceptable or usable as starting point. The results indicate that while some corrections were necessary, the DL approach significantly reduced the time required for segmentation and produced final contours of

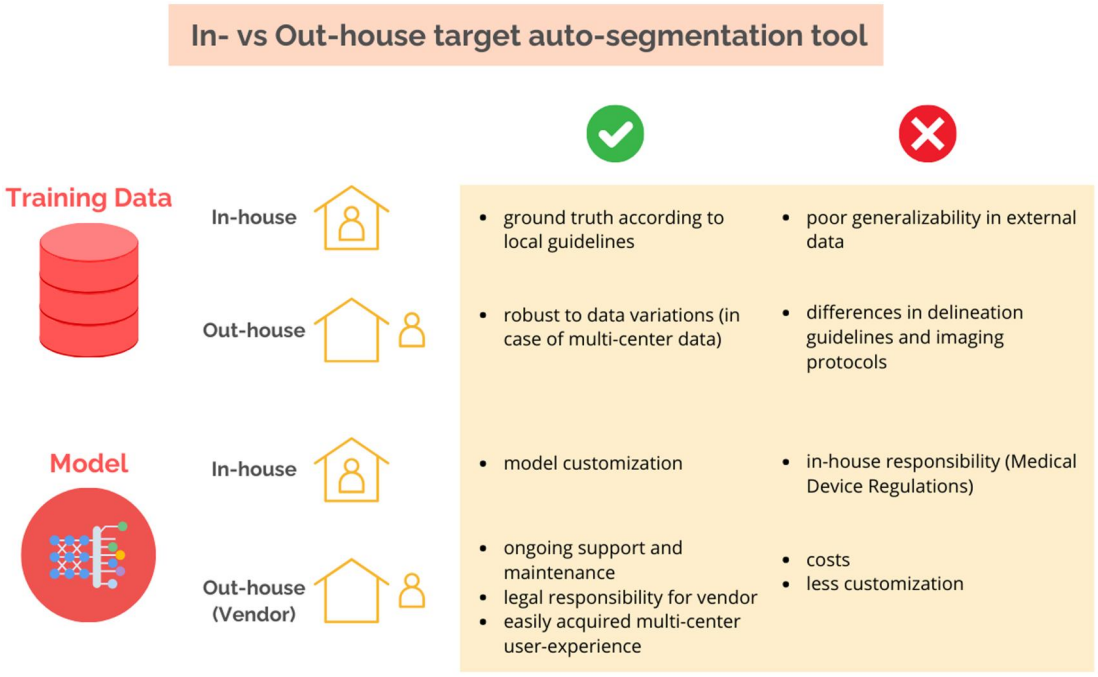


Figure 3. Advantages and disadvantages of using in-house versus out-house data or models in clinical implementation.

higher quality compared to the inter-observer variability found in similar studies.

Currently, the commercially available AI solutions for target auto-segmentation are limited to CTVs of a few tumour sites: the prostate[37-42], the breast[43], and the head and neck lymph nodes levels for elective irradiation[38]. Notably, all identified studies exclusively utilized single-modality CT. In some studies, the algorithms of these AI solutions were not exposed to the study cohort on which they were tested[38,41,42], while in other cases, the models were re-trained using a small centre-specific dataset[37,39,40]. To evaluate the results of AI-driven solutions, the auto-segmented structures are typically compared to manually delineated structures using commonly employed metrics. Notably, in almost all these studies[37,39-43], radiation oncologists were also invited to assess the clinical acceptability of the auto-segmentations. In all referenced studies except for[41], the evaluation was blinded to avoid bias. The criteria for clinical acceptability varied across studies, with some focusing on specific anatomical thresholds for deviations[39], while others assessed the extent of required corrections[37,40-43]. Most studies employed a Likert scale for qualitative assessment, though the scale's granularity varied: some used a 5-point scale[37,39], others a 4-point scale[40], and one used a 3-point scale[43]. The AI auto-segmentations were considered clinically acceptable for treatment planning under the supervision of radiation oncologists who could perform manual editing in outlier scenarios. The studies vary in expert involvement for evaluating auto-segmentations. Studies such as[37,40,42,43] utilized multiple experts, including panels of radiation oncologists and residents, to ensure a comprehensive assessment. In contrast, Duan et al[39] and Palazzo et al[41] relied on evaluations by individual experts, focusing on personal assessments of clinical acceptability and performance.

Why are AI solutions for target auto-segmentation only available for these tumour sites? Firstly, it is crucial to emphasize that prostate gland segmentation benefits from reduced anatomical site variation, as the prostate is an inherent part of anatomy rather than pathology. The same applies to the breast and head and neck elective lymph node levels, as the CTV in these cases does not include the pathologic nodes, but rather the anatomically defined whole breast or head and neck lymph node levels, respectively. Hence, the availability of AI solutions for auto-segmentation of OARs made these targets logical first candidates for target auto-segmentation in clinical practice. Notably, checking and manually editing auto-segmented structures within anatomical regions is significantly more straightforward than doing so within pathological regions, where consultation of multiple imaging modalities is often required.

What makes the implementation of target auto-segmentation so challenging in clinical practice? One of the foremost challenges still remains offering a *comprehensive performance evaluation* that incorporates both qualitative and quantitative metrics[40]. Currently, in most studies, auto-segmentation results are compared to manually performed contours performed by 1 or 2 radiation oncologists, which are considered the reference. In some cases, however, consensus contours are used, either provided directly by a larger group of experts or generated by combining several individual contours[44]. While this approach offers a more reliable benchmark, it is also more labour-intensive. This highlights the potential need for high-quality, open-source gold-standard datasets that could streamline algorithm validation and provide a robust foundation for consistent evaluation across studies. Overlapping metrics (DSC, precision, recall) are the most commonly used in the majority of technical studies, followed by distance-based metrics (Hausdorff distance, mean surface distance, surface DSC). Those metrics allow an easier and faster comparison of multiple trained DL models, especially during hyperparameter tuning; however, they do not always correlate with clinical acceptability[20,45]. In addition, in clinical practice, there are more factors to consider in order to prefer automatic systems over manual. It is necessary to prove that the new (or corrected) auto-segmentations allow a full coverage of the target (including dosimetric performance[21]) and ensure time-saving[45]. In a recent study from Ferreira Silvério et al[46], the authors provided such a comprehensive evaluation when assessing the performance and clinical usability of DL CTV auto-contouring in rectal cancer. They built and tested a framework in which they (1) quantitatively evaluated model performance both against the reference contours and the inter-observer variability; (2) performed expert review and correction of the auto-segmentations, quantifying the time required; (3) evaluated the impact on expected target coverage. In Figure 4, an adapted version of such a framework is provided, summarizing the key aspects for the evaluation of DL target auto-segmentation models.

According to Mackay et al[47], the *absence of a standardized approach for validating* auto-contouring systems is the root cause of the delay in clinical implementation. When the final goal is to demonstrate that auto-contouring systems can achieve human-level performance, it is crucial to ensure that fair comparisons are conducted across studies. In their review, they provided a multi-step solution consisting of 4 stages: model training and choice, commissioning (physics), commissioning (clinical), case-specific QA. In the first 2 stages, they suggested to produce the ground truth data to use for training, using guidelines and protocols. Additionally, they recommended using commonly employed evaluation metrics to select the best-performing model. The third stage involved a more comprehensive clinical evaluation of the auto-segmentation models, wherein a subset of cases is assessed. This evaluation included comparing the results with intra- and inter-observer variability, as well as evaluating dosimetry and time efficiency. Finally, ongoing QA was proposed as an essential component, involving continuous assessment of validity through quality checks conducted by clinicians for each case.

It is essential to consider broader guidelines or frameworks for the evaluation and continuous QA of auto-segmentation software in radiotherapy. A unified approach endorsed by international organizations could significantly enhance the consistency and reliability of these tools in clinical practice.

Risk analysis

When introducing a new technology into clinical workflows, it is extremely important to understand its associated risks. In a study from Nealon et al[48], the authors utilized the failure mode and effects analysis approach to quantify the risks associated with an automated contouring and treatment planning tool which was under development in their institute. Interestingly, among the 290 identified failure modes, the largest group regarded the contour edits, review, and upload process steps in the workflow. This analysis serves as an

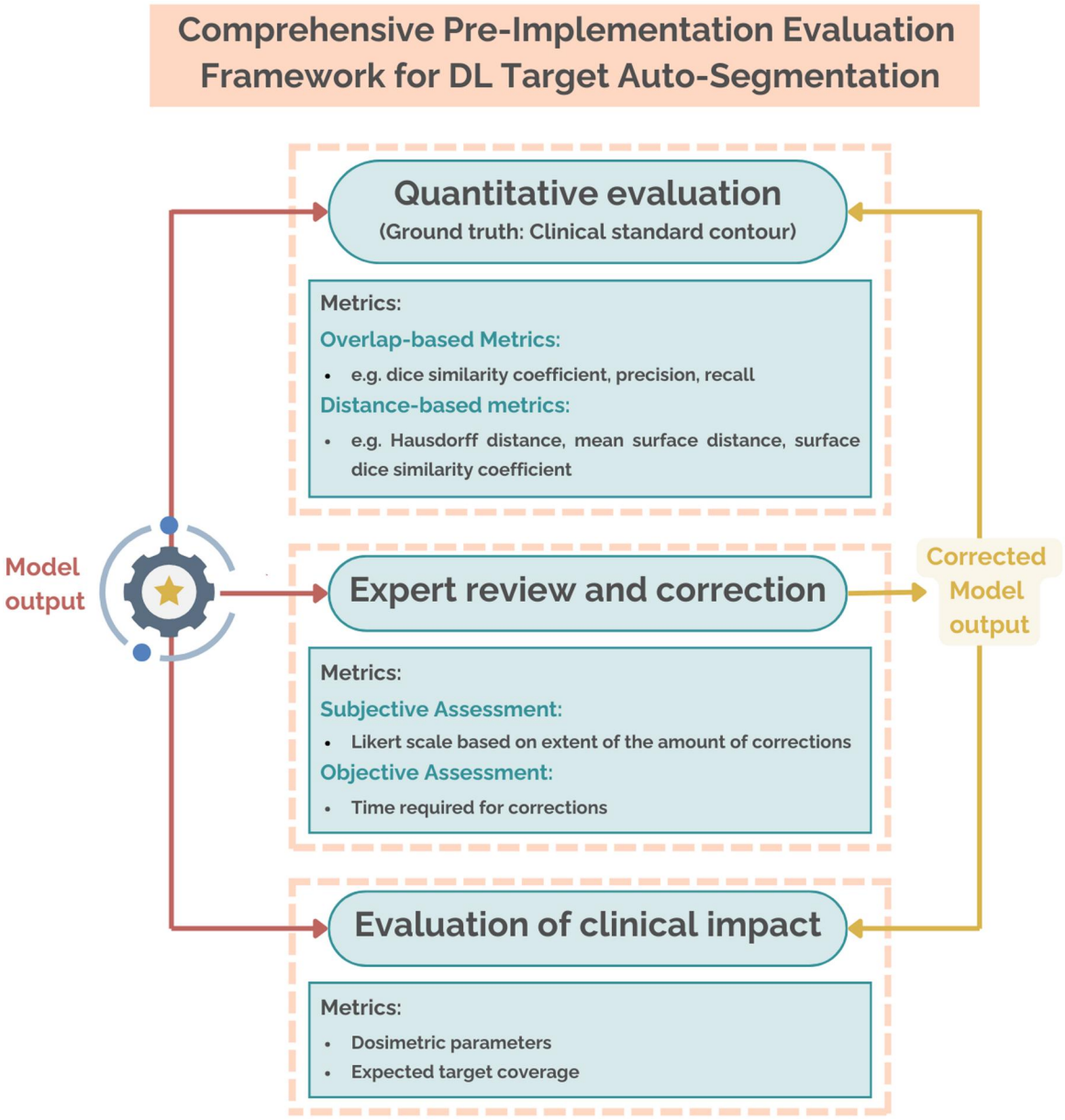


Figure 4. Framework for comprehensive evaluation of DL target auto-segmentation pre-implementation. The evaluation of a DL model's output for target auto-segmentation should include (1) a quantitative assessment, (2) expert review and correction, and (3) evaluation of the expected target coverage. Abbreviation: DL = deep learning.

exemplary model, demonstrating how clinics could approach risk assessments when implementing automated software.

Based on the findings from Nealon et al[48] and our own experience, we identified examples of potential failures specifically associated with the clinical use of DL target auto-segmentation models, potential consequences in the radiotherapy treatment setting and potential mitigation strategies. These failures can be attributed to the segmentation model itself or to the end user. Failures of target auto-segmentation models arise from several factors, including the type and quality of the input data, as well as the definition of the ground truth used for model training. Regarding the input data, factors such as image quality, resolution disparities, and variations in tumour or patient characteristics can introduce inconsistencies that challenge segmentation algorithms. Additionally, the presence of artefacts within imaging data can distort segmentation results, compromising the accurate delineation of the target structures. Furthermore, the complexity of defining target boundaries on scans, the segmentation guidelines inclined to clinical interpretation, and the lack of an adequate education can lead to subjective interpretations and discrepancies in segmentations[49]. Inconsistent segmentations used for model training may affect the quality of the model's output. This challenge is particularly pronounced in tumour sites where target segmentation involves integrating information from multiple imaging modalities and clinical factors, thus adding additional layers of complexity. Relying solely on clinical judgement to establish ground truth introduces subjectivity, as interpretations of a patient's condition may vary among practitioners. Addressing failures in auto-segmentation requires a comprehensive understanding of the diverse factors influencing both input data type/quality and the definition of a ground truth. Moreover, to

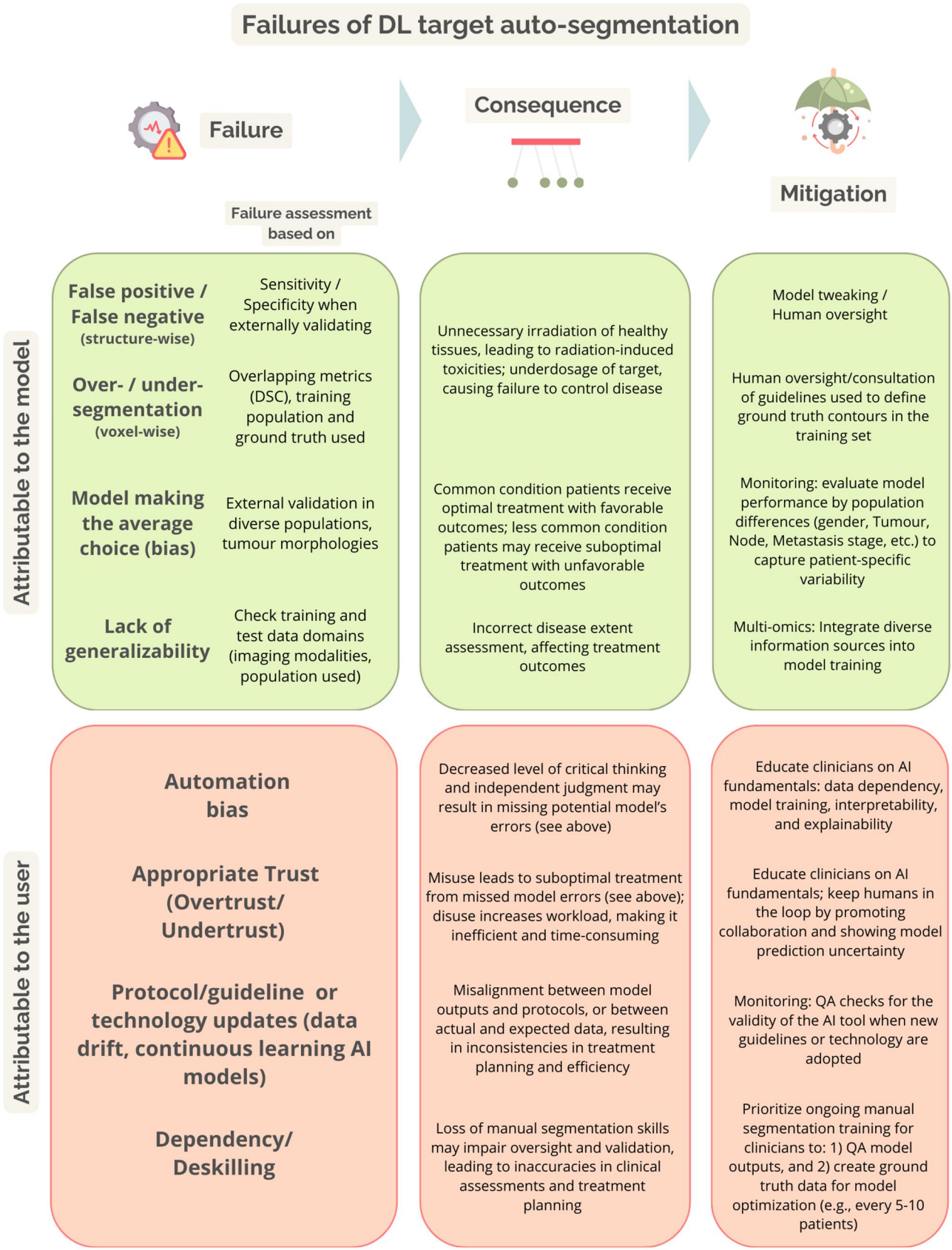


Figure 5. Overview of possible failures attributable to a DL target auto-segmentation model (in green) or to the user (in orange) of a DL target auto-segmentation model. Abbreviation: DL = deep learning.

mitigate associated clinical consequences and ensure patient safety, it is necessary to have information about the training data, the model performance in external sets, and those populations' characteristics (see Figure 5).

User interaction and utilization of models for target auto-segmentation may also lead to failures in clinical practice[48]. The clinical risks associated with these failures lie in the trust in target auto-segmentation models, in the potential loss of skills resulting from the adoption of such tools, and the need for continuous updates to align with evolving guidelines and protocols (seeFigure 5). Relying excessively (overtrust) on DL models may result in a decreased level of critical thinking in assessing output target contours, potentially resulting in sub-optimal treatment planning and patient outcomes. Conversely, insufficient trust (undertrust) may be an obstacle for technological progress, compromising the transition to innovative techniques like online adaptive radiotherapy where rapid target segmentation is crucial. The involvement of radiation oncologists and radiation therapists is pivotal to ensure that these applications facilitate personalized and high-quality treatment for patients[50]. However, it is necessary that professionals adapt their tasks alongside the technology. In the context of target auto-segmentation, professionals will need to master skills related to data dependency (input imaging and ground truths) and model training, explainability, and interpretability[51]. It is expected, in fact, that a higher experience with AI will lead to a more appropriate use of[52]. Finally, to ensure the accuracy and reliability of the target auto-segmentation model outputs over time, these skills need to be incorporated into robust QA practices.

Quality assurance

In radiotherapy, many tasks are prone to error that can have significant consequences for patients' treatment outcomes. Thus, QA procedures are essential standards of care to detect and prevent these potential failures. Human inspection, particularly in the form of peer review, plays a crucial role in traditional manual target segmentation QA[53]. A recent study showed that 38.5% of cases needed changes in head and neck cancer PTV definition after peer review volume rounds. Minor changes were recommended in 16% of cases, whereas major changes were advised in 22.5% of cases. Despite the often subjective nature of minor revisions, the study illustrates that manual contours could benefit from additional editing or consultations before treatment planning[54]. With the introduction of DL target auto-segmentation, it becomes important to routinely monitor DL systems to notice decreased performance due to shifts in protocols, guidelines, technology, or input data distribution. When guidelines change, the solution provided by the system may become outdated for the end user.

Claessens et al[55] provided a comprehensive overview of QA techniques for AI-based applications in radiotherapy. They delved into auto-segmentation and distinguished between *case-specific QA*, which involves assessing the quality of auto-segmentation on a case-by-case basis, and *routine QA*, which entails periodic monitoring of the model's output to ensure its overall quality and validity. In *case-specific QA*, human oversight remains indispensable for mitigating the potential associated clinical risks[27]. A common practice involves visual checking and slice-by-slice editing of the output of the auto-segmentation model[56], which, however, reduces the time-saving of the automated versus the manual segmentation. Considering the complexities of the target segmentation task, largely discussed in the first section, which often require the integration of multiple imaging modalities and careful consideration of patient data, it becomes apparent that manually adjusting auto-segmented target structures can be significantly more time-consuming and challenging compared to OARs.

To speed up the revision process, automated segmentation QA and intuitive human-machine interfaces can enhance time-saving and consistency in contouring. Statistical models are commonly used to detect errors in automatically segmented structures by identifying outliers. These outliers are segmentations that deviate from expected population-based characteristics such as volume, shape, and intensity[57]. These models rely on pre-defined statistical thresholds to find anomalies. Machine learning models are increasingly used to assess segmentation quality. These models are trained to classify whether a segmentation is accurate or not, using the same features but modelling more complex variations in the data[58]. Recently, the topic of uncertainty estimation and quantification drew considerable attention due to its potential dual benefits: (1) reducing the editing/checking time of performed contours by employing metrics that identify cases/areas for which the models are more uncertain in giving predictions; (2) enhancing clinicians' trust in AI-based auto-segmentation models, enabling a better understanding of where and why these uncertainties arise[19]. In the subsection 'Towards optimizing case-specific QA: uncertainty estimation', several solutions for uncertainty estimation will be discussed.

For *routine QA* and *monitoring*, benchmark datasets have been proposed to periodically test end-to-end performance[55]. It is crucial that the datasets accurately represent the real-world patient population[59], requiring an ongoing upgrade that regularly includes new patients, especially when new scans or guidelines for target segmentation are adopted. The frequency of monitoring should be determined based on these changes and considering potential clinical risks[60]. An alternative monitoring strategy involves conducting statistical analysis on data acquired after manual adjustments have been performed to auto-segmentations. In a recent paper by Nealon et al[61], the authors showed that tracking the magnitude of required edits over time can help identifying failures attributable to both the model and the end user. An increase in editing could signal deficiencies in the model, while a decrease might imply the end user's over-reliance on automation.

Towards optimizing case-specific QA: uncertainty estimation

To ensure that metrics derived from DL estimated uncertainty are suitable for case-specific QA, they must demonstrate well-calibrated relations with segmentation accuracy metrics[62]. Recent literature offers limited studies addressing this topic in target auto-segmentation for radiotherapy. In Table 2, we provide an overview of these studies for various tumour sites, including head and neck[63,64], lung[65], prostate[66], cervical, and rectal cancer[67]. While promising calibration was generally observed, most studies introduced new metrics to quantify both the estimated uncertainty and segmentation quality, underscoring the absence of a valid and standardized approach. The meaning and implications of uncertainty information for end users remain unclear. Therefore, some studies conducted additional investigations

Table 2. Overview of studies on quality assurance using DL uncertainty estimation.

Tumour site (Target)	Imaging modalities	Training dataset (*n*: # test cases)	Uncertainty estimation	Details	Calibration performance
Oropharynx (GTVp)[64]	CT, FDG PET	HECKTOR 2021 dataset (External MDA: $n = 67$)	Deep Ensemble and the Monte Carlo Dropout Ensemble	Segmentation accuracy: DSC Uncertainty quantification metrics: CV, structure expected entropy, structure predictive entropy, and structure mutual information, Dice-risk Calibration quantified by AvU metric and linear correlation between uncertainty and DSC	AvU for CV (both approaches): 0.87 Pearson correlation coefficients (both approaches): 0.72 CV
Oropharynx (GTVp and GTVln)[63]	CT, FDG PET	Internal dataset (Internal: $n = 104$, External (HECKTOR 2022): $n = 519$)	Deep Ensemble of not binarized sigmoid outputs	Output type: Probability maps obtained as the average of multiple outputs of models trained to perform multi-label segmentation. Segmentation accuracy: average of surface DSC calculated after thresholding using different probability thresholds, per patient and per structure. Uncertainty quantification metric: patient and structure-specific CV Calibration: linear correlation between structure-specific CV and average surface DSC	Pearson correlation coefficients: Internal test −0.63 GTVp −0.63 GTVln External test: −0.54 GTVp −0.66 GTVln
Lung (GTV)[65]	CT	Lung1 dataset ($n = 42$)	Monte Carlo dropout with softmax output	Output type: (1) Probability maps calculated as the average of the outputs; (2) Uncertainty maps obtained by normalized entropy Uncertainty quantification metric: MU, RUV Calibration: reliability diagrams Clinical evaluation: correlation between MU and RUV and clinician's uncertainty estimation (high, intermediate, low uncertainty).	AUC of proposed metrics: 0.849-0.876 MU 0.719-0.804 RUV Pearson correlation coefficients: 0.68 MU 0.49 RUV
Prostate (CTV)[66]	MRI	Multicentre dataset (External: $n = 49$)	Monte Carlo dropout	Output type: (1) the average segmentation of the multiple predictions; (2) a binary uncertainty band obtained thresholding the normalized standard deviation of the predictions Quality assessment: logistic regression that classified manual delineations (pass/discrepancy) based on their spatial association with both the network's outputs.	AUC = 0.92 TPR = 0.92 FPR = 0.09
Cervix (GTV)[67]	MRI (T2w)	Internal dataset ($n = 53$)	Not binarized softmax output	Output type: 'score map' Segmentation accuracy: DSC, 95th HD, MSD, surface DSC Uncertainty quantification metrics: HiS metric = mean of the score map values higher than a threshold ($\lambda = 0.3$). Calibration: correlation between HiS and accuracy metrics	Pearson correlation coefficients: 0.79 DSC −0.60 95th HD −0.66 MSD 0.67 surface DSC
Rectal (CTV)[67]	MRI (T2w)	Internal dataset ($n = 38$)			Pearson correlation coefficients: 0.76 DSC −0.53 95th HD −0.73 MSD 0.62 surface DSC

Abbreviations: AvU = accuracy versus uncertainty; CV = coefficient of variation; MU = mean uncertainty; RUV = relative uncertainty volume; AUC = area under the ROC curve; TPR = true positive rate; FPR = false positive rate; DSC = dice similarity coefficient; 95th HD = Hausdorff distance 95th percentile; MSD = mean surface distance.

to shed light on its significance. Interestingly, one study[65] compared DL network uncertainty with areas of uncertainty identified by 2 radiation oncologists, revealing high overall agreement between the 2. In a different work[63], authors investigated potential correlations between patient-specific characteristics (tumour volume, imaging features) and both segmentation accuracy and uncertainty metrics, showing similar results for the uncertainty in both internal and external test sets.

Future directions

Will DL target auto-segmentation software ever be independent of human oversight? With the implementation of the new regulations outlined in the EU AI Act[68], it is unlikely. High-risk AI systems, such as tumour auto-segmentation software, should be designed and developed in such a way that it can be regulated by humans. In line with the risks identified in this review, human oversight remains essential. Thus, research should be carried out to fasten and facilitate the human-machine interaction by proposing solutions that are interactive and efficient, and by combining clinically relevant metrics and visual guidance assessment.

In this context, addressing the challenge of limited annotated data becomes crucial. One promising approach is the Segment Anything Model (SAM)[69], which has transformed the landscape of image segmentation. SAM generates accurate object masks from natural images using a Vision Transformer-based encoder, a prompt encoder for user interactions, and a mask decoder for reconstructing segmentation masks. Users guide the segmentation process with prompts such as points or bounding boxes around the target. Recently, a new refined version was introduced that is fine-tuned on over 1 million medical image-mask pairs, called MedSAM[70]. Despite the technical progress, research should investigate if and how this type of interactive model can be useful and relevant to radiation oncologists to perform tumour segmentation in clinical practice.

Uncertainty and probability maps represent a promising advancement in facilitating the clinical implementation of target auto-segmentation. Patient-specific QA could be facilitated if adaptive techniques, such as interactive probability maps, would be used for target auto-segmentation. The visual aspect of these maps, combined with well-calibrated uncertainty quantification metrics, would introduce a dynamic element where users are actively engaged, mitigating the risk of over-reliance on automated processes. Recently, a study engaged clinicians in developing a user graphical interface facilitating interaction between radiation oncologists and probability maps derived from a DL GTVp auto-segmentation model trained on PET/CT images of head and neck cancer patients[71]. A user test was also conducted involving 9 expert radiation oncologists to assess the clinical utility of the output. Based on the analysis, end users perceived probability maps as more intuitive and transparent compared to the single contour outputs typically generated by auto-segmentation models. Furthermore, visualizing the model's confidence in its predictions facilitated a clearer comprehension of the decision-making process underlying the model. An immediate next step involves leveraging the insights from this study for the deployment of similar tools. Subsequently, investigating the potential benefits of such an approach in reducing time requirements and mitigating inter-observer variability is imperative.

In the interaction with the output of DL auto-segmentation models, users should not only consider the model's segmentation uncertainty (which is ideally informative of segmentation accuracy) but should have a focus on regions where differences in contouring have a larger impact on the patient treatment outcome. In a recent study from Roberfroid et al[72], the authors proposed the DIVE (Dosimetrically Informed Volume Edition) map, a 'heatmap' tool to visually guide clinicians in correcting auto-segmented contours, by only focusing on dosimetrically relevant regions. The tool was evaluated on auto-segmented bladder and rectum from 3 prostate cancer patients treated with adaptive radiotherapy. The preliminary results demonstrated that the tool was able to discard corrections with no dosimetric impact, although its applicability in scenarios different than adaptive radiotherapy may be more challenging to test.

Finally, a reflection is needed on the availability and validity of the ground truth data used to train DL target auto-segmentation models. While manual contours offer the best estimation of tumours based on clinical knowledge and imaging, they do not represent the true ground truth, the histopathological tumour does. Considering that DL models rely on these contours, this raises the question: do manual contours created for radiotherapy treatment accurately represent the histopathological tumour? Only a few studies in literature investigated this topic due to the complexity and uncertainty in aligning imaging with pathological images. Nevertheless, a recent study by Terzidis et al[73] revealed that PET/MRI imaging tends to overestimate the pathological tumour volume in head and neck squamous cell carcinoma. Improving manual contouring through clearer and more specific guidelines, as well as investigating the relation between imaging and pathology, could improve the consistency of the ground truth data leading to more robust auto-segmentation models.

Conclusion

The clinical adoption of DL target auto-segmentation is facing delays due to the numerous challenges and considerations it poses. There is a notable discrepancy in the pace at which commercially available software and AI researchers deliver solutions for target auto-segmentation in radiotherapy. In fact, while DL models offer promising solutions to improve efficiency and consistency in target delineation, significant hurdles remain in ensuring their safe and effective utilization. The lack of a standardized approach for QA and monitoring of DL target auto-segmentation models significantly hinders its clinical implementation. Moreover, in the absence of human oversight, DL models may produce target auto-segmentations that could result in suboptimal radiotherapy treatment (over- or under-dosage), with serious consequences for patient outcomes. User interaction with DL target auto-segmentation software may also introduce errors, as a result of overtrust or undertrust, and thereby potentially compromising treatment planning. To mitigate associated clinical risks and ensure patient safety while promoting effective model utilization, strategies include educating professionals, promoting collaborative decision-making, prioritizing ongoing training in manual segmentation skills, and developing robust QA practices. These practices should incorporate human oversight guided by well-calibrated uncertainty quantification metrics, adaptation of professional tasks alongside technological advancements, and ongoing monitoring for end-to-end performance assessment when data evolves.

Additionally, further research should delve into expanding the scope of QA to include geometric, dose-volume, and outcome-based performance monitoring. Overall, while DL target auto-segmentation holds great potential to revolutionize radiotherapy treatment planning, careful consideration of clinical risks and rigorous QA measures are imperative for its successful integration into clinical practice.

Funding

A.D.B. is on a PhD position funded by the Hanarth Foundation.

Conflicts of interest

The authors declare no conflict of interests.

References

1. Kashyap A, Rapsomaniki MA, Barros V, et al. Quantification of tumor heterogeneity: from data acquisition to metric generation. *Trends Biotechnol*. 2022;40(6):647-676. doi:10.1016/J.TIBTECH.2021.11.006
2. ICRU Report 50, Prescribing, Recording, and Reporting Photon Beam Therapy—ICRU. Accessed August 20, 2024. https://www.icru.org/report/prescribing-recording-and-reporting-photon-beam-therapy-report-50/
3. Niyazi M, Andratschke N, Bendszus M, et al. ESTRO-EANO guideline on target delineation and radiotherapy details for glioblastoma. *Radiother Oncol*. 2023;184:109663. doi:10.1016/J.RADONC.2023.109663
4. Salembier C, Villeirs G, De Bari B, et al. ESTRO ACROP consensus guideline on CT- and MRI-based target volume delineation for primary radiation therapy of localized prostate cancer. *Radiother Oncol*. 2018;127(1):49-61. doi:10.1016/j.radonc.2018.01.014
5. Grégoire V, Evans M, Le QT, et al. Delineation of the primary tumour Clinical Target Volumes (CTV-P) in laryngeal, hypopharyngeal, oropharyngeal and oral cavity squamous cell carcinoma: AIRO, CACA, DAHANCA, EORTC, GEORCC, GORTEC, HKNPCSG, HNCIG, IAG-KHT, LPRHHT, NCIC CTG, NCRI, NRG Oncology, PHNS, SBRT, SOMERA, SRO, SSHNO, TROG consensus guidelines. *Radiother Oncol*. 2018;126(1):3-24. doi:10.1016/j.radonc.2017.10.016
6. Le Pechoux C, Faivre-Finn C, Ramella S, et al. ESTRO ACROP guidelines for target volume definition in the thoracic radiation treatment of small cell lung cancer. *Radiother Oncol*. 2020;152:89-95. doi:10.1016/j.radonc.2020.07.012
7. 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(1):205-208. doi:10.1016/J.RADONC.2015.12.027
8. Vinod SK, Jameson MG, Min M, Holloway LC. Uncertainties in volume delineation in radiation oncology: a systematic review and recommendations for future studies. *Radiother Oncol*. 2016;121(2):169-179. doi:10.1016/J.RADONC.2016.09.009
9. Guzene L, Beddok A, Nioche C, et al. Assessing interobserver variability in the delineation of structures in radiation oncology: a systematic review. *Int J Radiat Oncol Biol Phys*. 2023;115(5):1047-1060. doi:10.1016/j.ijrobp.2022.11.021
10. Franco P, Arcadipane F, Trino E, et al. Variability of clinical target volume delineation for rectal cancer patients planned for neoadjuvant radiotherapy with the aid of the platform Anatom-e. *Clin Transl Radiat Oncol*. 2018;11:33-39. doi:10.1016/j.ctro.2018.06.002
11. Li F, Li Y, Wang X, et al. Inter-observer and intra-observer variability in gross tumor volume delineation of primary esophageal carcinomas based on different combinations of diagnostic multimodal images. *Front Oncol*. 2022;12:817413. doi:10.3389/FONC.2022.817413/BIBTEX
12. Liu X, Huang H, Zhu C, et al. Interobserver variations in target delineation in intensity-modulated radiation therapy for nasopharyngeal carcinoma and its impact on target dose coverage. *Technol Cancer Res Treat*. 2023;22:15330338231169592. doi:10.1177/15330338231169592
13. Lee N, Xia P, Fischbein NJ, Akazawa P, Akazawa C, Quivey JM. Intensity-modulated radiation therapy for head-and-neck cancer: the UCSF experience focusing on target volume delineation. *Int J Radiat Oncol Biol Phys*. 2003;57(1):49-60. doi:10.1016/S0360-3016(03)00405-X
14. Kim K, Chun M, Jin H, et al. Inter-institutional variation in intensity-modulated radiotherapy for breast cancer in Korea (KROG 19-01). *Anticancer Res*. 2021;41(6):3145-3152. doi:10.21873/ANTICANRES.15100
15. Lin D, Lapen K, Sherer MV, et al. A systematic review of contouring guidelines in radiation oncology: analysis of frequency, methodology, and delivery of consensus recommendations. *Int J Radiat Oncol Biol Phys*. 2020;107(4):827-835. doi:10.1016/j.ijrobp.2020.04.011
16. Staal FHE, Siang KNW, Brouwer CL, et al. Pretrial quality assurance for hypofractionated salvage radiation therapy after prostatectomy in the multi-institutional PERYTON-trial. *Adv Radiat Oncol*. 2024;9(2):101379. doi:10.1016/J.ADRO.2023.101379
17. Fuller CD, Nijkamp J, Duppen JC, et al. Radiation Oncology Committee of the Southwest Oncology Group. Prospective randomized double-blind pilot study of site-specific consensus atlas implementation for rectal cancer target volume delineation in the cooperative group setting. *Int J Radiat Oncol Biol Phys*. 2011;79(2):481-489. doi:10.1016/J.IJROBP.2009.11.012
18. Isaksson LJ, Summers P, Mastroleo F, et al. Automatic segmentation with deep learning in radiotherapy. *Cancers (Basel)*. 2023;15(17):4389. doi:10.3390/CANCERS15174389/S1
19. Savjani RR, Lauria M, Bose S, Deng J, Yuan Y, Andrearczyk V. Automated tumor segmentation in radiotherapy. *Semin Radiat Oncol*. 2022;32(4):319-329. doi:10.1016/J.SEMRADONC.2022.06.002
20. Rong Y, Chen Q, Fu Y, et al. NRG oncology assessment of artificial intelligence deep learning-based auto-segmentation for radiation therapy: current developments, clinical considerations, and future directions. *Int J Radiat Oncol Biol Phys*. 2024;119(1):261-280. doi:10.1016/J.IJROBP.2023.10.033
21. Matoska T, Patel M, Liu H, Beriwal S. Review of deep learning based autosegmentation for clinical target volume: current status and future directions. *Adv Radiat Oncol*. 2024;9(5):101470. doi:10.1016/J.ADRO.2024.101470
22. Hasler SW, Kallehauge JF, Hansen RH, et al. Geometric distortions in clinical MRI sequences for radiotherapy: insights gained from a multicenter investigation. *Acta Oncol (Madr)*. 2023;62(11):1551-1560. doi:10.1080/0284186X.2023.2266560
23. Ferreira A, Solak N, Li J, et al. How we won BraTS 2023 adult glioma challenge? Just faking it! Enhanced synthetic data augmentation and model ensemble for brain tumour segmentation. arXiv, https://doi.org/10.48550/arXiv.2402.17317, February 27, 2024, preprint: not peer reviewed.
24. Qu J, Zhang W, Shu X, et al. Construction and evaluation of a gated high-resolution neural network for automatic brain metastasis detection and segmentation. *Eur Radiol*. 2023;33(10):6648-6658. doi:10.1007/s00330-023-09648-3
25. Myronenko A, Siddiquee MMR, Yang D, He Y, Xu D. Automated head and neck tumor segmentation from 3D PET/CT HECKTOR 2022 challenge report. *Lect Notes Comput Sci*. 2023;13626LNCS:31-37. doi:10.1007/978-3-031-27420-6_2/TABLES/2
26. Luo X, Liao W, He Y, et al. Deep learning-based accurate delineation of primary gross tumor volume of nasopharyngeal carcinoma on heterogeneous magnetic resonance imaging: a large-scale and multi-center study. *Radiother Oncol*. 2023;180:109480-doi:10.1016/J.RADONC.2023.109480

27. Lee B. m, Kim JS, Chang Y, et al. Experience of implementing deep learning-based automatic contouring in breast radiation therapy planning: insights from over 2,000 cases. *Int J Radiat Oncol Biol Phys*. 2024;119(5):1579-1589. doi:10.1016/J.IJROBP.2024.02.041
28. Primakov SP, Ibrahim A, van Timmeren JE, et al. Automated detection and segmentation of non-small cell lung cancer computed tomography images. *Nat Commun*. 2022;13(1):3423. doi:10.1038/s41467-022-30841-3
29. Zhang S, Li K, Sun Y, et al. Deep learning for automatic gross tumor volumes contouring in esophageal cancer based on contrast-enhanced CT images: a multi-institutional study. *Int J Radiat Oncol Biol Phys*. 2024;119(5):1590-1600. doi:10.1016/J.IJROBP.2024.02.035
30. Myronenko A, Yang D, He Y, Xu D. Automated 3D segmentation of kidneys and tumors in MICCAI KiTS 2023 challenge. In: Heller N, Wood A, Isensee F, et al., eds. Kidney and Kidney Tumor Segmentation. KiTS 2023. Lecture Notes in Computer Science, Vol. 14540. Springer; 2024:1-7. doi:10.1007/978-3-031-54806-2_1
31. Dai S, Zhu Y, Jiang X, Yu F, Lin J, Yang D. TD-Net: trans-deformer network for automatic pancreas segmentation. *Neurocomputing*. 2023;517:279-293. doi:10.1016/J.NEUCOM.2022.10.060
32. Shui Y, Wang Z, Liu B, Wang W, Fu S, Li Y. A three-path network with multi-scale selective feature fusion, edge-inspiring and edge-guiding for liver tumor segmentation. *Comput Biol Med*. 2024;168:107841-doi:10.1016/j.compbiomed.2023.107841
33. Jimenez-Pastor A, Lopez-Gonzalez R, Fos-Guarinos B, et al. Automated prostate multi-regional segmentation in magnetic resonance using fully convolutional neural networks. *Eur Radiol*. 2023;33(7):5087-5096. doi:10.1007/s00330-023-09410-9
34. 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):1-10. doi:10.1186/S13014-023-02350-0/FIGURES/7
35. Ma CY, Zhou JY, Xu XT, et al. Deep learning-based auto-segmentation of clinical target volumes for radiotherapy treatment of cervical cancer. *J Appl Clin Med Phys*. 2022;23(2):e13470. doi:10.1002/ACM2.13470
36. Bakx N, Van der Sangen M, Theuws J, Bluemink J, Hurkmans C. Comparison of the use of a clinically implemented deep learning segmentation model with the simulated study setting for breast cancer patients receiving radiotherapy. *Acta Oncol*. 2024;63:477-481. doi:10.2340/1651-226X.2024.34986
37. Duan J, Vargas CE, Yu NY, et al. Incremental retraining, clinical implementation, and acceptance rate of deep learning auto-segmentation for male pelvis in a multiuser environment. *Med Phys*. 2023;50(7):4079-4091. doi:10.1002/MP.16537
38. 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. doi:10.1016/J.RADONC.2019.10.019
39. Duan J, Bernard M, Downes L, et al. Evaluating the clinical acceptability of deep learning contours of prostate and organs-at-risk in an automated prostate treatment planning process. *Med Phys*. 2022;49(4):2570-2581. doi:10.1002/MP.15525
40. Hobbis D, Yu NY, Mund KW, et al. First report on physician assessment and clinical acceptability of custom-retrained artificial intelligence models for clinical target volume and organs-at-risk auto-delineation for postprostatectomy patients. *Pract Radiat Oncol*. 2023;13(4):351-362. doi:10.1016/J.PRRO.2023.03.011
41. Palazzo G, Mangili P, Deantoni C, et al. Real-world validation of artificial intelligence-based computed tomography auto-contouring for prostate cancer radiotherapy planning. *Phys Imaging Radiat Oncol*. 2023;28:100501. doi:10.1016/J.PHRO.2023.100501
42. Hoque SMH, Pirrone G, Matrone F, et al. Clinical use of a commercial artificial intelligence-based software for autocontouring in radiation therapy: geometric performance and dosimetric impact. *Cancers (Basel)*. 2023;15(24):5735. doi:10.3390/CANCERS15245735
43. Bakx N, Rijkaart D, van der Sangen M, et al. Clinical evaluation of a deep learning segmentation model including manual adjustments afterwards for locally advanced breast cancer. *Tech Innov Patient Support Radiat Oncol*. 2023;26:100211. doi:10.1016/j.tipsro.2023.100211
44. Guzene L, Beddok A, Nioche C, et al. Assessing interobserver variability in the delineation of structures in radiation oncology: a systematic review. *Int J Radiation Oncol Biol Phys*. 2023;115(5):1047-1060. doi:10.1016/j.ijrobp.2022.11.021
45. Baroudi H, Brock KK, Cao W, et al. Automated contouring and planning in radiation therapy: what is 'clinically acceptable'? *Diagnostics*. 2023;13(4):667. doi:10.3390/DIAGNOSTICS13040667
46. Ferreira Silvério N, van den Wollenberg W, Betgen A, et al. Evaluation of deep learning clinical target volumes auto-contouring for magnetic resonance imaging-guided online adaptive treatment of rectal cancer. *Adv Radiat Oncol*. 2024;9(6):101483-doi:10.1016/j.adro.2024.101483
47. Mackay K, Bernstein D, Glocker B, Kamnitsas K, Taylor A. A review of the metrics used to assess auto-contouring systems in radiotherapy. *Clin Oncol*. 2023;35(6):354-369. doi:10.1016/j.clon.2023.01.016
48. Nealon KA, Balter PA, Douglas RJ, et al. Using failure mode and effects analysis to evaluate risk in the clinical adoption of automated contouring and treatment planning tools. *Pract Radiat Oncol*. 2022;12(4):e344-e353. doi:10.1016/J.PRRO.2022.01.003
49. Stelmes JJ, Vu E, Grégoire V, et al. Quality assurance of radiotherapy in the ongoing EORTC 1420 "Best of" trial for early stage oropharyngeal, supraglottic and hypopharyngeal carcinoma: results of the benchmark case procedure. *Radiat Oncol*. 2021;16(1):1-10. doi:10.1186/S13014-021-01809-2/FIGURES/3
50. 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. doi:10.1016/J.PHRO.2020.11.002
51. Jones S, Thompson K, Porter B, et al. Automation and artificial intelligence in radiation therapy treatment planning. *J Med Radiat Sci*. 2024;71(2):290-298. doi:10.1002/jmrs.729
52. Horowitz MC, Kahn L. Bending the automation bias curve: a study of human and AI-based decision making in national security contexts. *Int Stud Q*. 2024;68(2):20. doi:10.1093/ISQ/SQAE020
53. Luk SMH, Ford EC, Phillips MH, Kalet AM. Improving the quality of care in radiation oncology using artificial intelligence. *Clin Oncol (R Coll Radiol)*. 2022;34(2):89-98. doi:10.1016/j.clon.2021.11.011
54. Hesse J, Chen L, Yu Y, et al. Peer review of head and neck cancer planning target volumes in radiation oncology. *Adv Radiat Oncol*. 2022;7(3):100917. doi:10.1016/J.ADRO.2022.100917
55. Claessens M, Oria CS, Brouwer CL, et al. Quality assurance for AI-based applications in radiation therapy. *Semin Radiat Oncol*. 2022;32(4):421-431. doi:10.1016/j.semradonc.2022.06.011
56. Brouwer CL, Boukerroui D, Oliveira J, et al. Assessment of manual adjustment performed in clinical practice following deep learning contouring for head and neck organs at risk in radiotherapy. *Phys Imaging Radiat Oncol*. 2020;16:54-60. doi:10.1016/J.PHRO.2020.10.001
57. Isaksson LJ, Summers P, Bhalerao A, et al. Quality assurance for automatically generated contours with additional deep learning. *Insights Imaging*. 2022;13(1):137. doi:10.1186/S13244-022-01276-7/FIGURES/7
58. Claessens M, Vanreusel V, De Kerf G, et al. Machine learning-based detection of aberrant deep learning segmentations of target and organs at risk for prostate radiotherapy using a secondary segmentation algorithm. *Phys Med Biol*. 2022;67(11):115014. doi:10.1088/1361-6560/AC6FAD
59. Jin X, Hao Y, Hilliard J, et al. A quality assurance framework for routine monitoring of deep learning cardiac substructure

computed tomography segmentation models in radiotherapy. *Med Phys*. 2024;51(4):2741-2758. doi:10.1002/MP.16846
60. Mahmood U, Shukla-Dave A, Chan HP, et al. Artificial intelligence in medicine: mitigating risks and maximizing benefits via quality assurance, quality control, and acceptance testing. *BJR Artif Intell*. 2024;1(1):ubae003. doi:10.1093/BJRAI/UBAE003
61. Nealon KA, Han EY, Kry SF, et al. Basic original report monitoring variations in the use of automated contouring software. *Pract Radiat Oncol*. 2024;14(1):e75-e85. doi:10.1016/j.prro.2023.09.004
62. Wahid KA, Sahlsten J, Jaskari J, et al. Harnessing uncertainty in radiotherapy auto-segmentation quality assurance. *Phys Imaging Radiat Oncol*. 2024;29:100526. doi:10.1016/J.PHRO.2023.100526
63. De Biase A, Sijtsema NM, van Dijk L, Steenbakkers RJHM, Langendijk JA, van Ooijen P. Uncertainty-aware deep learning for segmentation of primary tumour and pathologic lymph nodes in oropharyngeal cancer: insights from a multi-centre cohort. Preprint (Version 1), available at Research Square. Published online March 14, 2024. doi:10.21203/RS.3.RS-4091632/V1
64. Sahlsten J, Jaskari J, Wahid KA, et al. Application of simultaneous uncertainty quantification for image segmentation with probabilistic deep learning: performance benchmarking of oropharyngeal cancer target delineation as a use-case. medRxiv, https://doi.org/10.1101/2023.02.20.23286188, February 24, 2023, preprint: not peer reviewed.
65. Maruccio FC, Eppinga W, Laves MH, et al. Clinical assessment of deep learning-based uncertainty maps in lung cancer segmentation. *Phys Med Biol*. 2024;69(3):35007. doi:10.1088/1361-6560/ad1a26
66. Min H, Dowling J, Jameson MG, et al. Clinical target volume delineation quality assurance for MRI-guided prostate radiotherapy using deep learning with uncertainty estimation. *Radiother Oncol*. 2023;186:109794. doi:10.1016/J.RADONC.2023.109794
67. Rodríguez Outeiral R, Ferreira Silvério N, González PJ, et al. A network score-based metric to optimize the quality assurance of automatic radiotherapy target segmentations. *Phys Imaging Radiat Oncol*. 2023;28:100500. doi:10.1016/j.phro.2023.100500
68. European Commission. AI Act | Shaping Europe's digital future. European Commission. 2024. Accessed April 17, 2024. https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai
69. Kirillov A, Mintun E, Ravi N, et al. Segment anything. 2023 IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, October 1-6, 2023:3992-4003. doi:10.1109/ICCV51070.2023.00371
70. Ma J, He Y, Li F, Han L, You C, Wang B. Segment anything in medical images. *Nat Commun*. 2024;15(1):654. doi:10.1038/s41467-024-44824-z
71. De Biase A, Ziegfeld L, Sijtsema NM, et al. Probability maps for deep learning-based head and neck tumor segmentation: graphical user interface design and test. *Comput Biol Med*. 2024;177:108675-. doi:10.1016/J.COMPBIOMED.2024.108675
72. Roberfroid B, Lee JA, Geets X, Sterpin E, Barragán-Montero AM. DIVE-ART: a tool to guide clinicians towards dosimetrically informed volume editions of automatically segmented volumes in adaptive radiation therapy. *Radiother Oncol*. 2024;192:110108. doi:10.1016/j.radonc.2024.110108
73. Terzidis E, Friborg J, Vogelius IR, et al. Tumor volume definitions in head and neck squamous cell carcinoma—comparing PET/MRI and histopathology. *Radiother Oncol*. 2023;180:109484. doi:10.1016/J.RADONC.2023.109484

© The Author(s) 2024. Published by Oxford University Press on behalf of the British Institute of Radiology.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.
BJR|Artificial Intelligence, 2024, 1, 1–14
https://doi.org/10.1093/bjrai/ubae015
Review