A review of the metrics used to assess auto-contouring systems in radiotherapy

Authors:

Katherine Mackay [a, b]

David Bernstein [a, b]

Ben Glocker [c]

Konstantinos Kamnitsas [c, d]

Alexandra Taylor [a, b]

a: The Institute of Cancer Research, 123 Old Brompton Road, London SW7 3RP, United Kingdom
b: The Royal Marsden Hospital, 203 Fulham Road, London, SW3 6JJ, United Kingdom
c: Department of Computing, Imperial College London, South Kensington Campus, London, SW7 2AZ, United Kingdom
d: Department of Engineering Science, University of Oxford, Parks Road, Oxford, OX1 3PJ, United Kingdom

Corresponding author:
Katherine Mackay
Katherine.mackay@icr.ac.uk
K Mackay, The Institute of Cancer Research, 123 Old Brompton Road, London SW7 3RP, United Kingdom

Acknowledgements:
This publication represents independent research supported by the National Institute for Health and Care Research (NIHR) Biomedical Research Centre at The Royal Marsden NHS Foundation Trust and the Institute of Cancer Research, London. This research was funded by the Lady Garden Foundation. The views expressed are those of the author(s) and not necessarily those of the NHS, NIHR, the Department of Health and Social Care or the Lady Garden Foundation.

Declarations of interest:
Alexandra Taylor: MSD Advisory Board

Author Contributions:

Katherine Mackay 1, 2, 3, 5, 6, 7, 8

David Bernstein 2, 3, 7, 8

Ben Glocker 2, 7, 8

Konstantinos Kamnitsas 2, 7, 8

Alexandra Taylor 2, 3, 7, 8

1 guarantor of integrity of the entire study
2 study concepts and design
3 literature research
4 clinical studies- N/A
5 experimental studies / data analysis
6 statistical analysis
7 manuscript preparation
8 manuscript editing

Abstract

Introduction:
Auto-contouring could revolutionise future planning of radiotherapy treatment. Lack of consensus on how to assess and validate auto-contouring systems currently limits clinical use. This review formally quantifies the assessment metrics used in studies published during one calendar year and assesses the need for standardised practice.

Methods:
A PubMed literature search was undertaken for papers evaluating radiotherapy auto-contouring published during 2021. Papers were assessed for types of metric and the methodology used to generate ground-truth comparators.

Results:
Our PubMed search identified 212 studies, of which 117 met the criteria for clinical review. Geometric assessment metrics were used in 116 of 117 studies (99.1%). This includes the Dice Similarity Coefficient used in 113 (96.6%) studies. Clinically relevant metrics, such as qualitative, dosimetric and time-saving metrics were less frequently used in 22 (18.8%), 27 (23.1%) and 18 (15.4%) of 117 studies respectively. There was heterogeneity within each category of metric. Over 90 different names for geometric measures were used. Methods for qualitative assessment were different in all but two papers. Variation existed in the methods used to generate radiotherapy plans for dosimetric assessment. Consideration of editing time was only given in 11 (9.4%) papers. A single manual contour as a ground-truth comparator was used in 65 (55.6%) studies. Only 31 (26.5%) studies compared auto-contours to usual inter- and/or intra-observer variation.

Conclusions:
Significant variation exists in how research papers currently assess the accuracy of automatically generated contours. Geometric measures are the most popular, however their clinical utility is unknown. There is heterogeneity in the methods used to perform clinical assessment. Considering the different stages of system implementation may provide a framework to decide the most appropriate metrics. This analysis supports the need for a consensus on the clinical implementation of auto-contouring.

Introduction

Auto-contouring of radiotherapy target volumes and organs-at-risk using artificial intelligence (AI) could revolutionise radiotherapy treatments [1]. Before use in clinical practice, the performance of an auto-contouring system should be evaluated, to confirm efficacy on local clinical data [2-4]. Any contours produced by an AI algorithm should be checked by a clinician prior to clinical use and the validation process should not become susceptible to automation bias [5, 6].

Despite exponentially increasing research into the field, auto-contouring software is not yet widely used in clinical practice. A 2020 survey of medical physicists reported that significant barriers to the use of AI included a lack of information or knowledge on how to implement AI into the clinical workflow and a lack of resources [7]. Limited guidance on how auto-contouring systems should be validated may be a contributory factor to the delay in their widespread adoption.

Auto-contouring Evaluation Metrics

The goals of auto-contouring include reducing contouring time, reducing inter-observer variability and improving dose consistency and accuracy [8]. A wide range of metrics can be used to assess the quality of automatically generated contours [8-12]. Four categories of metric can be used; geometric, dosimetric, time-based and qualitative (table 1) [8].

Ground-truth

Most metrics rely on the comparison of an auto-contour to a “ground-truth” in order to ascertain whether the system performs accurate segmentation. Unfortunately, a biologically perfect ground-truth contour is impossible to produce as it is defined on medical imaging which has contrast and resolution limitations. Instead, either a single manual contour or multiple manual contours have been used as reference or proxy of the underlying ground-truth.

Multiple contours are used in ground-truth generation to account for variation in manual contouring [13]. Variation may exist between different operators (inter-observer variation) and at different times (intra-observer variation). Since intra- and inter-observer variation exist, there may be a range to what constitutes an acceptable ground-truth.

To minimise the impact of this variation, a single contour can be peer reviewed by one or more clinicians, or it can be generated by a group of clinicians (consensus contour). A popular method for creating a ground-truth contour is the Simultaneous Truth and Performance Level Estimation (STAPLE) [14]. This uses an algorithm to create a probabilistic estimate of the true contour, using the input of multiple contours. This creates a statistical ground truth from input data, but importantly does not consider the underlying spatial context or protocol for contouring. The ability of the STAPLE to normalise contour variation also depends on the number of input contours used, with smaller numbers being less able to appropriately establish a consensus contour.

Auto-contouring implementation

The range in metrics reflects the variety of approaches used to evaluate auto-contouring systems. Robert *et al* recently studied the implementation of auto-contouring models in three French centres which used a wide range of assessment metrics and ground-truths [15]. In total, seven different metrics were assessed, with only one metric (the Volumetric Dice Similarity Coefficient) used by all three centres.

Guidance from the 3rd ESTRO physics workshop on the implementation of AI techniques recommends using a combination of qualitative and quantitative evaluation metrics and aiming for accuracy comparable to usual inter- or intra-observer variability [4]. However, this 2019 guidance does not specify exactly which metrics should be chosen.

Auto-contouring research currently exists on a spectrum, from computer science-based method development to the clinical implementation of auto-contouring systems.. Gooding *et al* acknowledge that the assessment metrics used should relate to the overall study objective [12]. They propose that for computer science development studies, quantitative (geometric) measures are the most appropriate evaluation metrics. In contrast, methods evaluating clinical impact are needed for clinical commissioning studies. This is of inherent importance as there is evidence that geometric assessment metrics do not correlate with clinical or dosimetric acceptability of contours [16-18].

This systematic review was undertaken to identify which auto-contouring evaluation metrics were used in literature published in 2021. The aim of this review was to assess current practice and identify whether there is a need for a standardised framework in the evaluation of auto-contouring tools in research and clinical practice.

Methods

A literature search was carried out using PubMed as a search engine with the following search terms: 'radiotherapy' and 'auto- contouring' or 'auto-contouring' or 'autocontouring' or 'auto contouring' or 'automatic contouring' or 'auto-segmentation' or 'auto-segmentation' or 'autosegmentation' or 'auto segmentation' or 'automatic segmentation' or 'auto- delineation' or 'auto-delineation' or 'autodelineation' or 'auto delineation' or 'automatic delineation'.

All papers published between 1st January 2021 and 31st December 2021 were reviewed. Papers evaluating target or organ-at-risk auto-contours for radiotherapy in humans were included. Papers not published in the English language were excluded. Other exclusion criteria were: review articles, case reports, studies assessing only automated planning or quality assurance and studies using delineation not for the purpose of radiotherapy (e.g. for diagnostic purposes, radiomics, prediction of recurrence and automatic detection of needles in brachytherapy). Studies that used auto-contours but did not evaluate them were also excluded.

All papers were reviewed by 1 reviewer. Any papers where it was not certain if they met inclusion criteria were reviewed by 2 additional reviewers.

For each paper, data was collected regarding whether the study was using a newly developed model, the tumour type and the type of structures contoured. The method used to perform auto-contouring was recorded, based on the classifications set out by Harrison *et al* [1]. These are intensity analysis and shape modelling, atlas-based, non-deep machine learning and deep learning. The assessment metrics, the method for generation of a ground-truth contour and whether there was any comparison to inter- or intra-observer variation were also recorded.

Results

There were 212 studies identified by the PubMed search of which 95 were excluded. Common reasons for exclusion included review articles and studies using auto-contouring not for the purpose of radiotherapy (see figure 1). In total, 117 papers were reviewed, with 91 publications assessing a newly developed auto-contouring model. The other 26 papers evaluated a previously published or commercially available model, often with a clinical focus.

Demographics

The majority of papers (89/117) evaluated auto-contouring models built with deep learning architecture (76.1%). An additional 11 studies (9.4%) compared deep learning models to other types of models (atlas-based or intensity analysis/ shape modelling). Thirteen studies (11.1%) analysed atlas-based models and 4 (3.4%) studies used intensity analysis and shape modelling as their auto-contouring method.

The most commonly investigated tumour site was head and neck (30.8% papers), followed by breast (14.5%), lung (13.7%), prostate (12.0%) and brain (11.1%). 40.2% studies analysed auto-contours for target structures, 42.7% studies analysed contours for organs-at-risk and 16.2% looked at both (not specified in 0.9%).

Overall

The most frequently used type of assessment metric was geometric assessment metrics, being used in 99.1% of studies. The percentage of studies using each category of assessment metrics is summarised in Figure 2.

Geometric

The different geometric metrics and the number of studies they were used in is set out in table 2. All of the 91 studies presenting a new auto-contouring model reported geometric evaluation metrics compared with 96.2% (25/26) of studies using previously published or commercial models. The median number of geometric metrics used in the first group was 3 (range 1-9), compared to 2.5 (range 0-23) in the second group. Two studies in the latter group used 23 and 18 metrics respectively to ascertain if any geometric metrics correlate with dosimetry [18, 19]. If these studies were excluded, the range of geometric metrics used would be 0-5 (median 2).

Of all 117 studies, 115 (98.3%) published at least one overlap metric. Variations of each overlap metric existed and these are listed in Table 2. A variant of the Dice Similarity coefficient was published in 113 studies (96.6%), making it the most commonly used metric.

A surface-based metric was analysed in 90 studies (76.9%). Volume statistics were published in 25 studies (21.4%) and classification accuracy statistics in 28 (23.9%). Over 30 different classification accuracy metrics were used in total. Nine studies (7.7%) used a measure of estimated editing and 7 (6.0%) used metrics that compared the location of the centre of a structure.

Qualitative

Qualitative assessment was performed in 22/117 studies (18.8%), in 19/91 using a new model and 3/26 using a previously established model. Ten of these studies performed more than one type of qualitative test. All 22 studies used a Likert scale to give a numerical value for qualitative assessment. In 20 studies, this was based on clinical acceptability. Other scales were based on estimated helpfulness of auto-contours, estimated difference of auto-contours to manual contours and clinician satisfaction. Additional Turing tests assessing clinician contour preference or contour source were each performed in 5 studies. Some groups presented pictures of auto-contours as “qualitative” results however these were not always accompanied by an assessment.

Of the 22 studies using a Likert scale for a qualitative analysis, 11 used a 4-point scale. The number of denominators on the scale ranged from 2 to 11 (table 3). Different scales with different descriptors were used in all but two studies published by the same research group (supplementary information) [20, 21].

The number of observers used to perform qualitative assessment varied between studies. The median number of observers performing assessment was 3 and the range of observers was 1 to 39 (not specified in 1 study), however many did not score the same cases. Two studies repeated the qualitative tests a few weeks later to measure reproducibility and consistency amongst the observers [21, 22].

Time-saving

A comparison of auto-contouring time to manual contouring time was used in 18/117 (15.4%) total studies, 14/91 studies presenting a new model and 4/26 studies using a previously presented model. Of these studies, 7 of the former and all of the latter accounted for how long it would take a clinician to check and edit auto-contours when calculating an overall time-saving benefit. Time-processing statistics for the auto-contouring model were more frequently reported in 32/117 (27.4%) studies; however, some of these studies did not consider how long it would take to perform manual contouring.

Dosimetric

The dosimetric impact of using auto-contours was measured in 23.1% (27/117) total studies including 15/91 new model studies and 12/26 studies using a previously published model. In 16 papers, radiotherapy plans were generated using manual contours or manual beam selection (for tangential field breast radiotherapy). Auto-contours were then transposed onto these plans to assess the dose that would be received by these structures. Nine studies created new radiotherapy plans based on the automated contours and compared these to different plans generated from manual contours. One study performed planning just using auto-contours while the contour source used to generate a plan was not specified in 1 study.

Nine studies assessed the impact on target volume coverage by using auto-contours and 24 studies assessed the impact on organ-at-risk coverage. Each study reported the differences in dose for either Dmean, Dmax or other important dose constraints for each structure generated using auto-contours and manual contours. Some studies performed additional 3D gamma analysis [16, 23] and calculated the homogeneity index and conformity index for plans [16, 24].

Ground-Truth

The most common method of creating a ground-truth for comparison against auto-contours was a single manual contour, used in 55.6% (65/117) studies. This was either the clinical contour used for treatment or a contour drawn by one clinician. A peer-reviewed contour was used in 26.5% (31/117) cases, a consensus contour in 12.8% (15/117) of cases while a STAPLE contour was only used in two studies (1.7%).

Multiple clinical contours, either drawn by the same or different clinicians were used to perform all evaluations of auto-contours in 3.4% (4/117) of studies. 27/117 (23.1%) analysed multiple contours on a subset of cases, meaning 26.5% (31/117) of all studies considered inter- and/ or intra-observer variation in some way. The full breakdown of methods used to generate a ground-truth for each paper is included in the supplementary information.

Discussion

This review highlights that a wide range of assessment metrics were used to evaluate auto-contours in literature published in 2021. The use of AI in radiotherapy planning has been exponentially increasing since 2012 [25]. It is for this reason that the most recent calendar year was chosen for this review. The heterogeneity demonstrated in literature from 2021 is likely to be present in 2022 and beyond without specific guidance. Geometric measures were clearly the most popular assessment metrics, being used in 99.1% of studies. Qualitative, dosimetric and time-saving assessment metrics were less popular, being used in 18.8%, 23.1% and 15.4% studies respectively. Variation existed in the methodology for each metric category and only 26.5% of studies formally assessed auto-contours in the context of intra- and inter-observer variation. The Dice Similarity Coefficient was used in 96.6% studies, making it the most “standardised” metric to be used. Despite its popularity, the Dice Similarity Coefficient should not be presumed as the best metric. In the clinical setting, all

assessment metrics need to be considered with regards to their strengths and weaknesses, and whether they demonstrate that the aims of auto-contouring have been met [8].

The ultimate goal for assessment in AI is to demonstrate similarity to human-level performance. The popularity of geometric assessment metrics may be related to their relative ease of calculation and that they quantify similarity to a manual contour [1, 12]. This should enable direct comparisons between studies. This review however detected that over 90 different names were used to describe geometric metrics, with multiple names sometimes used to describe the same or a very similar metric. The formulae used to create these metrics were sometimes not reported or were different. Variations can also exist for each metric. This is clearly demonstrated by the Hausdorff distance, which has been reported as the maximum, minimum, median, mean, 80th percentile, 90th percentile or 95th percentile Hausdorff distance. This choice of metrics could potentially result in bias, with studies choosing to report only their most favourable results. Having such variation makes it difficult to compare between studies or establish what should be deemed an acceptable result.

Using numbers can be helpful when comparing different iterations of a model during the development stage [12]. However, if being used to clinically evaluate a model, these numbers should be interpreted with caution since it is difficult to attribute clinical meaning. Some groups have attempted this; for example Zhang *et al* used a Dice Similarity Coefficient value of 0.8 and a mean distance to agreement of less than 2 mm to define acceptable slices [26]. This was based upon image registration and fusion algorithm recommendations by the American Association of Physicists in Medicine (AAPM) task group, but ignores the AAPM recommendation to consider the size of the structure if using Dice [27]. The Dice Similarity Coefficient is influenced by the overall size of a structure and this means a threshold to predict clinical acceptability would change for each structure. Several studies have also now reported that geometric auto-contour evaluation metrics do not predict whether an auto-contouring system produces clinically useful contours [16-18, 23, 28], and geometric thresholds should therefore be used with caution when clinically commissioning a system.

When considering clinically relevant auto-contour assessment, qualitative metrics emulate real-world clinical practice in which the decision about whether a contour is acceptable for clinical use is ultimately a subjective process. Despite this, these metrics are less popular than geometric metrics and there is significant heterogeneity, with all but two studies (from the same research group) using a different method.

The most common method of qualitative assessment was using a scale to assess clinical acceptability. This generally quantified the amount of editing required to make an auto-contour safe to use. Interestingly, papers do not attempt to report “accuracy” or “time-saving potential”, which are ultimately what a utility assessment is trying to demonstrate. The qualitative scales in use currently are varied in terms of numbers and instructions. Limited instructions may lead to inconsistencies in operator interpretation [12]. A lack of consistency in scoring between observers was clearly demonstrated in 5 studies [20, 21, 29-31]. Interestingly, consistency did not improve by increasing the numbers of observers from 2 to 9 [20, 21, 29]. Ying *et al* introduced a new scale whereby a score was allocated based on the proportion of slices on which a contour needed to be edited [32]. Different

proportions were recommended based on the size of the structure. Unfortunately, only one observer was used in this study, so there was no consistency assessment. Cardenas *et al* recognised that different clinicians have their own contouring styles. In their trial, the best qualitative scores for auto-contours were given out by the clinician who had performed the manual delineation to train the auto-contouring model [30]. It therefore may be difficult to generate consistency in qualitive scoring whilst there is disagreement in how manual contouring should be performed [33].

There is potential to standardise and improve the utility of qualitative scoring. Using a 5-point scoring system has been suggested, in line with scoring adverse events in clinical trials [8, 34, 35]. Performing a blinded Turing Test also introduces a manual contour as a control [36]. Finally, adjusting the scoring system to allocate scores based on the relative importance of different structures within different radiotherapy treatment protocols may improve clinical utility of qualitative scoring. Poel *et al* demonstrated that the direction of contour variation with respect to target volume location has a strong correlation with the effect on dose [18]. Vaassen *et al* concluded that delineation errors of the heart only need to be corrected if they overlap with the planning target volume [37]. Target structures, serial organs-at-risk and parallel organs-at-risk may have different editing priorities and scoring systems could be re-designed to reflect this.

This review reveals that heterogeneity also exists for time-saving and dosimetric assessment metrics. Consideration was given to editing time in only 11/32 studies reporting a time-based metric. Auto-contouring is not clinically helpful if it takes longer to check and correct an auto-contour than to contour manually. Studies just reporting processing time may therefore not capture whether an actual time-saving benefit is achieved. Similarly, for dosimetric assessment, the method most frequently used was transposing auto-contours onto manual contour-based plans to compare dose to different structures. This does not reveal if a plan made using auto-contours would directly produce a dosimetrically safe and effective radiotherapy plan.

Most studies used a single clinical contour as a ground-truth. Although incorporating an inter- or intra-observer assessment is best practice, only 26.5% studies used multiple contours at some point in their analysis. Inter-observer studies require significant resource and may not be feasible for each department wishing to implement auto-contouring locally. Some studies have attempted to create inter-observer surrogates by producing an auto-contour uncertainty range [38]; producing auto-contour inner and outer boundaries [39]; by shifting manual contours by 5 mm [40] or by applying a tolerance factor (e.g. 3mm) to predict editing [41, 42]. However, in reality, tolerance for editing is likely to vary at different points within a structure and using a fixed distance may not predict clinically relevant edits.

It was not always possible to segregate papers into the discrete categories of computer-science method development and clinical studies due to frequent overlap: for example, a study publishing a new model could attempt to validate it for clinical practice. There was a distinction between studies presenting a system for the first time, and those presenting a previously published or commercially available system, with dosimetric and editing-time assessment metrics used relatively more frequently in the latter category

The key factor in deciding the most appropriate tests to use when evaluating an auto-contouring system should be the aim of the study. We propose a possible approach as presented in figure 3. When an AI model is being trained and compared with other models, geometric metrics may have a role in identifying models that have potential clinical use. When a model is being clinically commissioned, a clinically relevant comparison to usual practice is needed. For the ongoing quality assurance, qualitative checks will be performed by a clinician on each case. If qualitative assessment methods can be improved, this could support standardised quality assurance processes and provide the much-needed guidance for clinicians to safely use auto-contours.

Conclusions

There is currently significant variation in how auto-contouring systems are assessed which makes education, research, training and clinical implementation challenging. There is a lack of consensus over which metrics are most clinically useful and how ground-truth comparators can be created in a straightforward, reproducible way.

Auto-contouring is anticipated to be used widely in the future and systems need to be appropriately validated to ensure they are safe to use in clinical practice [3, 43] Prior to clinical use, a clinically relevant assessment should be performed, ideally assessing the dosimetric impact resulting from the use of any uncorrected auto-contours and/or an assessment of the time required to correct auto-contours in comparison to drawing them from scratch.

There is a clear need to develop standardised approaches for the validation of auto-contouring systems and this should become a priority for the auto-contouring research community.

Figures

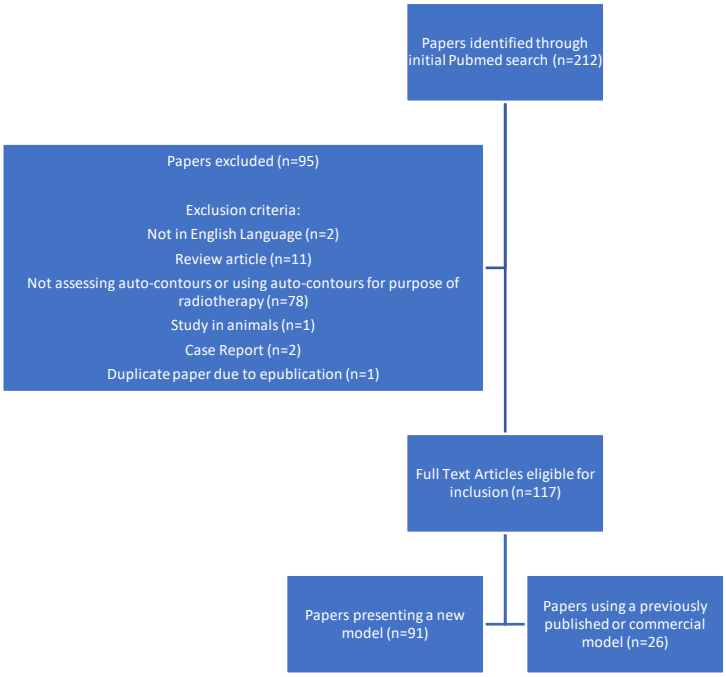


Figure 1: Flow diagram to summarise the literature search process

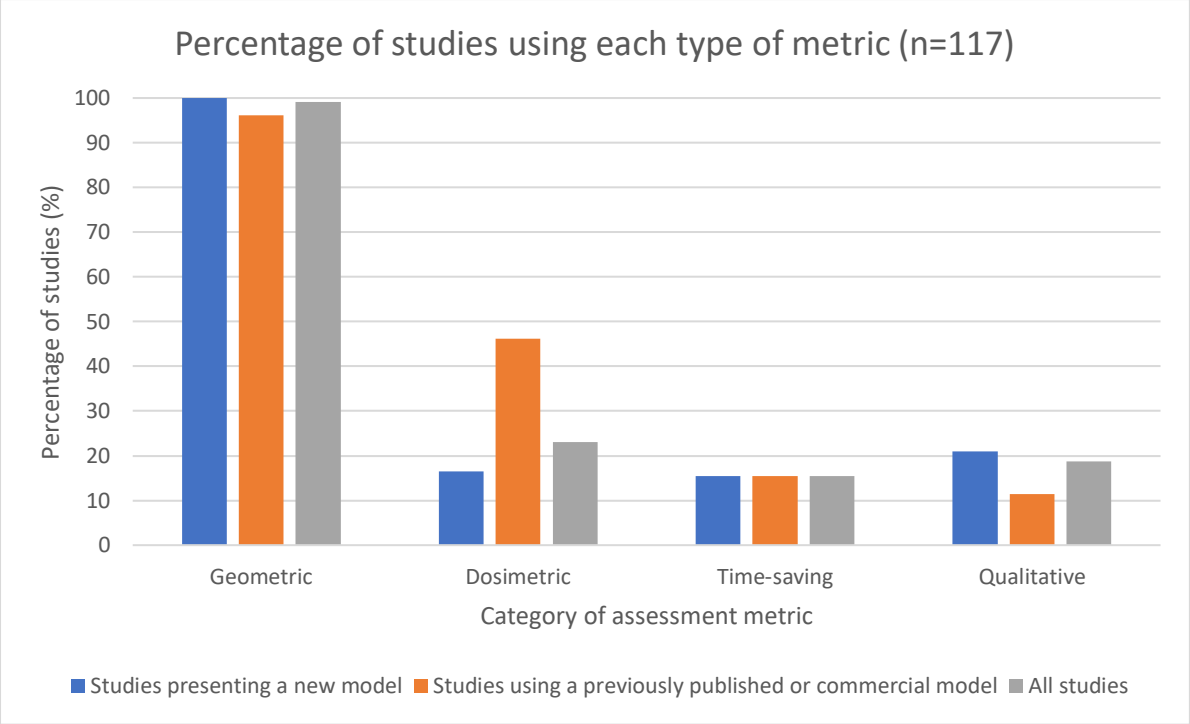


Figure 2: A bar chart showing the percentage of studies using each category of assessment metric

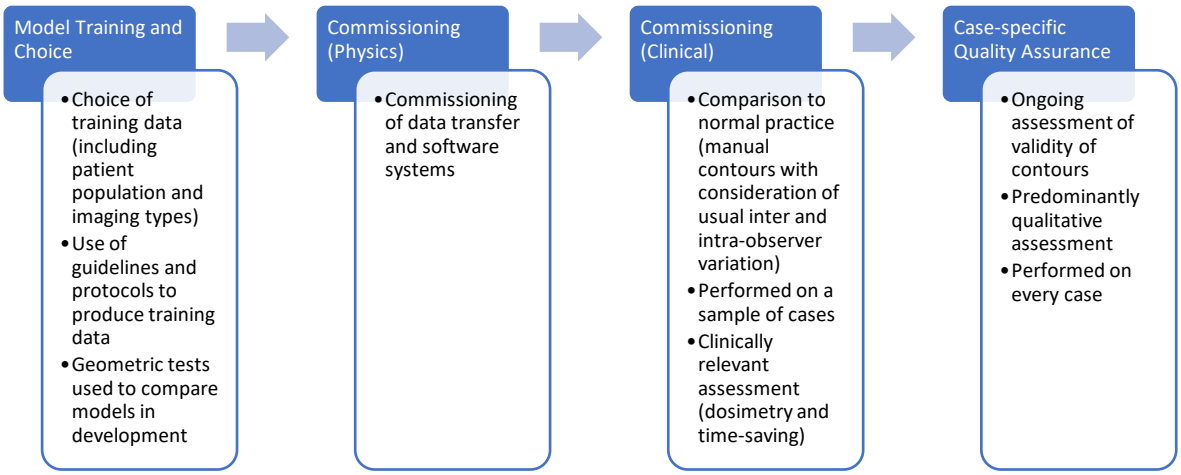


Figure 3: A flowchart describing the proposed stages of auto-contouring implementation and the types of evaluation required at each stage

Tables

Table 1: The types of assessment metric currently available to evaluate auto-contours as discussed in review articles [8-12]. Note some metrics such as overlap metrics fit into more than one category however for simplicity they are just discussed in the most relevant category.
Key: Blue contour=auto-contour and red contour= ground-truth contour

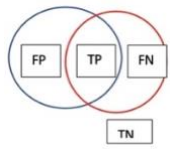
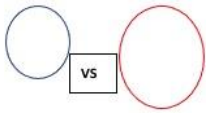
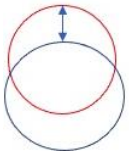
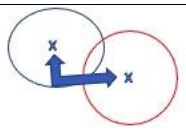
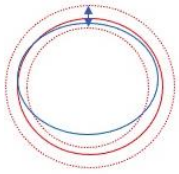
Type of Assessment Metric	**Subtype of Assessment Metric and Explanations**					
Geometric Metrics Compare an auto-contour to a "ground-truth contour	*Classification Accuracy*	*Overlap Based*	*Volume Based*	*Surface Based*	*Moment Based*	*Measures of Estimated Editing [41, 42]*
	Assesses if voxels within and outside the auto-contour have been correctly labelled.	Assesses the overlap between an auto-contour and a reference structure using different formulae	Compares the volume of an auto-contour to the volume of a manually labelled structure	Compares the distance between two structure surfaces (either maximum distance, average distance or distance at a set percentile of ordered distances)	Compares the location of an auto-contour structure centre to a reference structure centre in the x, y and z dimensions.	Compares the anticipated amount of editing a structure may require by incorporating a tolerance parameter to the reference structure. (n.b. tolerance defined by blue arrow and dotted lines around ground-truth contour below)
						
Example formulae	Sensitivity= (TP/(TP+FN)	Dice Similarity Coefficient = $2(V_A \cap V_M)/(V_A + V_M)$	Volume ratio= Volume A ÷ Volume M	Hausdorff Distance= (max*d*(a,M), max*d*(A,m))	Distance from centre in x, y and z dimensions	Surface Dice= $\lvert V_A \cap V_M^{\,t} \rvert + \lvert V_M \cap V_A^{\,t} \rvert / \lvert V_A \rvert + \lvert V_M \rvert$
Dosimetric Metrics	Compares radiotherapy plans generated for: • Manual Contours, un-edited auto- contours and edited auto-contours. Up to 9 sets of Dose Volume Histograms (DVH) can be compared, by transposing each set of contours onto each plan. Dose Constraints (e.g. Dmean, Dmax, V20) can be compared for each type of structure or plan.					
Time-saving Metrics	Compares how long it takes to contour structures manually and how long it takes to produce, check and edit auto-contours.					
Qualitative Assessment Metrics	*Scale Scoring*			*Blinded Tests (Turing Tests) [36]*		
	Uses a Likert scale (scale with varying descriptors) to grade the quality of auto-contours. The number of descriptive points on the scale can vary (commonly between 2 and 10). Can be used to assess auto-contours alone or auto-contours and manual contours. Can be performed for whole structures or single slices.			Clinicians are blinded and shown a sample of manual contours and automatic contours for the same cases. Clinicians are asked to identify if: • The contour has been generated by a human or computer • Which contour is better If the auto-contours are comparable to manual contours, they will be indistinguishable from manual contours and the correct source will only be identified 50% of the time.		

Table 2: A table demonstrating the frequency of geometric assessment metrics used

Metric Used	**Variations on Metric Used**	**Frequency (number of studies and percentage)**			References
		Studies publishing a new model (n=91)	**Studies using a prior model (n=26)**	**Total (n=117)**	
Overlap- based metrics					
Dice Similarity Coefficient	Dice per Case, Dice Global Score, Percentage Dice Similarity Coefficient 2D Dice, 3D Dice, Volumetric Dice Similarity Coefficient	90	23	113 (96.6%)	New Model: [20-24, 29-33, 38, 44-122] Prior Model: [16, 18, 19, 35, 37, 39, 40, 123-138]
Jaccard Index	Jaccard Similarity Coefficient, Concordance Index, Intersection over Union- (mean and frequency weighted), Jaccard Conformity Index (for multiple analyses)	8	2	10 (8.5%)	New Model: [57, 68, 71, 75, 83, 84, 116, 139] Prior Model: [16, 18]
Overlap Index	Sensitivity Index	2	1	3 (2.6%)	New Model: [74, 77] Prior Model: [136]
Jaccard Distance		2	0	2 (1.7%)	New Model: [77, 121] Prior Model:
Simpson's Coefficient	Overlap coefficient	0	1	1 (0.9%)	Prior Model: [140]
Inclusiveness Index	Coverage Fraction	2	0	2 (1.7%)	New Model: [77] [97]
Surface- based metrics					
Average Surface Distance	Mean Surface Distance, Median Surface Distance, Average Symmetric Surface Distance, Mean absolute surface to surface distance, Symmetric Mean Boundary Distance, Mean Contour Distance, Residual mean square distance, Minimum Average Distance, Median Hausdorff, Average (mean Hausdorff), Mean slice-wise Hausdorff Distance	38	8	46 (39.3%)	New Model: [30, 32, 38, 45, 50, 53, 54, 56, 61-64, 67, 71, 72, 75-77, 79-82, 85, 90-92, 94, 96-98, 102-106, 109, 110, 113] Prior Model: [18, 19, 37, 39, 40, 127, 135, 137]
Hausdorff Distance	Maximum Hausdorff, 2D Hausdorff, 3D Hausdorff	28	7	35 (29.9%)	New Model: [23, 30, 32, 44, 46, 48, 50, 56, 57, 60, 68, 73-75, 77, 78, 80, 82, 86, 88, 92, 98, 101, 103, 108, 110, 115, 121] Prior Model: [16, 18, 39, 125, 127, 133, 137]
Percentile Hausdorff Distance	80% Hausdorff, 90% Hausdorff, 95% Hausdorff, 90th percentile symmetric surface distance	37	3	40 (34.2%)	New Model: [20-22, 29, 31, 48, 49, 52-54, 58, 61-64, 67, 69, 71, 76, 79, 81, 85, 87, 92-94, 97, 98, 100, 104, 105, 110-112, 118, 120, 122] Prior Model: [19, 132, 134]
Distance to Agreement	Mean distance to agreement	4	4	8 (6.8%)	New Model: [23, 60, 65, 68] Prior Model: [16, 124, 131, 133]

Other	Maximum Diameter Difference, Shortest Distance to ITV, Landmark point difference, Manhabolis Distance, Probabalistic Distance, ComGrad Distance, Global consistency error, Variation of information	3	2	5 (4.3%)	New Model: [60, 107, 139] Prior Model: [18, 130]
Volume-based metrics					
Volume Difference	Volume difference, Relative volume, Percentage difference in volume, Change in volume within a defined area, volume scatter plots	17	8	25 (21.4%)	New Model: [32, 33, 50, 60, 61, 66-69, 73, 74, 77, 94, 98, 103, 107, 121] Prior Model: [19, 123, 126, 128, 131, 136, 138, 141]
Classification Accuracy metrics					
Sensitivity and Specificity	ROC curve, AUC (of ROC curve), Structure-wise sensitivity, Voxel wise recall rate, Recall	15	3	18 (15.4%)	New Model: [47, 63, 64, 69, 70, 72, 83, 93, 97, 99, 100, 104, 110, 115, 116] Prior Model: [18, 19, 129]
Precision	PR curve (Precision- recall curve), Positive Predictive Value	14	1	15 (12.8%)	New Model: [47, 63, 69, 70, 83, 97-100, 104, 110, 114-116] Prior Model: [18]
False Positive or False Negative	False Positive Dice, False Negative Dice	5	1	6 (5.1%)	New Model: [30, 69, 81, 97, 110] Prior Model: [19]
Accuracy		3	1	4 (3.4%)	New Model: [64, 93, 115] Prior Model: [18]
F1 measure		2	1	3 (2.6%)	New Model: [83, 84] Prior Model: [18]
Cohen Kappa coefficient		1	1	2 (1.7%)	New Model: [46] Prior Model: [18]
Other	True positive Volume Fraction, Volumetric Overlap Error, Relative Volume Error, False Detection Rate, Root Mean Square Error, Sensitivity/ Specificity Ratio, True Volume, False Volume, C factor, Deviance, Fallout, Rand index, Adjusted Rand index, Interclass correlation, Mutual information, Jacobian minimum and maximum, Matthews correlation coefficient, mean pixel accuracy	4	4	8 (6.8%)	New Model: [47, 84, 86, 90] Prior Model: [18, 19, 131, 133]
Measures of estimated editing					
Surface Dice Coefficient	Surface Dice score at 1mm	5	3	8 (6.8%)	New Model: [33, 59, 100, 110, 122] Prior Model: [18, 19, 35]
Added Path Length		1	2	3 (2.6%)	New Model: [108] Prior Model: [19, 35]
Moment Based metrics					
Centre of mass difference	Centre of mass shift, Centroid Distance differences (in all planes)	3	4	7 (6.0%)	New Model: [32, 67, 86] Prior Model: [19, 123, 131, 132]

Table 3: A table showing the types of qualitative assessments performed on auto-contours

Number (and Percentage) of studies using each type of qualitative assessment:	
Type of qualitative assessment	Number (and percentage of studies)
Clinical Acceptability	20 (90.9%)
Preference of Contour (Turing Test)	5 (22.7%)
Source of Contour (Turing Test)	5 (22.7%)
Estimated assistance of contours	2 (9.1%)
Estimated difference to manual contours	1 (4.5%)
Satisfaction rating	1 (4.5%)
Number (and Percentage) of studies using each size of qualitative assessment scale	
Number of points on qualitative assessment scale	Number (and percentage) of studies
2	2 (9.1%)
3	5 (22.7%)
4	11 (50%)
5	1 (4.5%)
6 - 11	3 (13.6%)

Supplementary Information

Supplementary Table 1: The different types of qualitative assessment used in 2021

Study	Type of Qualitative Assessment	No. of points on scale	Scale used	Number of Assessors used	Number of scans used for assessment
Studies publishing first presentation of a model					
[24]	Clinical Acceptability	2	Auto- segmentation and auto-plan are clinically acceptable OR Auto-segmentation and auto-plan need editing	1	40
[31]	Clinical Acceptability	3	Acceptable with no corrections Acceptable with minor corrections OR Unacceptable	6 (3 groups of 2)	222
[107]	Clinical Acceptability	3	1) No editing required 2) Minor editing required 3) Major editing required and not useful in clinical practice	4	10
[30]	Clinical Acceptability	3	1) Clinically acceptable without requiring edits 2) Requiring minor edits (can be corrected within 2 minutes and/ or are acceptable to use if a CTV to PTV margin of 4mm is to be used) 3) Requiring major edits (would affect the likelihood of cure, adverse events or locoregional control)	3	32
[64]	Clinical Acceptability	3	1) Accepted as is 2) Needs manual correction 3) Failed	1	35
[38]	Clinical Acceptability AND Source of Contour	4	4) Acceptable without changes 3) Acceptable with minor changes (i.e. does not miss important pathology) 2) Acceptable with major changes (the contour needs significant revision and the treatment should not proceed without contour correction) 1) Completely unacceptable	Not specified	30
[59]	Clinical Acceptability	4	1) Requires corrections, large obvious errors 2) Requires corrections, minor errors 3) Clinically acceptable, errors not clinically significant 4) Clinically acceptable, contours are highly accurate	3	99
[29]	Clinical Acceptability	4	3) No need to be edited 2) Number of layers to be edited </=4 1) Number of layers to be edited >/=4 0) Not acceptable	2	20 for OAR 10 for CTV
[20]	Clinical Acceptability AND Preference of Contour	4	3) No revision- segmentation is perfect and completely acceptable for treatment 2) Minor revision- the segmentation needs a few minor edits but has no significant clinical impact without correction 1) Major revision- the segmentation needs significant revision. Treatment planning should not proceed without contour correction. 0) Rejection- the segmentation is unacceptable and needs to be redrawn	2	10 100 slices used for preference test
[21]	Clinical Acceptability	4	3) No revision	9	10

	AND Preference of Contour		2) Minor revision (no significant clinical impact without correction) 1) Major revision (treatment planning cannot proceed until corrected) 0) Rejection					200 slices used for preference test
[22]	Clinical Acceptability AND Preference of Contour	4	0) Rejected- the contour is unacceptable and requires re-drawing 1) Major revision- the contour requires significant revision and treatment planning should not proceed without correction 2) Minor revision- the contour should be revised with a few minor edits but has no significant effect on treatment without correction 3) The contour is perfect and completely acceptable for treatment.				10	10 contours (5 AI, 5 manual) for each of 10 patients (100 total contours) for scoring 100 slices from 10 patients for Turing Test
[92]	Clinical Acceptability	4	1) The segmentation does not need to be modified and can be used in clinical practice 2) The algorithm can be used as an auxillary contouring tool, since the segmentation result can be used in clinical practice after minor modifications 3) The algorithm can be used as an auxillary contouring tool and the segmentation result can be used in clinical practice after significant modifications 4) The algorithm has no auxillary contouring value. In addition, perceived errors in segmentation results have been identified.				9	56
[81]	Clinical Acceptability AND Source of Contour AND Preference of contour	4	1) Precise 2) Minor error without revision 3) Minor error with revision 4) Major error with revision.				26	20 cases, using 4 different methods (manual, 2xDL, 1x DIR)- 198 comparison scenarios total shown
[32]	Clinical Acceptability	4	Descriptor	>10 slices (% to be modified)	3-10 slices (slice number to be modified)	<3 slices (slice number to be modified	1	20

			1) Auto-segmentation is not recommended 2) Many manual modifications are required after auto-segmentation 3) Some manual modifications are required after auto-segmentation 4) Auto-segmentation can completely replace manual delineation	20-100% 10-20% 0-10% 0	>3 2-3 1 0	3 2 1 0		
[110]	Clinical Acceptability AND Source of Contour	4	1) Requires corrections- large errors 2) Requires corrections- minor errors 3) Clinically acceptable- errors not clinically significant 4) Clinically acceptable- highly accurate				3	30
[120]	Clinical Acceptability AND Source of Contour AND Preference of Contour	4	For all contours (blinded manual and automatic): Would you a) require it to be corrected- there are large, obvious errors b) require it to be corrected, there are minor errors c) accept it as it is, but it needs a small amount of editing d) accept it as it is, the contour is very precise. Which contour do you prefer? Results classified as 1) strong tendency to manual 2) more inclined to manual 3) no tendency 4) more inclined to autosegmentation 5) strong inclination to autosegmentation				2-3	30
[65]	Clinical Acceptability	7	1= Good agreement 5= moderate manual edits needed in 20-50% of slices to be clinically acceptable 7= Gross error n.b </=5 determined as clinically acceptable.				3	28
[97]	Estimated helpfulness AND Source of contour	10	1= delineation with little to no clinical value 10= unable to identify whether CNN or human (implying high value and indistinguishable from manual delineation)				1	15
[52]	Difference between contours AND	11	What score would you give for the differences between manually delineated contours and auto-segmented contours? (0= most different, 10= least different)				26	19

	Estimated assistance of contours		How much do you think auto-segmentation would assist you in real-world clinical practice? (0= not helpful, 10= very helpful)		
Studies presenting a previously published model					
[130]	Clinical Acceptability	2	1) Meet 2) Fail- contour has to be corrected	2	10 cases- 97 sets of CBCT
[35]	Clinical Acceptability	3	1) Acceptable without edits 2) Need for minor edits 3) Major edits	18	43
[134]	Clinical Acceptability AND Satisfaction rating	5	Editing rating 1= minimal editing 5= significant editing. Overall satisfaction rating 1= minimal 5= significant	39	174

Supplementary Table 2: A table showing the different methods to generate a ground-truth to compare auto-contours to

Method of Ground Truth	**Use of additional inter- or intra-observer studies**	**Total number of studies**	**References**
STAPLE	No additional interobserver/ intraobserver study	0	
	Intra-observer study only	0	
	Inter-observer study only	2	New Model: [68, 86]
	Intra- and Interobserver study	0	
Consensus contour	No additional interobserver/ intraobserver study	7	New Model: [31, 54, 57, 63, 84, 97, 104]
	Intra-observer study only	0	
	Inter-observer study only	6	New Model: [100, 107] Prior Model: [18, 39, 125, 142]
	Intra- and Interobserver study	2	New Model: [93] Prior Model: [129]
Peer reviewed contour	No additional interobserver/ intraobserver study	26	New Model: [20, 21, 29, 30, 46, 47, 49, 58, 59, 67, 74, 75, 77, 89, 91, 94, 95, 102, 103, 105, 112, 113, 122] Prior Model: [16, 130, 140]
	Intra-observer study only	0	
	Inter-observer study only	4	New Model: [23, 66, 92, 106]
	Intra- and Interobserver study	1	Prior Model: [40]
Multiple manual contours	No additional interobserver/ intraobserver study	0	
	Intra-observer study only	0	
	Inter-observer study only	4	New Model: [96, 139] [33]* Prior Model: [128] *peer reviewed
	Intra- and Interobserver study	0	
Single manual contour	No additional interobserver/ intraobserver study	53	New Model: [22, 24, 32, 44, 45, 48, 51, 53, 55, 61, 62, 64, 65, 69-73, 76, 78, 79, 81, 83, 87, 88, 90, 98, 99, 101, 108-111, 114, 116-121] Prior Model: [19, 35, 123, 124, 126, 127, 131-134, 136, 137, 141]
	Intra-observer study only	2	New Model: [50] Prior Model: [37]
	Inter-observer study only	8	New Model: [38, 52, 60, 80, 82, 85] Prior Model: [135, 138]
	Intra- and Interobserver study	2	New Model: [56, 115]

References

[1] Harrison K, Pullen H, Welsh C, Oktay O, Alvarez-Valle J, Jena R. Machine Learning for Auto-Segmentation in Radiotherapy Planning. Clin Oncol (R Coll Radiol). 2022;34:74-88. 10.1016/j.clon.2021.12.003.
[2] Valentini V, Boldrini L, Damiani A, Muren LP. Recommendations on how to establish evidence from auto-segmentation software in radiotherapy. Radiother Oncol. 2014;112:317-20. 10.1016/j.radonc.2014.09.014.
[3] Chen M, Wu S, Zhao W, Zhou Y, Zhou Y, Wang G. Application of deep learning to auto-delineation of target volumes and organs at risk in radiotherapy. Cancer Radiother. 2022;26:494-501. 10.1016/j.canrad.2021.08.020.
[4] Vandewinckele L, Claessens M, Dinkla A, Brouwer C, Crijns W, Verellen D, et al. Overview of artificial intelligence-based applications in radiotherapy: Recommendations for implementation and quality assurance. Radiother Oncol. 2020;153:55-66. 10.1016/j.radonc.2020.09.008.
[5] Radiologists RCo. Radiotherapy Target Volume Definition and Peer Review- RCR guidance, https://www.rcr.ac.uk/publication/radiotherapy-target-volume-definition-and-peer-review-second-edition-rcr-guidance; 2022 [accessed 24th November, 2022]
[6] Straw I. The automation of bias in medical Artificial Intelligence (AI): Decoding the past to create a better future. Artif Intell Med. 2020;110:101965. 10.1016/j.artmed.2020.101965.
[7] Brouwer CL, Dinkla AM, Vandewinckele L, Crijns W, Claessens M, Verellen D, et al. Machine learning applications in radiation oncology: Current use and needs to support clinical implementation. Phys Imaging Radiat Oncol. 2020;16:144-8. 10.1016/j.phro.2020.11.002.
[8] Sherer MV, Lin D, Elguindi S, Duke S, Tan LT, Cacicedo J, et al. Metrics to evaluate the performance of auto-segmentation for radiation treatment planning: A critical review. Radiother Oncol. 2021;160:185-91. 10.1016/j.radonc.2021.05.003.
[9] Fotina I, Lutgendorf-Caucig C, Stock M, Potter R, Georg D. Critical discussion of evaluation parameters for inter-observer variability in target definition for radiation therapy. Strahlenther Onkol. 2012;188:160-7. 10.1007/s00066-011-0027-6.
[10] Jameson MG, Holloway LC, Vial PJ, Vinod SK, Metcalfe PE. A review of methods of analysis in contouring studies for radiation oncology. J Med Imaging Radiat Oncol. 2010;54:401-10. 10.1111/j.1754-9485.2010.02192.x.
[11] Taha AA, Hanbury A. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. BMC Med Imaging. 2015;15:29. 10.1186/s12880-015-0068-x.
[12] Gooding MJ. On the Evaluation of Auto-Contouring in Radiotherapy. In: Yang J, Sharp GC, Gooding MJ, editors. Auto-segmentation for Radiation Oncology: State of the Art. 1st ed: Taylor and Francis Group; 2021. p. 217-52.
[13] Segedin B, Petric P. Uncertainties in target volume delineation in radiotherapy - are they relevant and what can we do about them? Radiol Oncol. 2016;50:254-62. 10.1515/raon-2016-0023.
[14] Warfield SK, Zou KH, Wells WM. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. IEEE Trans Med Imaging. 2004;23:903-21. 10.1109/TMI.2004.828354.

[15] Robert C, Munoz A, Moreau D, Mazurier J, Sidorski G, Gasnier A, et al. Clinical implementation of deep-learning based auto-contouring tools-Experience of three French radiotherapy centers. Cancer Radiother. 2021;25:607-16. 10.1016/j.canrad.2021.06.023.
[16] Guo H, Wang J, Xia X, Zhong Y, Peng J, Zhang Z, et al. The dosimetric impact of deep learning-based auto-segmentation of organs at risk on nasopharyngeal and rectal cancer. Radiat Oncol. 2021;16:113. 10.1186/s13014-021-01837-y.
[17] Kosmin M, Ledsam J, Romera-Paredes B, Mendes R, Moinuddin S, de Souza D, et al. Rapid advances in auto-segmentation of organs at risk and target volumes in head and neck cancer. Radiother Oncol. 2019;135:130-40. 10.1016/j.radonc.2019.03.004.
[18] Poel R, Rüfenacht E, Hermann E, Scheib S, Manser P, Aebersold DM, et al. The predictive value of segmentation metrics on dosimetry in organs at risk of the brain. Med Image Anal. 2021;73:102161. 10.1016/j.media.2021.102161.
[19] Thor M, Apte A, Haq R, Iyer A, LoCastro E, Deasy JO. Using Auto-Segmentation to Reduce Contouring and Dose Inconsistency in Clinical Trials: The Simulated Impact on RTOG 0617. Int J Radiat Oncol Biol Phys. 2021;109:1619-26. 10.1016/j.ijrobp.2020.11.011.
[20] Liu Z, Liu F, Chen W, Liu X, Hou X, Shen J, et al. Automatic Segmentation of Clinical Target Volumes for Post-Modified Radical Mastectomy Radiotherapy Using Convolutional Neural Networks. Front Oncol. 2020;10:581347. 10.3389/fonc.2020.581347.
[21] Liu Z, Chen W, Guan H, Zhen H, Shen J, Liu X, et al. An Adversarial Deep-Learning-Based Model for Cervical Cancer CTV Segmentation With Multicenter Blinded Randomized Controlled Validation. Front Oncol. 2021;11:702270. 10.3389/fonc.2021.702270.
[22] Wu Y, Kang K, Han C, Wang S, Chen Q, Chen Y, et al. A blind randomized validated convolutional neural network for auto-segmentation of clinical target volume in rectal cancer patients receiving neoadjuvant radiotherapy. Cancer Med. 2022;11:166-75. 10.1002/cam4.4441.
[23] Rigaud B, Anderson BM, Yu ZH, Gobeli M, Cazoulat G, Söderberg J, et al. Automatic Segmentation Using Deep Learning to Enable Online Dose Optimization During Adaptive Radiation Therapy of Cervical Cancer. Int J Radiat Oncol Biol Phys. 2021;109:1096-110. 10.1016/j.ijrobp.2020.10.038.
[24] Xia X, Wang J, Li Y, Peng J, Fan J, Zhang J, et al. An Artificial Intelligence-Based Full-Process Solution for Radiotherapy: A Proof of Concept Study on Rectal Cancer. Front Oncol. 2020;10:616721. 10.3389/fonc.2020.616721.
[25] Meyer P, Noblet V, Mazzara C, Lallement A. Survey on deep learning for radiotherapy. Comput Biol Med. 2018;98:126-46. 10.1016/j.compbiomed.2018.05.018.
[26] Zhang Y, Paulson E, Lim S, Hall WA, Ahunbay E, Mickevicius NJ, et al. A Patient-Specific Autosegmentation Strategy Using Multi-Input Deformable Image Registration for Magnetic Resonance Imaging-Guided Online Adaptive Radiation Therapy: A Feasibility Study. Adv Radiat Oncol. 2020;5:1350-8. 10.1016/j.adro.2020.04.027.
[27] Brock KK, Mutic S, McNutt TR, Li H, Kessler ML. Use of image registration and fusion algorithms and techniques in radiotherapy: Report of the AAPM Radiation Therapy Committee Task Group No. 132. Med Phys. 2017;44:e43-e76. 10.1002/mp.12256.
[28] Voet PW, Dirkx ML, Teguh DN, Hoogeman MS, Levendag PC, Heijmen BJ. Does atlas-based autosegmentation of neck levels require subsequent manual contour editing to avoid risk of severe target underdosage? A dosimetric analysis. Radiother Oncol. 2011;98:373-7. 10.1016/j.radonc.2010.11.017.
[29] Liu Z, Liu F, Chen W, Tao Y, Liu X, Zhang F, et al. Automatic Segmentation of Clinical Target Volume and Organs-at-Risk for Breast Conservative Radiotherapy Using a

Convolutional Neural Network. Cancer Manag Res. 2021;13:8209-17. 10.2147/cmar.S330249.
[30] Cardenas CE, Beadle BM, Garden AS, Skinner HD, Yang J, Rhee DJ, et al. Generating High-Quality Lymph Node Clinical Target Volumes for Head and Neck Cancer Radiation Therapy Using a Fully Automated Deep Learning-Based Approach. Int J Radiat Oncol Biol Phys. 2021;109:801-12. 10.1016/j.ijrobp.2020.10.005.
[31] Dai Z, Zhang Y, Zhu L, Tan J, Yang G, Zhang B, et al. Geometric and Dosimetric Evaluation of Deep Learning-Based Automatic Delineation on CBCT-Synthesized CT and Planning CT for Breast Cancer Adaptive Radiotherapy: A Multi-Institutional Study. Front Oncol. 2021;11:725507. 10.3389/fonc.2021.725507.
[32] Ying Y, Wang H, Chen H, Cheng J, Gu H, Shao Y, et al. A novel specific grading standard study of auto-segmentation of organs at risk in thorax: subjective-objective-combined grading standard. Biomed Eng Online. 2021;20:54. 10.1186/s12938-021-00890-8.
[33] Nikolov S, Blackwell S, Zverovitch A, Mendes R, Livne M, De Fauw J, et al. Clinically Applicable Segmentation of Head and Neck Anatomy for Radiotherapy: Deep Learning Algorithm Development and Validation Study. J Med Internet Res. 2021;23:e26151. 10.2196/26151.
[34] Ghooi RB, Bhosale N, Wadhwani R, Divate P, Divate U. Assessment and classification of protocol deviations. Perspect Clin Res. 2016;7:132-6. 10.4103/2229-3485.184817.
[35] Cha E, Elguindi S, Onochie I, Gorovets D, Deasy JO, Zelefsky M, et al. Clinical implementation of deep learning contour autosegmentation for prostate radiotherapy. Radiother Oncol. 2021;159:1-7. 10.1016/j.radonc.2021.02.040.
[36] Gooding MJ, Smith AJ, Tariq M, Aljabar P, Peressutti D, van der Stoep J, et al. Comparative evaluation of autocontouring in clinical practice: A practical method using the Turing test. Med Phys. 2018;45:5105-15. 10.1002/mp.13200.
[37] Vaassen F, Hazelaar C, Canters R, Peeters S, Petit S, van Elmpt W. The impact of organ-at-risk contour variations on automatically generated treatment plans for NSCLC. Radiother Oncol. 2021;163:136-42. 10.1016/j.radonc.2021.08.014.
[38] Balagopal A, Nguyen D, Morgan H, Weng Y, Dohopolski M, Lin MH, et al. A deep learning-based framework for segmenting invisible clinical target volumes with estimated uncertainties for post-operative prostate cancer radiotherapy. Med Image Anal. 2021;72:102101. 10.1016/j.media.2021.102101.
[39] Finnegan R, Laugaard Lorenzen E, Dowling J, Thwaites D, Delaney G, Brink C, et al. Validation of a new open-source method for automatic delineation and dose assessment of the heart and LADCA in breast radiotherapy with simultaneous uncertainty estimation. Phys Med Biol. 2021;66:035014. 10.1088/1361-6560/abcb1d.
[40] Jung JW, Mille MM, Ky B, Kenworthy W, Lee C, Yeom YS, et al. Application of an automatic segmentation method for evaluating cardiac structure doses received by breast radiotherapy patients. Phys Imaging Radiat Oncol. 2021;19:138-44. 10.1016/j.phro.2021.08.005.
[41] Nikolov S, Blackwell S, Zverovitch A, Mendes R, Livne M, De Fauw J, et al. Deep learning to achieve clinically applicable segmentation of head and neck anatomy for radiotherapy. arXiv:180904430. 2018.
[42] Vaassen F, Hazelaar C, Vaniqui A, Gooding M, van der Heyden B, Canters R, et al. Evaluation of measures for assessing time-saving of automatic organ-at-risk segmentation in radiotherapy. Phys Imaging Radiat Oncol. 2020;13:1-6. 10.1016/j.phro.2019.12.001.

[43] Cardenas CE, Yang J, Anderson BM, Court LE, Brock KB. Advances in Auto-Segmentation. Semin Radiat Oncol. 2019;29:185-97. 10.1016/j.semradonc.2019.02.001.
[44] Aoyama T, Shimizu H, Kitagawa T, Yokoi K, Koide Y, Tachibana H, et al. Comparison of atlas-based auto-segmentation accuracy for radiotherapy in prostate cancer. Phys Imaging Radiat Oncol. 2021;19:126-30. 10.1016/j.phro.2021.08.002.
[45] Brion E, Léger J, Barragán-Montero AM, Meert N, Lee JA, Macq B. Domain adversarial networks and intensity-based data augmentation for male pelvic organ segmentation in cone beam CT. Comput Biol Med. 2021;131:104269. 10.1016/j.compbiomed.2021.104269.
[46] Cao R, Pei X, Ge N, Zheng C. Clinical Target Volume Auto-Segmentation of Esophageal Cancer for Radiotherapy After Radical Surgery Based on Deep Learning. Technol Cancer Res Treat. 2021;20:15330338211034284. 10.1177/15330338211034284.
[47] Cao Y, Vassantachart A, Ye JC, Yu C, Ruan D, Sheng K, et al. Automatic detection and segmentation of multiple brain metastases on magnetic resonance image using asymmetric UNet architecture. Phys Med Biol. 2021;66:015003. 10.1088/1361-6560/abca53.
[48] Chang Y, Wang Z, Peng Z, Zhou J, Pi Y, Xu XG, et al. Clinical application and improvement of a CNN-based autosegmentation model for clinical target volumes in cervical cancer radiotherapy. J Appl Clin Med Phys. 2021;22:115-25. 10.1002/acm2.13440.
[49] Chen X, Sun S, Bai N, Han K, Liu Q, Yao S, et al. A deep learning-based auto-segmentation system for organs-at-risk on whole-body computed tomography images for radiation therapy. Radiother Oncol. 2021;160:175-84. 10.1016/j.radonc.2021.04.019.
[50] Christiansen RL, Johansen J, Zukauskaite R, Hansen CR, Bertelsen AS, Hansen O, et al. Accuracy of automatic structure propagation for daily magnetic resonance image-guided head and neck radiotherapy. Acta Oncol. 2021;60:589-97. 10.1080/0284186x.2021.1891282.
[51] Chun J, Park JC, Olberg S, Zhang Y, Nguyen D, Wang J, et al. Intentional deep overfit learning (IDOL): A novel deep learning strategy for adaptive radiation therapy. Med Phys. 2022;49:488-96. 10.1002/mp.15352.
[52] Chung SY, Chang JS, Choi MS, Chang Y, Choi BS, Chun J, et al. Clinical feasibility of deep learning-based auto-segmentation of target volumes and organs-at-risk in breast cancer patients after breast-conserving surgery. Radiat Oncol. 2021;16:44. 10.1186/s13014-021-01771-z.
[53] Dai X, Lei Y, Wang T, Dhabaan AH, McDonald M, Beitler JJ, et al. Head-and-neck organs-at-risk auto-delineation using dual pyramid networks for CBCT-guided adaptive radiotherapy. Phys Med Biol. 2021;66:045021. 10.1088/1361-6560/abd953.
[54] Dai X, Lei Y, Wang T, Zhou J, Roper J, McDonald M, et al. Automated delineation of head and neck organs at risk using synthetic MRI-aided mask scoring regional convolutional neural network. Med Phys. 2021;48:5862-73. 10.1002/mp.15146.
[55] Fang Y, Wang J, Ou X, Ying H, Hu C, Zhang Z, et al. The impact of training sample size on deep learning-based organ auto-segmentation for head-and-neck patients. Phys Med Biol. 2021;66. 10.1088/1361-6560/ac2206.
[56] Friedrich F, Hörner-Rieber J, Renkamp CK, Klüter S, Bachert P, Ladd ME, et al. Stability of conventional and machine learning-based tumor auto-segmentation techniques using undersampled dynamic radial bSSFP acquisitions on a 0.35 T hybrid MR-linac system. Med Phys. 2021;48:587-96. 10.1002/mp.14659.
[57] Gan W, Wang H, Gu H, Duan Y, Shao Y, Chen H, et al. Automatic segmentation of lung tumors on CT images based on a 2D & 3D hybrid convolutional neural network. Br J Radiol. 2021;94:20210038. 10.1259/bjr.20210038.

[58] Gao Y, Huang R, Yang Y, Zhang J, Shao K, Tao C, et al. FocusNetv2: Imbalanced large and small organ segmentation with adversarial shape constraint for head and neck CT images. Med Image Anal. 2021;67:101831. 10.1016/j.media.2020.101831.
[59] Garrett Fernandes M, Bussink J, Stam B, Wijsman R, Schinagl DAX, Monshouwer R, et al. Deep learning model for automatic contouring of cardiovascular substructures on radiotherapy planning CT images: Dosimetric validation and reader study based clinical acceptability testing. Radiother Oncol. 2021;165:52-9. 10.1016/j.radonc.2021.10.008.
[60] Ghandourh W, Dowling J, Chlap P, Oar A, Jacob S, Batumalai V, et al. Assessing tumor centrality in lung stereotactic ablative body radiotherapy (SABR): the effects of variations in bronchial tree delineation and potential for automated methods. Med Dosim. 2021;46:94-101. 10.1016/j.meddos.2020.09.004.
[61] Gonzalez Y, Shen C, Jung H, Nguyen D, Jiang SB, Albuquerque K, et al. Semi-automatic sigmoid colon segmentation in CT for radiation therapy treatment planning via an iterative 2.5-D deep learning approach. Med Image Anal. 2021;68:101896. 10.1016/j.media.2020.101896.
[62] Groendahl AR, Moe YM, Kaushal CK, Huynh BN, Rusten E, Tomic O, et al. Deep learning-based automatic delineation of anal cancer gross tumour volume: a multimodality comparison of CT, PET and MRI. Acta Oncol. 2022;61:89-96. 10.1080/0284186x.2021.1994645.
[63] Groendahl AR, Skjei Knudtsen I, Huynh BN, Mulstad M, Moe YM, Knuth F, et al. A comparison of methods for fully automatic segmentation of tumors and involved nodes in PET/CT of head and neck cancers. Phys Med Biol. 2021;66:065012. 10.1088/1361-6560/abe553.
[64] Gu H, Gan W, Zhang C, Feng A, Wang H, Huang Y, et al. A 2D-3D hybrid convolutional neural network for lung lobe auto-segmentation on standard slice thickness computed tomography of patients receiving radiotherapy. Biomed Eng Online. 2021;20:94. 10.1186/s12938-021-00932-1.
[65] Hague C, McPartlin A, Lee LW, Hughes C, Mullan D, Beasley W, et al. An evaluation of MR based deep learning auto-contouring for planning head and neck radiotherapy. Radiother Oncol. 2021;158:112-7. 10.1016/j.radonc.2021.02.018.
[66] Han X, Hong J, Reyngold M, Crane C, Cuaron J, Hajj C, et al. Deep-learning-based image registration and automatic segmentation of organs-at-risk in cone-beam CT scans from high-dose radiation treatment of pancreatic cancer. Med Phys. 2021;48:3084-95. 10.1002/mp.14906.
[67] Harms J, Lei Y, Tian S, McCall NS, Higgins KA, Bradley JD, et al. Automatic delineation of cardiac substructures using a region-based fully convolutional network. Med Phys. 2021;48:2867-76. 10.1002/mp.14810.
[68] Hearn N, Blazak J, Vivian P, Vignarajah D, Cahill K, Atwell D, et al. Prostate cancer GTV delineation with biparametric MRI and (68)Ga-PSMA-PET: comparison of expert contours and semi-automated methods. Br J Radiol. 2021;94:20201174. 10.1259/bjr.20201174.
[69] Hsu DG, Ballangrud Å, Shamseddine A, Deasy JO, Veeraraghavan H, Cervino L, et al. Automatic segmentation of brain metastases using T1 magnetic resonance and computed tomography images. Phys Med Biol. 2021;66. 10.1088/1361-6560/ac1835.
[70] Huang D, Wang M, Zhang L, Li H, Ye M, Li A. Learning rich features with hybrid loss for brain tumor segmentation. BMC Med Inform Decis Mak. 2021;21:63. 10.1186/s12911-021-01431-y.

[71] Huang S, Cheng Z, Lai L, Zheng W, He M, Li J, et al. Integrating multiple MRI sequences for pelvic organs segmentation via the attention mechanism. Med Phys. 2021;48:7930-45. 10.1002/mp.15285.
[72] Huang YJ, Dou Q, Wang ZX, Liu LZ, Jin Y, Li CF, et al. 3-D RoI-Aware U-Net for Accurate and Efficient Colorectal Tumor Segmentation. IEEE Trans Cybern. 2021;51:5397-408. 10.1109/tcyb.2020.2980145.
[73] Jiang J, Luo Y, Wang F, Fu Y, Yu H, He Y. Evaluation on Auto-segmentation of the Clinical Target Volume (CTV) for Graves' Ophthalmopathy (GO) with a Fully Convolutional Network (FCN) on CT Images. Curr Med Imaging. 2021;17:404-9. 10.2174/1573405616666200910141323.
[74] Jiang X, Wang F, Chen Y, Yan S. RefineNet-based automatic delineation of the clinical target volume and organs at risk for three-dimensional brachytherapy for cervical cancer. Ann Transl Med. 2021;9:1721. 10.21037/atm-21-4074.
[75] Jin D, Guo D, Ho TY, Harrison AP, Xiao J, Tseng CK, et al. DeepTarget: Gross tumor and clinical target volume segmentation in esophageal cancer radiotherapy. Med Image Anal. 2021;68:101909. 10.1016/j.media.2020.101909.
[76] Jin X, Thomas MA, Dise J, Kavanaugh J, Hilliard J, Zoberi I, et al. Robustness of deep learning segmentation of cardiac substructures in noncontrast computed tomography for breast cancer radiotherapy. Med Phys. 2021;48:7172-88. 10.1002/mp.15237.
[77] Ju Z, Guo W, Gu S, Zhou J, Yang W, Cong X, et al. CT based automatic clinical target volume delineation using a dense-fully connected convolution network for cervical Cancer radiation therapy. BMC Cancer. 2021;21:243. 10.1186/s12885-020-07595-6.
[78] Kano Y, Ikushima H, Sasaki M, Haga A. Automatic contour segmentation of cervical cancer using artificial intelligence. J Radiat Res. 2021;62:934-44. 10.1093/jrr/rrab070.
[79] Kazemimoghadam M, Chi W, Rahimi A, Kim N, Alluri P, Nwachukwu C, et al. Saliency-guided deep learning network for automatic tumor bed volume delineation in post-operative breast irradiation. Phys Med Biol. 2021;66. 10.1088/1361-6560/ac176d.
[80] Kieselmann JP, Fuller CD, Gurney-Champion OJ, Oelfke U. Cross-modality deep learning: Contouring of MRI data from annotated CT data only. Med Phys. 2021;48:1673-84. 10.1002/mp.14619.
[81] Kim N, Chun J, Chang JS, Lee CG, Keum KC, Kim JS. Feasibility of Continual Deep Learning-Based Segmentation for Personalized Adaptive Radiation Therapy in Head and Neck Area. Cancers (Basel). 2021;13. 10.3390/cancers13040702.
[82] Korte JC, Hardcastle N, Ng SP, Clark B, Kron T, Jackson P. Cascaded deep learning-based auto-segmentation for head and neck cancer patients: Organs at risk on T2-weighted magnetic resonance imaging. Med Phys. 2021;48:7757-72. 10.1002/mp.15290.
[83] Li CC, Wu MY, Sun YC, Chen HH, Wu HM, Fang ST, et al. Ensemble classification and segmentation for intracranial metastatic tumors on MRI images based on 2D U-nets. Sci Rep. 2021;11:20634. 10.1038/s41598-021-99984-5.
[84] Li D, Chu X, Cui Y, Zhao J, Zhang K, Yang X. Improved U-Net based on contour prediction for efficient segmentation of rectal cancer. Comput Methods Programs Biomed. 2022;213:106493. 10.1016/j.cmpb.2021.106493.
[85] Li Z, Li R, Kiser KJ, Giancardo L, Zheng WJ. Segmenting Thoracic Cavities with Neoplastic Lesions: A Head-to-head Benchmark with Fully Convolutional Neural Networks. Acm bcb. 2021;2021. 10.1145/3459930.3469564.

[86] Liang X, Bibault JE, Leroy T, Escande A, Zhao W, Chen Y, et al. Automated contour propagation of the prostate from pCT to CBCT images via deep unsupervised learning. Med Phys. 2021;48:1764-70. 10.1002/mp.14755.
[87] Lin M, Momin S, Lei Y, Wang H, Curran WJ, Liu T, et al. Fully automated segmentation of brain tumor from multiparametric MRI using 3D context deep supervised U-Net. Med Phys. 2021;48:4365-74. 10.1002/mp.15032.
[88] Liu C, Zhang X, Si W, Ni X. Multiview Self-Supervised Segmentation for OARs Delineation in Radiotherapy. Evid Based Complement Alternat Med. 2021;2021:8894222. 10.1155/2021/8894222.
[89] Liu Z, Sun C, Wang H, Li Z, Gao Y, Lei W, et al. Automatic segmentation of organs-at-risks of nasopharynx cancer and lung cancer by cross-layer attention fusion network with TELD-Loss. Med Phys. 2021;48:6987-7002. 10.1002/mp.15260.
[90] Luan S, Xue X, Ding Y, Wei W, Zhu B. Adaptive Attention Convolutional Neural Network for Liver Tumor Segmentation. Front Oncol. 2021;11:680807. 10.3389/fonc.2021.680807.
[91] Luximon DC, Abdulkadir Y, Chow PE, Morris ED, Lamb JM. Machine-assisted interpolation algorithm for semi-automated segmentation of highly deformable organs. Med Phys. 2022;49:41-51. 10.1002/mp.15351.
[92] Ma CY, Zhou JY, Xu XT, Guo J, Han MF, Gao YZ, et al. Deep learning-based auto-segmentation of clinical target volumes for radiotherapy treatment of cervical cancer. J Appl Clin Med Phys. 2022;23:e13470. 10.1002/acm2.13470.
[93] Marin T, Zhuo Y, Lahoud RM, Tian F, Ma X, Xing F, et al. Deep learning-based GTV contouring modeling inter- and intra- observer variability in sarcomas. Radiother Oncol. 2022;167:269-76. 10.1016/j.radonc.2021.09.034.
[94] Matkovic LA, Wang T, Lei Y, Akin-Akintayo OO, Abiodun Ojo OA, Akintayo AA, et al. Prostate and dominant intraprostatic lesion segmentation on PET/CT using cascaded regional-net. Phys Med Biol. 2021;66. 10.1088/1361-6560/ac3c13.
[95] Men K, Chen X, Yang B, Zhu J, Yi J, Wang S, et al. Automatic segmentation of three clinical target volumes in radiotherapy using lifelong learning. Radiother Oncol. 2021;157:1-7. 10.1016/j.radonc.2020.12.034.
[96] Milo MLH, Nyeng TB, Lorenzen EL, Hoffmann L, Møller DS, Offersen BV. Atlas-based auto-segmentation for delineating the heart and cardiac substructures in breast cancer radiation therapy. Acta Oncol. 2022;61:247-54. 10.1080/0284186x.2021.1967445.
[97] Moe YM, Groendahl AR, Tomic O, Dale E, Malinen E, Futsaether CM. Deep learning-based auto-delineation of gross tumour volumes and involved nodes in PET/CT images of head and neck cancer patients. Eur J Nucl Med Mol Imaging. 2021;48:2782-92. 10.1007/s00259-020-05125-x.
[98] Mohammadi R, Shokatian I, Salehi M, Arabi H, Shiri I, Zaidi H. Deep learning-based auto-segmentation of organs at risk in high-dose rate brachytherapy of cervical cancer. Radiother Oncol. 2021;159:231-40. 10.1016/j.radonc.2021.03.030.
[99] Naser MA, van Dijk LV, He R, Wahid KA, Fuller CD. Tumor Segmentation in Patients with Head and Neck Cancers Using Deep Learning Based-on Multi-modality PET/CT Images. Head Neck Tumor Segm (2020). 2021;12603:85-98. 10.1007/978-3-030-67194-5_10.
[100] Oreiller V, Andrearczyk V, Jreige M, Boughdad S, Elhalawani H, Castelli J, et al. Head and neck tumor segmentation in PET/CT: The HECKTOR challenge. Med Image Anal. 2022;77:102336. 10.1016/j.media.2021.102336.
[101] Oya M, Sugimoto S, Sasai K, Yokoyama K. Investigation of clinical target volume segmentation for whole breast irradiation using three-dimensional convolutional neural

networks with gradient-weighted class activation mapping. Radiol Phys Technol. 2021;14:238-47. 10.1007/s12194-021-00620-8.
[102] Pan K, Zhao L, Gu S, Tang Y, Wang J, Yu W, et al. Deep learning-based automatic delineation of the hippocampus by MRI: geometric and dosimetric evaluation. Radiat Oncol. 2021;16:12. 10.1186/s13014-020-01724-y.
[103] Qiu Q, Yang Z, Wu S, Qian D, Wei J, Gong G, et al. Automatic segmentation of hippocampus in hippocampal sparing whole brain radiotherapy: A multitask edge-aware learning. Med Phys. 2021;48:1771-80. 10.1002/mp.14760.
[104] Ren J, Eriksen JG, Nijkamp J, Korreman SS. Comparing different CT, PET and MRI multi-modality image combinations for deep learning-based head and neck tumor segmentation. Acta Oncol. 2021;60:1399-406. 10.1080/0284186x.2021.1949034.
[105] Rodríguez Outeiral R, Bos P, Al-Mamgani A, Jasperse B, Simões R, van der Heide UA. Oropharyngeal primary tumor segmentation for radiotherapy planning on magnetic resonance imaging using deep learning. Phys Imaging Radiat Oncol. 2021;19:39-44. 10.1016/j.phro.2021.06.005.
[106] Shi J, Ding X, Liu X, Li Y, Liang W, Wu J. Automatic clinical target volume delineation for cervical cancer in CT images using deep learning. Med Phys. 2021;48:3968-81. 10.1002/mp.14898.
[107] Spoor DS, Sijtsema NM, van den Bogaard VAB, van der Schaaf A, Brouwer CL, Ta BDP, 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. 10.1016/j.radonc.2021.07.025.
[108] Trimpl MJ, Boukerroui D, Stride EPJ, Vallis KA, Gooding MJ. Interactive contouring through contextual deep learning. Med Phys. 2021;48:2951-9. 10.1002/mp.14852.
[109] van Velzen SGM, Bruns S, Wolterink JM, Leiner T, Viergever MA, Verkooijen HM, et al. AI-Based Quantification of Planned Radiation Therapy Dose to Cardiac Structures and Coronary Arteries in Patients With Breast Cancer. Int J Radiat Oncol Biol Phys. 2022;112:611-20. 10.1016/j.ijrobp.2021.09.009.
[110] Wahid KA, Ahmed S, He R, van Dijk LV, Teuwen J, McDonald BA, et al. Evaluation of deep learning-based multiparametric MRI oropharyngeal primary tumor auto-segmentation and investigation of input channel effects: Results from a prospective imaging registry. Clin Transl Radiat Oncol. 2022;32:6-14. 10.1016/j.ctro.2021.10.003.
[111] Wang T, Lei Y, Roper J, Ghavidel B, Beitler JJ, McDonald M, et al. Head and neck multi-organ segmentation on dual-energy CT using dual pyramid convolutional neural networks. Phys Med Biol. 2021;66. 10.1088/1361-6560/abfce2.
[112] Wong J, Huang V, Giambattista JA, Teke T, Kolbeck C, Giambattista J, et al. Training and Validation of Deep Learning-Based Auto-Segmentation Models for Lung Stereotactic Ablative Radiotherapy Using Retrospective Radiotherapy Planning Contours. Front Oncol. 2021;11:626499. 10.3389/fonc.2021.626499.
[113] Xie X, Song Y, Ye F, Yan H, Wang S, Zhao X, et al. Prior information guided auto-contouring of breast gland for deformable image registration in postoperative breast cancer radiotherapy. Quant Imaging Med Surg. 2021;11:4721-30. 10.21037/qims-20-1141.
[114] Xu L, Hu J, Song Y, Bai S, Yi Z. Clinical target volume segmentation for stomach cancer by stochastic width deep neural network. Med Phys. 2021;48:1720-30. 10.1002/mp.14733.
[115] Yuan C, Zhang M, Huang X, Xie W, Lin X, Zhao W, et al. Diffuse large B-cell lymphoma segmentation in PET-CT images via hybrid learning for feature fusion. Med Phys. 2021;48:3665-78. 10.1002/mp.14847.

[116] Zhang J, Gu L, Han G, Liu X. AttR2U-Net: A Fully Automated Model for MRI Nasopharyngeal Carcinoma Segmentation Based on Spatial Attention and Residual Recurrent Convolution. Front Oncol. 2021;11:816672. 10.3389/fonc.2021.816672.
[117] Zhang J, Yang Y, Shao K, Bai X, Fang M, Shan G, et al. Fully convolutional network-based multi-output model for automatic segmentation of organs at risk in thorax. Sci Prog. 2021;104:368504211020161. 10.1177/00368504211020161.
[118] Zhang S, Wang H, Tian S, Zhang X, Li J, Lei R, et al. A slice classification model-facilitated 3D encoder-decoder network for segmenting organs at risk in head and neck cancer. J Radiat Res. 2021;62:94-103. 10.1093/jrr/rraa094.
[119] Zhao J, Chen Z, Wang J, Xia F, Peng J, Hu Y, et al. MV CBCT-Based Synthetic CT Generation Using a Deep Learning Method for Rectal Cancer Adaptive Radiotherapy. Front Oncol. 2021;11:655325. 10.3389/fonc.2021.655325.
[120] Zhong Y, Yang Y, Fang Y, Wang J, Hu W. A Preliminary Experience of Implementing Deep-Learning Based Auto-Segmentation in Head and Neck Cancer: A Study on Real-World Clinical Cases. Front Oncol. 2021;11:638197. 10.3389/fonc.2021.638197.
[121] Zhou H, Li Y, Gu Y, Shen Z, Zhu X, Ge Y. A deep learning based automatic segmentation approach for anatomical structures in intensity modulation radiotherapy. Math Biosci Eng. 2021;18:7506-24. 10.3934/mbe.2021371.
[122] Jiang J, Riyahi Alam S, Chen I, Zhang P, Rimner A, Deasy JO, et al. Deep cross-modality (MR-CT) educed distillation learning for cone beam CT lung tumor segmentation. Med Phys. 2021;48:3702-13. 10.1002/mp.14902.
[123] Barrett S, Simpkin AJ, Walls GM, Leech M, Marignol L. Geometric and Dosimetric Evaluation of a Commercially Available Auto-segmentation Tool for Gross Tumour Volume Delineation in Locally Advanced Non-small Cell Lung Cancer: a Feasibility Study. Clin Oncol (R Coll Radiol). 2021;33:155-62. 10.1016/j.clon.2020.07.019.
[124] Boyd R, Basavatia A, Tomé WA. Validation of accuracy deformable image registration contour propagation using a benchmark virtual HN phantom dataset. J Appl Clin Med Phys. 2021;22:58-68. 10.1002/acm2.13246.
[125] Byun HK, Chang JS, Choi MS, Chun J, Jung J, Jeong C, et al. Evaluation of deep learning-based autosegmentation in breast cancer radiotherapy. Radiat Oncol. 2021;16:203. 10.1186/s13014-021-01923-1.
[126] Duma MN, Kulms T, Knippen S, Teichmann T, Wittig A. Breast clinical target volume: HU-based glandular CTVs and ESTRO CTVs in modern and historical radiotherapy treatment planning. Strahlenther Onkol. 2022;198:229-35. 10.1007/s00066-021-01839-5.
[127] Finnegan RN, Orlandini L, Liao X, Yin J, Lang J, Dowling J, et al. Feasibility of using a novel automatic cardiac segmentation algorithm in the clinical routine of lung cancer patients. PLoS One. 2021;16:e0245364. 10.1371/journal.pone.0245364.
[128] Giaj-Levra N, Figlia V, Cuccia F, Mazzola R, Nicosia L, Ricchetti F, et al. Reduction of inter-observer differences in the delineation of the target in spinal metastases SBRT using an automatic contouring dedicated system. Radiat Oncol. 2021;16:197. 10.1186/s13014-021-01924-0.
[129] Lu SL, Xiao FR, Cheng JC, Yang WC, Cheng YH, Chang YC, et al. Randomized multi-reader evaluation of automated detection and segmentation of brain tumors in stereotactic radiosurgery with deep neural networks. Neuro Oncol. 2021;23:1560-8. 10.1093/neuonc/noab071.

[130] Posiewnik M, Piotrowski T. Utility of deformable image registration for adaptive prostate cancer treatment. Analysis and comparison of two commercially available algorithms. Z Med Phys. 2021. 10.1016/j.zemedi.2021.10.001.
[131] Schmidt RM, Delgadillo R, Ford JC, Padgett KR, Studenski M, Abramowitz MC, et al. Assessment of CT to CBCT contour mapping for radiomic feature analysis in prostate cancer. Sci Rep. 2021;11:22737. 10.1038/s41598-021-02154-w.
[132] Thor M, Iyer A, Jiang J, Apte A, Veeraraghavan H, Allgood NB, et al. Deep learning auto-segmentation and automated treatment planning for trismus risk reduction in head and neck cancer radiotherapy. Phys Imaging Radiat Oncol. 2021;19:96-101. 10.1016/j.phro.2021.07.009.
[133] Urago Y, Okamoto H, Kaneda T, Murakami N, Kashihara T, Takemori M, et al. Evaluation of auto-segmentation accuracy of cloud-based artificial intelligence and atlas-based models. Radiat Oncol. 2021;16:175. 10.1186/s13014-021-01896-1.
[134] Wong J, Huang V, Wells D, Giambattista J, Giambattista J, Kolbeck C, et al. Implementation of deep learning-based auto-segmentation for radiotherapy planning structures: a workflow study at two cancer centers. Radiat Oncol. 2021;16:101. 10.1186/s13014-021-01831-4.
[135] Zabel WJ, Conway JL, Gladwish A, Skliarenko J, Didiodato G, Goorts-Matthews L, et al. Clinical Evaluation of Deep Learning and Atlas-Based Auto-Contouring of Bladder and Rectum for Prostate Radiation Therapy. Pract Radiat Oncol. 2021;11:e80-e9. 10.1016/j.prro.2020.05.013.
[136] Chen W, Wang C, Zhan W, Jia Y, Ruan F, Qiu L, et al. A comparative study of auto-contouring softwares in delineation of organs at risk in lung cancer and rectal cancer. Sci Rep. 2021;11:23002. 10.1038/s41598-021-02330-y.
[137] Huang K, Rhee DJ, Ger R, Layman R, Yang J, Cardenas CE, et al. Impact of slice thickness, pixel size, and CT dose on the performance of automatic contouring algorithms. J Appl Clin Med Phys. 2021;22:168-74. 10.1002/acm2.13207.
[138] Knobe S, Dzierma Y, Wenske M, Berdel C, Fleckenstein J, Melchior P, et al. Feasibility and clinical usefulness of modelling glioblastoma migration in adjuvant radiotherapy. Z Med Phys. 2021. 10.1016/j.zemedi.2021.03.004.
[139] Tibdewal A, Patil M, Misra S, Purandare N, Rangarajan V, Mummudi N, et al. Optimal Standardized Uptake Value Threshold for Auto contouring of Gross Tumor Volume using Positron Emission Tomography/Computed Tomography in Patients with Operable Nonsmall-Cell Lung Cancer: Comparison with Pathological Tumor Size. Indian J Nucl Med. 2021;36:7-13. 10.4103/ijnm.IJNM_134_20.
[140] Okada H, Ito M, Minami Y, Nakamura K, Asai A, Adachi S, et al. Automatic one-click planning for hippocampal-avoidance whole-brain irradiation in RayStation. Med Dosim. 2022;47:98-102. 10.1016/j.meddos.2021.09.003.
[141] Moazzezi M, Rose B, Kisling K, Moore KL, Ray X. Prospects for daily online adaptive radiotherapy via ethos for prostate cancer patients without nodal involvement using unedited CBCT auto-segmentation. J Appl Clin Med Phys. 2021;22:82-93. 10.1002/acm2.13399.
[142] Gan Y, Langendijk JA, Oldehinkel E, Scandurra D, Sijtsema NM, Lin Z, et al. A novel semi auto-segmentation method for accurate dose and NTCP evaluation in adaptive head and neck radiotherapy. Radiother Oncol. 2021;164:167-74. 10.1016/j.radonc.2021.09.019.