Published in final edited form as:
Radiother Oncol. 2021 July ; 160: 185–191. doi:10.1016/j.radonc.2021.05.003.

Metrics to evaluate the performance of auto-segmentation for radiation treatment planning: A critical review

Michael V. Sherer[a,1], **Diana Lin**[b,1], **Sharif Elguindi**[c], **Simon Duke**[d], **Li-Tee Tan**[d], **Jon Cacicedo**[e], **Max Dahele**[f], **Erin F. Gillespie**[b,*]

[a]Department of Radiation Medicine and Applied Sciences, University of California San Diego, La Jolla, United States;

[b]Department of Radiation Oncology, Memorial Sloan Kettering Cancer Center, New York, United States;

[c]Department of Medical Physics, Memorial Sloan Kettering Cancer Center, New York, United States;

[d]Department of Oncology, Cambridge University Hospitals, United Kingdom;

[e]Department of Radiation Oncology, Cruces University Hospital/BioCruces Health Research Institute, Osakidetza, Barakaldo, Spain;

[f]Department of Radiation Oncology, Amsterdam University Medical Center, Amsterdam, The Netherlands

Abstract

Advances in artificial intelligence-based methods have led to the development and publication of numerous systems for auto-segmentation in radiotherapy. These systems have the potential to decrease contour variability, which has been associated with poor clinical outcomes and increased efficiency in the treatment planning workflow. However, there are no uniform standards for evaluating auto-segmentation platforms to assess their efficacy at meeting these goals. Here, we review the most frequently used evaluation techniques which include geometric overlap, dosimetric parameters, time spent contouring, and clinical rating scales. These data suggest that many of the most commonly used geometric indices, such as the Dice Similarity Coefficient, are not well correlated with clinically meaningful endpoints. As such, a multi-domain evaluation, including composite geometric and/or dosimetric metrics with physician-reported assessment, is necessary to gauge the clinical readiness of auto-segmentation for radiation treatment planning.

Keywords

Auto-segmentation; Contouring; Treatment planning; Quality assurance

[*]Corresponding author at: Department of Radiation Oncology, Memorial Sloan Kettering Cancer Center, 1275 York Ave, Box 22, New York, NY 10065, United States. efgillespie@ucsd.edu (E.F. Gillespie).

[1]Authors contributed equally.

Disclosures
E.F.G. is a co-founder of the educational website eContour.org.

Contouring target volumes and surrounding organs-at-risk (OARs), also referred to as segmentation or delineation, is a critical step in radiation treatment planning. It is often performed manually by trained radiation oncology professionals (e.g. radiation oncologists/ trainees, physicists, dosimetrists, therapists) and is a time-consuming and subjective process. Variation among providers is common and likely driven by numerous underlying factors including experience [1] as well as the availability, quality, and interpretation of diagnostic imaging used to assist in delineation [2]. This variation impacts plan quality and patient outcomes, with studies across multiple disease sites associating inadequate contouring with worse disease control and increased toxicity [3–5]. In cooperative group trials, radiation protocol deviations have been associated with inferior survival in head and neck [6] and gastrointestinal [7] cancers, with an increasing number of deviations due to contouring errors [3]. This has led to the publication of numerous contouring guidelines designed to assist practitioners with consistent delineation on clinical trials and in routine practice [8,9].

Auto-segmentation, broadly defined as the generation of contours reflecting the boundary of normal structures and/or target volumes on a digital image by a computer algorithm, has the potential to improve dosimetric consistency and clinical outcomes. While the general concept of automated contouring has been under investigation for more than 20 years [10–12], new approaches such as artificial intelligence-based algorithms [13–15] continue to emerge with improved capabilities. The methodologies used to generate auto-segmented contours have been recently reviewed [13]. In this article, we focus on methods to assess the clinical utility and consistency of the resulting contours to guide implementation for the purposes of radiation therapy treatment planning. Auto-segmentation also has potential applications for diagnostic imaging and radiomics [16–18], which are outside the scope of this review.

To decide on a strategy for evaluating the performance of auto-segmentation, it is useful to consider the *goals* of auto-segmentation. These include reducing contouring time, decreasing interobserver variability, and improving dose consistency and accuracy [19–22]. We have classified evaluation metrics into four domains that can be used to assess these goals (Fig. 1): Geometric, dosimetric, time-based, and qualitative scoring. Measures of geometric overlap are useful to assess contour variability, dosimetric calculations can assess the impact of contouring on the treatment plan, and measurement of time saving is important for understanding impact on clinical workflow. Qualitative scoring of contours by end-users (e.g. physicians) provides an overall assessment of clinical acceptability.

Herein we discuss the benefits and limitations of each evaluation strategy, and investigate which metrics, or combinations thereof, may be best suited to evaluate the performance of auto-segmentation. This topic is both clinically relevant and contemporary: automation in radiotherapy is rapidly gathering pace, not only in contouring, but in the development of online adaptive workflows [23,24], treatment planning, and quality assurance [25]. It is important that technologies are robustly evaluated to ensure they are both fit for purpose and that they do what they claim to.

Methodology

This is a narrative review in which articles were identified through a selective literature search of the PubMed database alongside bibliography searches of relevant articles. Searches focused on identifying publications that address the performance of measures used to evaluate auto-segmentation. Abstracts were screened and if deemed to be relevant, the full text was reviewed by multiple authors (MVS, DL, EFG).

Benchmarking

Most metrics used to evaluate contours require comparison to a benchmark "gold standard" [26]. The importance of gold standard contour selection is highlighted in a recent analysis of cervical cancer brachytherapy planning [27]. In this study, multiple treatment plans were applied to two different gold standard contour sets, which resulted in different estimates of both the mean dose and dose variability to the clinical target volume depending on the benchmark. While the simplest version of a benchmark is a single contour that has been approved for clinical use [28–31], this is most subject to variability given known intra- and inter-observer variations in contour quality seen in clinical practice. To circumvent this, benchmarks incorporating multiple ("expert") contours have been proposed, such as a consensus contour derived by an interdisciplinary expert panel using the simultaneous truth and performance level estimation (STAPLE) [20,32–34]. Another approach is to assess the variation in (edited) auto-segmentation volumes against the variation in manual contours from multiple experts, which eliminates the need for a single "gold standard" contour [19,35]. Use of either of these approaches is recommended to avoid reliance on a single set of contours for benchmarking.

Geometric analysis

The geometric measures used in contour assessment can be subdivided into several classes: volumetric overlap metrics, average or maximal boundary distances, and newer methods based on path length or surface agreement [11,36]. These are illustrated in Fig. 1. Among the most frequently used overlap metrics is the Dice Similarity Coefficient (DSC; or Sørensen–Dice coefficient) [37], which evaluates the intersection of more than one delineated volume over the sum of their total volume, and is scored from zero to one, with one suggesting a perfect overlap:

$$Dice\ Similarity\ Coefficient = \frac{2|\mathrm{A} \cap \mathrm{B}|}{|\mathrm{A}| + |\mathrm{B}|}$$

The DSC is one of the simplest metrics for the assessment of auto-segmented volumes and has been frequently used in the literature [15,37,38]. Despite its popularity, volumetric DSC may not predict the clinical adequacy of contours, as demonstrated in a recent evaluation of the clinical utility of auto-segmentation for prostate cancer [39]. Specifically, it cannot differentiate between systematic and random errors [26] and does not take into account proximity to critical structures [40]. This is also the case for the Jaccard Similarity Coefficient (JSC), a related metric which is defined by the ratio of the intersection of two volumes over their union [41]:

$$Jaccard\ Similarity\ Coefficient = \frac{|A \cap B|}{|A \cup B|}$$

A recent study by Duke et al. reported a significant discordance between the JSC and clinician-rated acceptability scores [42]. Although the contours in this study were manually generated, the findings illustrate the relationship (or lack thereof) between a geometric index and expert evaluation. Using a JSC cut-off of 0.7, only 45% of all contoured volumes considered adequate by experts would have passed (true-positive) while 55% would have failed (false-negative). In addition, 13% of the delineations that failed expert assessment would have passed (false-positive).

Likewise, an analysis of prostate cancer plans found that acceptance rates for manual contours of the bladder and rectum were significantly higher than for automatically generated contours using qualitative clinical evaluation, while surface distance and DSC indicated no difference [43]. Similar results have also been demonstrated in brachytherapy planning for cervical cancer, where a geometric concordance index (which may be described as a generalized version of the DSC/JSC used when comparing more than two volumes [44]) was not predictive of important dosimetric parameters [27]. Together, these studies indicate that conventional volumetric overlap indices, such as JSC and volumetric DSC, provide limited clinical context and correlation with clinical or dosimetric quality.

To overcome the limitations of volume-based metrics, some authors suggest the use of spatial distance-based metrics, which are more sensitive to boundary errors [11,41]. These are generated by calculating the closest distance from each point in the reference contour to the experimental contour. The largest of these distances is the maximum surface distance, also called the Hausdorff distance (HD) [45]. The average distance between the two contours can also calculated. Both metrics are calculated as a distance with zero indicating perfect overlap [46]. It is worth noting that volumetric overlap and distance metrics are often not highly correlated, and therefore, potentially complementary [11].

Evaluations of surface distance metrics as a predictor for clinical acceptability, dosimetric consistency, or time saved in the clinical workflow are limited and demonstrate mixed results. For example, Vassen et al. showed only moderate correlation between the maximum HD and time savings in contours of thoracic organs [36], although this was better than the volumetric DSC, which was poorly correlated. Other studies have also failed to demonstrate significant association between surface distances and qualitative contour scoring or time needed to manually adjust contours [43,47].

Another concern when using geometric indices, whether based on volume or distance, is the identification of a threshold value for “acceptable” segmentation. As referenced above, the JSC cut-off value of 0.7 used by Duke et al. [42] was selected based on results from a study by Fokas et al. comparing investigator-delineated contours to “gold standard” contours, which showed the median JSC of investigator planning target volume (PTV) to be 0.75 (IQR: 0.71–0.79) [48]. However, there is no evidence that this or any other cut-off value for JSC is correlated with clinical acceptability and should be considered a valid benchmark. As such, we advise against a universal cut-off to indicate clinical acceptability of a particular contour due to (1) the challenge of the variation of “acceptable” cut-offs by disease site and specific contour region of interest, and (2) the weak correlation of geometric indices with dosimetric measures, time savings, and physician ratings.

These results have helped to drive efforts to develop novel geometric performance metrics for auto-segmentation that are also clinically meaningful. For example, Vaassen et al. introduced the APL, which was defined as the absolute cumulative length of a contour that had to be added or removed during editing [36]. The APL accounts for the number of slices an organ encompasses and is *not normalized by volume*, which is particularly helpful for volumetrically small but elongated organs with poorly visualized boundaries, such as the esophagus. This study also evaluated the use of the surface DSC described by Nikolov et al. [49], which provides a measure of the agreement between just the surfaces of two structures above a clinically determined tolerance parameter, **τ**, as shown in the equation below (Fig. 1).

$$Surface\ DSC = \frac{\left|S_1 \cap B_{2,\tau}\right| + \left|S_2 \cap B_{1,\tau}\right|}{\left|S_1\right| + \left|S_2\right|}$$

Vaassen et al. concluded that (1) both the APL and surface DSC were better predictors of relative and absolute time saving than volumetric DSC and HD, and (2) APL and surface DSC provided additional quantifiable surrogates for the assessment of clinical utility and quality of automatically-generated contours. While promising, even these metrics may be limited by the inherent inability of geometric measures to distinguish where the variation is located within a contour. In isolation, we would favor use of the APL or surface DSC given evidence of a correlation with time savings. A composite of multiple geometric indices is hypothesized to further improve utility, though none exists in the current literature to recommend. Importantly, given the limited data available on corresponding clinical utility, incorporation of additional non-geometric assessments appears warranted.

Dosimetric analysis

A clinical limitation of geometric analysis is that spatial variations in contouring may or may not translate to meaningful changes in radiation dose delivered, depending on the relationship between the dose gradient and the structures in question. Some authors have calculated the dose delivered to auto-segmented and manually-segmented structures as a means of assessing clinical validity. However, this strategy introduces another element of variability, namely the treatment planning process used to generate the dosimetric indices. This process is complex and must account for multiple variables, including the geometry of tumor volumes, position of OARs relative to targets, beam arrangements, and clinical dosimetric requirements [50]. Some studies try to overcome this variability using automated knowledge-based planning to minimize subjectivity that could affect dosimetric parameters [50,51]. Alternatively, a previously generated treatment plan can be overlaid on the auto-segmented contours [19].

A recent study by Kaderka et al. used both geometric indices (e.g. DSC) and multiple dosimetric endpoints to analyze the auto-segmentation of cardiac structures in breast cancer patients [52]. In general, they observed a high degree of concordance between the dosimetry of plans based on auto- and manually-generated contours. However, for certain small substructures such as the left anterior descending artery, the DSC was very low and the dosimetric agreement high. Fig. 2 illustrates this scenario as well as the converse, where OAR contours with nearly complete geometric overlap have significant dosimetric variation. This will be influenced by, among other factors, the dose conformality around a given structure.

Dosimetric calculations appear particularly important when evaluating auto-segmented contours of target volumes. For example, one study utilized auto-segmentation to deform a set of initial contours to a CT scan acquired during treatment in head and neck cancer [53]. Plans based on the auto-segmented contours delivered less coverage (defined as D95% and V95%) to both the clinically-approved gross tumor volume (GTV) and clinical target volume (CTV). The DSC was not correlated with target coverage, and the authors warned against its use as a surrogate for plan quality. Another analysis of auto-segmented head and neck plans found large PTV under-dosing [54] with observed reductions in D99% averaging over 14 Gy despite DSCs ≥ 0.8 and mean HD ≤ 1 mm between auto-segmented and manually-generated target contours. Ultimately, such findings highlight the potential negative impact of uncorrected auto-segmentation errors on radiation dosimetry and plan quality. They also emphasize the importance of dose calculations (ideally with measures to reduce planning variability, such as knowledge-based planning) when evaluating the effectiveness of auto-segmentation platforms.

The specific dosimetric parameter to select for evaluation is dependent on the disease site and clinical scenario. But in contrast to geometric indices, numerous dosimetric parameters (or dose constraints) have been shown to correlate with clinical outcomes – for example, mean dose to the parotid gland affects the risk of xerostomia when treating head and neck cancer [55]. From first principles, maximum doses are typically used for serially arranged OARs (i.e. spinal cord and optic structures), while parallel organs (i.e. lungs, kidneys) and target volume coverage are often evaluated by looking at the percentage of the volume receiving a given fraction of the prescription dose [56]. It is important to recognize that many dosimetric parameters used in both clinical trials and routine practice have *not* been shown to correlate with clinical outcomes. When evaluating the dosimetric impact of auto-segmented contours, priority should be given to evidence-based dose constraints for the disease site in question.

Time analysis

Reduction in contouring time is a clinically meaningful way to evaluate automated delineation. Strategies for measuring time spent contouring include manual timing in a test environment [21], self-reporting by providers [22,29,39,57–59], and automatic measurements using software [29,60,61]. One specific platform called “Big Brother” has been used to analyze variation in human contouring behaviors [60]. This software accounted for user inactivity or distraction by discounting any pauses in input activity longer than a specified time interval (which ranged from 30 s to 5 min). The results showed good correlation between automatically recorded (6.2 min/case) and self-reported (5.5 min/case) time saving. Since manual timing requires a simulated environment that is often not feasible, and access to the "Big Brother" software may be limited, self-reporting is therefore a reliable method to evaluate time spent contouring.

It is important to note that both relative and absolute time savings should be reported as both estimates have value to the audience. One randomized trial compared time spent contouring among residents assigned to (1) automated delineation with editing or (2) manual segmentation, and reported a 30% time reduction with atlas-based auto-segmentation of OARs [22]. However, mean time savings per resident varied from 2.3 to 15.7 min, as senior residents were faster at contouring and thus received a smaller absolute benefit.

Time savings is inherently valuable as it allows radiation oncology professionals to redirect their efforts to other meaningful activities, such as direct patient care or peer review. Data support the importance of peer review for optimizing contour quality in clinical practice, which is currently hampered by a lack of available time [62]. However, speed alone should not be used to declare auto-segmentation successful without a concurrent rigorous validation ensuring that the quality of such contours is not compromised.

Qualitative scoring systems

In clinical practice, the ultimate acceptability of contours, whether automated or manually generated, is determined by physician judgment. While this introduces potential for subjectivity and interrater variability, there is clinical trial quality assurance data that physician-assessed protocol deviations, of which contours are an important component [63], do correlate with patient outcomes. Multiple meta-analyses of cooperative group trials have found that protocol deviations were associated with increased mortality and treatment failure [63,64]. Exact rating systems vary between trials, however amongst the most common is a three-point scale consisting of: (1) Accept/per protocol; (2) Minor deviation, (3) Major deviation [3,63]. The relevance of contour review is also recognized in routine clinical practice, and has led to, for example, implementation of peer review chart rounds in which an inadequate contour is decided by consensus [65,66].

As discussed above, Duke et al. found that geometric overlap was an inadequate predictor of expert-assessed scores on a three-point scale [42]. Given the difficulty in finding surrogates for expert review, several auto-segmentation studies have implemented physician review with similar scoring scales to clinical trials [31,47,57]. Others have expanded this approach by using a broader seven-point categorization system [67]. In a recent evaluation of the three-point system, expert case reviewers noted that a five-point system might have provided greater ability to differentiate contour quality, since no physicians reported "unacceptable/-major edits" [39].

Instead of asking end users to rate the quality of auto-segmented volumes, an alternative approach is to ask users to distinguish the origin of a contour (auto-segmented or manual). A study by Gooding et al. [47] used this approach, inspired by Turing's Imitation Game [68], to analyze segmentation of six thoracic OARs. Their data confirmed that the "misclassification rate" (user inability to judge the source of contour) appeared to be a better predictor of the time needed for contouring edits than the DSC, suggesting this strategy could be incorporated into the evaluation of auto-segmentation platforms.

Given concerns regarding the potential subjectivity and reproducibility of qualitative scoring systems, some studies have proposed multiple rounds of reviews by panels of experts [69]. McCarroll et al. conducted a study where multiple physicians reviewed the same automated contours for eight OARs on a three-point scale (no edit, minor edit, major edit) [70]. They found that only a small minority of automated contours generated substantial disagreement, suggesting reasonable interobserver variability in clinical qualitative scoring. Quality assurance programs for clinical trials are variable [71] but some protocols include review by at least two experts [72]. At present, an optimal contour scoring system likely involves at minimum a blinded physician qualitative review. Our recommendations align with those proposed in a recently published framework for evaluation of treatment planning studies [73], which also noted the importance of clinician evaluation.

While physician scoring may be a reliable approach, it is time-consuming and may be difficult to implement in some settings. Centralized quality assurance programs for clinical trials face similar logistical challenges. Proposed solutions have included review of only the first ten cases from a given center, as this is when the majority of improvement is reported [74], or the use of a benchmark case [75]. Such approaches could similarly be applied to the clinical implementation of auto-segmentation [39]. Regardless, physician feedback and participation in the process of implementation is as important as, in the authors' experience, algorithms that are often implemented without providing definite utility. These challenges highlight the ongoing need to develop and validate appropriate surrogate measures for clinician assessment to facilitate a thorough evaluation and comparison of auto-segmentation platforms.

Conclusion

Auto-segmentation algorithms offer the potential to improve the consistency and speed of delineation in radiation oncology. As the methods for generating auto-segmented contours continue to evolve, new approaches and algorithms will emerge, and it is critical for these to be evaluated by metrics that reflect clinically meaningful outcomes. Recent studies question the correlation of commonly used measures of geometric overlap such as the DSC and HD, with dose delivered, clinical acceptability, and time saved. As such, these measures should preferably not be utilized as the sole determinant of contour quality. If clinically relevant contour quality is the endpoint, then the gold standard remains physician evaluation, which is supported by the strongest evidence for correlation with clinical outcomes. Since this is a time-consuming and labor-intensive approach, there is an unmet need for the development and validation of adequate surrogate measures to allow for more efficient evaluation of automated platforms. The approach to evaluation also needs to consider the purpose of the evaluation, such as anatomical "correctness", suitability for segmentation of non-critical or critical structures in routine clinical practice, or time savings. If the goal of the auto-segmentation is more limited, then a specific geometric, dosimetric or time-based metric may be adequate to address the particular question being posed.

Funding

This work is supported by an MSK Core Grant (P30 CA008748), as well as the Radiologic Society of North American (RSNA) (EI1902, E.F.G), Agency for Healthcare Research and Quality (AHRQ) (R18 HS026881, E.F.G.), and Varian Medical Systems (M.D.).

Conflict of interest

EFG is a cofounder of the educational website eContour.org. EFG reports funding from a Radiologic Society of North America (RSNA) Innovation grant. MD reports research funding from Varian Medical Systems outside the scope of this work. There are no other conflicts of interest to report.

References

[1]. Boero IJ, Paravati AJ, Xu B, Cohen EEW, Mell LK, Le QT, et al. Importance of radiation oncologist experience among patients with head-and-neck cancer treated with intensity-modulated radiation therapy. J Clin Oncol 2016;34:684–90. [PubMed: 26729432]

[2]. Dalah E, Moraru I, Paulson E, Erickson B, Li XA. Variability of target and normal structure delineation using multimodality imaging for radiation therapy of pancreatic cancer. Int J Radiat Oncol Biol Phys 2014;89:633–40. [PubMed: 24755533]

[3]. Kachnic LA, Winter K, Myerson RJ, Goodyear MD, Willins J, Esthappan J, et al. RTOG 0529: a phase 2 evaluation of dose-painted intensity modulated radiation therapy in combination with 5-fluorouracil and mitomycin-C for the reduction of acute morbidity in carcinoma of the anal canal. Int J Radiat Oncol Biol Phys 2013;86:27–33. [PubMed: 23154075]

[4]. Berry SL, Boczkowski A, Ma R, Mechalakos J, Hunt M. Interobserver variability in radiation therapy plan output: results of a single-institution study. Pract Radiat Oncol 2016;6:442–9. [PubMed: 27374191]

[5]. 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. [PubMed: 27679540]

[6]. Peters LJ, O'Sullivan B, Giralt J, Fitzgerald TJ, Trotti A, Bernier J, et al. Critical impact of radiotherapy protocol compliance and quality in the treatment of advanced head and neck cancer: results from TROG 02.02. J Clin Oncol 2010;28:2996–3001. [PubMed: 20479390]

[7]. Abrams RA, Winter KA, Regine WF, Safran H, Hoffman JP, Lustig R, et al. Failure to adhere to protocol specified radiation therapy guidelines was associated with decreased survival in RTOG 9704–a phase III trial of adjuvant chemotherapy and chemoradiotherapy for patients with resected adenocarcinoma of the pancreas. Int J Radiat Oncol Biol Phys 2012;82:809–16. [PubMed: 21277694]

[8]. Lin D, Lapen K, Sherer MV, Kantor J, Zhang Z, Boyce LM, 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;107:827–35. [PubMed: 32311418]

[9]. Mir R, Kelly SM, Xiao Y, Moore A, Clark CH, Clementel E, et al. Organ at risk delineation for radiation therapy clinical trials: Global Harmonization Group consensus guidelines. Radiother Oncol 2020;150:30–9. [PubMed: 32504762]

[10]. Chaney EL, Pizer SM. Autosegmentation of images in radiation oncology. J Am Coll Radiol 2009;6:455–8. [PubMed: 19467494]

[11]. Sharp G, Fritscher KD, Pekar V, Peroni M, Shusharina N, Veeraraghavan H, et al. Vision 20/20: perspectives on automated image segmentation for radiotherapy. Med Phys 2014;41:050902. 10.1118/1.4871620. [PubMed: 24784366]

[12]. Elliott PJ, Knapman JM, Schlegel W. Interactive image segmentation for radiation treatment planning. IBM Syst J 1992;31:620–34.

[13]. Cardenas CE, Yang J, Anderson BM, Court LE, Brock KB. Advances in auto-segmentation. Semin Radiat Oncol 2019;29:185–97. [PubMed: 31027636]

[14]. Mak RH, Endres MG, Paik JH, Sergeev RA, Aerts H, Williams CL, et al. Use of crowd innovation to develop an artificial intelligence-based solution for radiation therapy targeting. JAMA Oncol 2019;5:654. [PubMed: 30998808]

[15]. Vrtovec T, Močnik D, Strojan P, Pernuš F, Ibragimov B. Auto-segmentation of organs at risk for head and neck radiotherapy planning: from atlas-based to deep learning methods. Med Phys 2020;47.

[16]. Maleki F, Le WT, Sananmuang T, Kadoury S, Forghani R. Machine learning applications for head and neck imaging. Neuroimaging Clin N Am 2020;30:517–29. [PubMed: 33039001]

[17]. Hatt M, Lee JA, Schmidtlein CR, Naqa IE, Caldwell C, De Bernardi E, et al. Classification and evaluation strategies of auto-segmentation approaches for PET: Report of AAPM task group No. 211. Med Phys. 2017;44:e1–e42. [PubMed: 28120467]

[18]. Polan DF, Brady SL, Kaufman RA. Tissue segmentation of computed tomography images using a Random Forest algorithm: a feasibility study. Phys Med Biol 2016;61:6553–69. [PubMed: 27530679]

[19]. Tao C-J, Yi J-L, Chen N-Y, Ren W, Cheng J, Tung S, et al. Multi-subject atlas-based auto-segmentation reduces interobserver variation and improves dosimetric parameter consistency for organs at risk in nasopharyngeal carcinoma: a multi-institution clinical study. Radiother Oncol 2015;115:407–11. [PubMed: 26025546]

[20]. Hwee J, Louie AV, Gaede S, Bauman G, D'Souza D, Sexton T, et al. Technology assessment of automated atlas based segmentation in prostate bed contouring. Radiat Oncol 2011;6:110. [PubMed: 21906279]

[21]. Young AV, Wortham A, Wernick I, Evans A, Ennis RD. Atlas-based segmentation improves consistency and decreases time required for contouring postoperative endometrial cancer nodal volumes. Int J Radiat Oncol Biol Phys 2011;79:943–7. [PubMed: 21281897]

[22]. Walker GV, Awan M, Tao R, Koay EJ, Boehling NS, Grant JD, et al. Prospective randomized double-blind study of atlas-based organ-at-risk autosegmentation-assisted radiation planning in head and neck cancer. Radiother Oncol 2014;112:321–5. [PubMed: 25216572]

[23]. Pathmanathan AU, van As NJ, Kerkmeijer LGW, Christodouleas J, Lawton CAF, Vesprini D, et al. Magnetic resonance imaging-guided adaptive radiation therapy: a "game changer" for prostate treatment?. Int J Radiat Oncol Biol Phys 2018;100:361–73. [PubMed: 29353654]

[24]. Tetar S, Bruynzeel A, Bakker R, Jeulink M, Slotman BJ, Oei S, et al. Patient-reported outcome measurements on the tolerance of magnetic resonance imaging-guided radiation therapy. Cureus. 2018;10:e2236. [PubMed: 29719739]

[25]. 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. [PubMed: 32920005]

[26]. 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. [PubMed: 25315862]

[27]. Bell L, Holloway L, Bruheim K, Petrič P, Kirisits C, Tanderup K, et al. Dose planning variations related to delineation variations in MRI-guided brachytherapy for locally advanced cervical cancer. Brachytherapy 2020;19:146–53. [PubMed: 32067884]

[28]. Yang J, Beadle BM, Garden AS, Gunn B, Rosenthal D, Ang K, et al. Auto-segmentation of low-risk clinical target volume for head and neck radiation therapy. Pract Radiat Oncol 2014;4:e31–7. [PubMed: 24621429]

[29]. Daisne JF, Blumhofer A. Atlas-based automatic segmentation of head and neck organs at risk and nodal target volumes: a clinical validation. Radiat Oncol 2013;8:154. [PubMed: 23803232]

[30]. La Macchia M, Fellin F, Amichetti M, Cianchetti M, Gianolini S, Paola V, et al. Systematic evaluation of three different commercial software solutions for automatic segmentation for adaptive therapy in head-and-neck, prostate and pleural cancer. Radiat Oncol 2012;7:160. [PubMed: 22989046]

[31]. Zhu M, Bzdusek K, Brink C, Eriksen JG, Hansen O, Jensen HA, et al. Multi-institutional quantitative evaluation and clinical validation of Smart Probabilistic Image Contouring Engine (SPICE) autosegmentation of target structures and normal tissues on computer tomography

images in the head and neck, thorax, liver, and male pelvis areas. Int J Radiat Oncol Biol Phys 2013;87:809–16. [PubMed: 24138920]

[32]. Stapleford LJ, Lawson JD, Perkins C, Edelman S, Davis L, McDonald MW, et al. Evaluation of automatic atlas-based lymph node segmentation for head-and-neck cancer. Int J Radiat Oncol Biol Phys 2010;77:959–66. [PubMed: 20231069]

[33]. Thomson D, Boylan C, Liptrot T, Aitkenhead A, Lee L, Yap B, et al. Evaluation of an automatic segmentation algorithm for definition of head and neck organs at risk. Radiat Oncol 2014;9:173. 10.1186/1748-717X-9-173. [PubMed: 25086641]

[34]. 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. [PubMed: 15250643]

[35]. Vinod SK, Min M, Jameson MG, Holloway LC. A review of interventions to reduce inter-observer variability in volume delineation in radiation oncology. J Med Imaging Radiat Oncol 2016;60:393–406. [PubMed: 27170216]

[36]. 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. [PubMed: 33458300]

[37]. Taha AA, Hanbury A. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. BMC Med Imaging 2015;15:29. [PubMed: 26263899]

[38]. Zou KH, Warfield SK, Bharatha A, Tempany CM, Kaus MR, Haker SJ, et al. Statistical validation of image segmentation quality based on a spatial overlap index. Acad Radiol 2004;11:178–89. [PubMed: 14974593]

[39]. 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. [PubMed: 33667591]

[40]. Rohlfing T, Brandt R, Menzel R, Maurer CR. Evaluation of atlas selection strategies for atlas-based image segmentation with application to confocal microscopy images of bee brains. Neuroimage 2004;21:1428–42. [PubMed: 15050568]

[41]. Hanna GG, Hounsell AR, O'Sullivan JM. Geometrical analysis of radiotherapy target volume delineation: a systematic review of reported comparison methods. Clin Oncol 2010;22:515–25.

[42]. Duke SL, Tan L-T, Jensen NBK, Rumpold T, De Leeuw AAC, Kirisits C, et al. Implementing an online radiotherapy quality assurance programme with supporting continuous medical education - report from the EMBRACE-II evaluation of cervix cancer IMRT contouring. Radiother Oncol 2020;147:22–9. [PubMed: 32240907]

[43]. Gautam A, Weiss E, Williamson J, Ford J, Sleeman W, Jan N, et al. Assessing the correlation between quantitative measures of contour variability and physician's qualitative measure for clinical usefulness of auto-segmentation in prostate cancer radiotherapy. Med Phys 2013;40.

[44]. Kouwenhoven E, Giezen M, Struikmans H. Measuring the similarity of target volume delineations independent of the number of observers. Phys Med Biol 2009;54:2863–73. [PubMed: 19384002]

[45]. Christiaens M, Collette S, Overgaard J, Gregoire V, Kazmierska J, Castadot P, et al. Quality assurance of radiotherapy in the ongoing EORTC 1219-DAHANCA-29 trial for HPV/p16 negative squamous cell carcinoma of the head and neck: results of the benchmark case procedure. Radiother Oncol 2017;123:424–30. [PubMed: 28478912]

[46]. Sim DG, Kwon OK, Park RH. Object matching algorithms using robust Hausdorff distance measures. IEEE Trans Image Process 1999;8:425–9. [PubMed: 18262885]

[47]. Gooding MJ, Smith AJ, Tariq M, Aljabar P, Peressutti D, Stoep J, et al. Comparative evaluation of autocontouring in clinical practice: a practical method using the Turing test. Med Phys 2018;45:5105–15. [PubMed: 30229951]

[48]. Fokas E, Clifford C, Spezi E, Joseph G, Branagan J, Hurt C, et al. Comparison of investigator-delineated gross tumor volumes and quality assurance in pancreatic cancer: analysis of the pretrial benchmark case for the SCALOP trial. Radiother Oncol 2015;117:432–7. [PubMed: 26328939]

[49]. Nikolov S Blackwell S Mendes R De Fauw J Meyer C Hughes C et al. Deep learning to achieve clinically applicable segmentation of head and neck anatomy for radiotherapy ArXiv. 2018;1809.04430.

[50]. Fung NTC, Hung WM, Sze CK, Lee MCH, Ng WT. Automatic segmentation for adaptive planning in nasopharyngeal carcinoma IMRT: time, geometrical, and dosimetric analysis. Med Dosim 2020;45:60–5. [PubMed: 31345672]

[51]. Delaney AR, Dahele M, Slotman BJ, Verbakel WFAR. Is accurate contouring of salivary and swallowing structures necessary to spare them in head and neck VMAT plans?. Radiother Oncol 2018;127:190–6. [PubMed: 29605479]

[52]. Kaderka R, Gillespie EF, Mundt RC, Bryant AK, Sanudo-Thomas CB, Harrison AL, et al. Geometric and dosimetric evaluation of atlas based auto-segmentation of cardiac structures in breast cancer patients. Radiother Oncol 2019;131:215–20. [PubMed: 30107948]

[53]. Tsuji SY, Hwang A, Weinberg V, Yom SS, Quivey JM, Xia P. Dosimetric evaluation of automatic segmentation for adaptive IMRT for head-and-neck cancer. Int J Radiat Oncol Biol Phys 2010;77:707–14. [PubMed: 20231063]

[54]. Voet PWJ, Dirkx MLP, Teguh DN, Hoogeman MS, Levendag PC, Heijmen BJM. 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. [PubMed: 21269714]

[55]. Deasy JO, Moiseenko V, Marks L, Chao KSC, Nam J, Eisbruch A. Radiotherapy dose-volume effects on salivary gland function. Int J Radiat Oncol Biol Phys 2010;76:S58–63. [PubMed: 20171519]

[56]. Lee AW, Ng WT, Pan JJ, Chiang C-L, Poh SS, Choi HC, et al. International guideline on dose prioritization and acceptance criteria in radiation therapy planning for nasopharyngeal carcinoma. Int J Radiat Oncol Biol Phys 2019;105:567–80. [PubMed: 31276776]

[57]. Lustberg T, van Soest J, Gooding M, Peressutti D, Aljabar P, van der Stoep J, et al. Clinical evaluation of atlas and deep learning based automatic contouring for lung cancer. Radiother Oncol 2018;126:312–7. [PubMed: 29208513]

[58]. Reed VK, Woodward WA, Zhang L, Strom EA, Perkins GH, Tereffe W, et al. Automatic segmentation of whole breast using atlas approach and deformable image registration. Int J Radiat Oncol Biol Phys 2009;73:1493–500. [PubMed: 18804333]

[59]. van der Veen J, Willems S, Deschuymer S, Robben D, Crijns W, Maes F, et al. Benefits of deep learning for delineation of organs at risk in head and neck cancer. Radiother Oncol 2019;138:68–74. [PubMed: 31146073]

[60]. Steenbakkers RJHM, Duppen JC, Fitton I, Deurloo KEI, Zijp L, Uitterhoeve ALJ, et al. Observer variation in target volume delineation of lung cancer related to radiation oncologist-computer interaction: a ‘Big Brother’ evaluation. Radiother Oncol 2005;77:182–90. [PubMed: 16256231]

[61]. Multi-Institutional Target Delineation in Oncology G. Human-computer interaction in radiotherapy target volume delineation: a prospective, multi-institutional comparison of user input devices. J Digit Imaging. 2011;24:794–803. [PubMed: 20978922]

[62]. Cha E, Brower J, Sherer MV, Golden D, Chimonas S, Korenstein D, et al. Assessment of contouring practices and econtour use among US radiation oncologists: a mixed methods study. ROECSG 2020 Spring Symposium. Virtual2020. p. 10.

[63]. Weber DC, Tomsej M, Melidis C, Hurkmans CW. QA makes a clinical trial stronger: evidence-based medicine in radiation therapy. Radiother Oncol 2012;105:4–8. [PubMed: 22985777]

[64]. Ohri N, Shen X, Dicker AP, Doyle LA, Harrison AS, Showalter TN. Radiotherapy protocol deviations and clinical outcomes: a meta-analysis of cooperative group clinical trials. J Natl Cancer Inst 2013;105:387–93. [PubMed: 23468460]

[65]. Lawrence YR, Whiton MA, Symon Z, Wuthrick EJ, Doyle L, Harrison AS, et al. Quality assurance peer review chart rounds in 2011: a survey of academic institutions in the United States. Int J Radiat Oncol Biol Phys 2012;84:590–5. [PubMed: 22445006]

[66]. Marks LB, Adams RD, Pawlicki T, Blumberg AL, Hoopes D, Brundage MD, et al. Enhancing the role of case-oriented peer review to improve quality and safety in radiation oncology: executive summary. Pract Radiat Oncol 2013;3:149–56. [PubMed: 24175002]

[67]. Greenham S, Dean J, Fu CKK, Goman J, Mulligan J, Tune D, et al. Evaluation of atlas-based auto-segmentation software in prostate cancer patients. J Med Radiat Sci 2014;61:151–8. [PubMed: 26229651]

[68]. Turing A Computing machinery and intelligence. Mind. 1950;59:433.

[69]. Teguh DN, Levendag PC, Voet PWJ, Al-Mamgani A, Han X, Wolf TK, et al. Clinical validation of atlas-based auto-segmentation of multiple target volumes and normal tissue (swallowing/mastication) structures in the head and neck. Int J Radiat Oncol Biol Phys 2011;81:950–7. [PubMed: 20932664]

[70]. McCarroll RE, Beadle BM, Balter PA, Burger H, Cardenas CE, Dalvie S, et al. Retrospective validation and clinical implementation of automated contouring of organs at risk in the head and neck: a step toward automated radiation treatment planning for low- and middle-income countries. J Glob Oncol 2018:1–11.

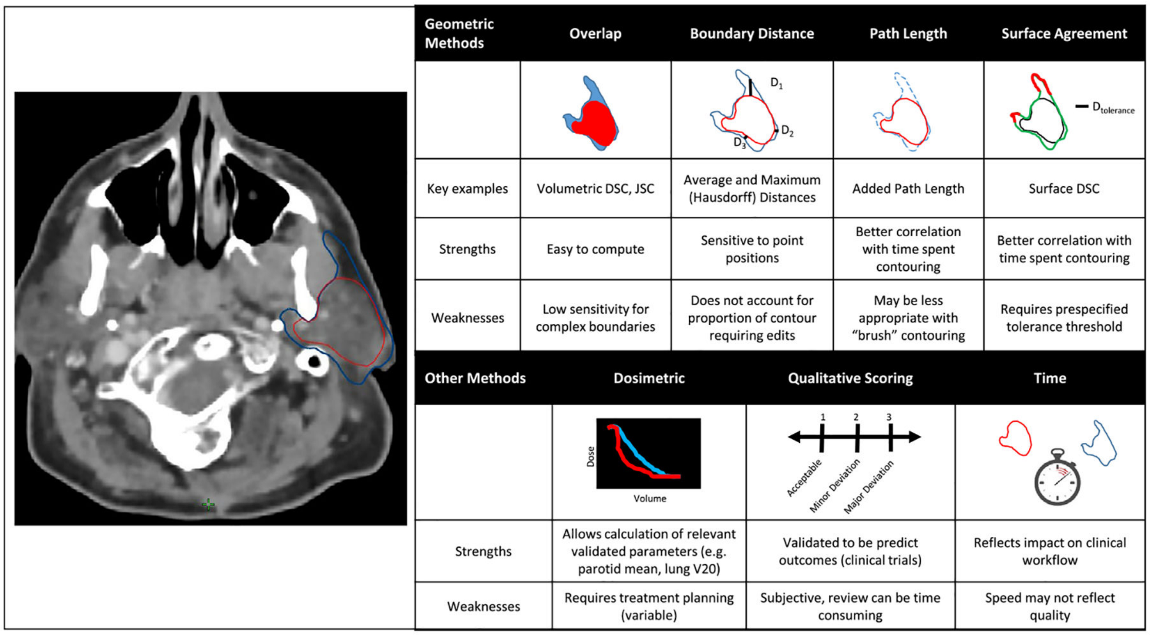
[71]. Fairchild A, Straube W, Laurie F, Followill D. Does quality of radiation therapy predict outcomes of multicenter cooperative group trials? A literature review. Int J Radiat Oncol Biol Phys 2013;87:246–60. [PubMed: 23683829]

[72]. Sanuki-Fujimoto N, Ishikura S, Hayakawa K, Kubota K, Nishiwaki Y, Tamura T. Radiotherapy quality assurance review in a multi-center randomized trial of limited-disease small cell lung cancer: the Japan Clinical Oncology Group (JCOG) trial 0202. Radiat Oncol 2009;4:16. [PubMed: 19490617]

[73]. Hansen CR, Crijns W, Hussein M, Rossi L, Gallego P, Verbakel W, et al. Radiotherapy Treatment plannINg study Guidelines (RATING): a framework for setting up and reporting on scientific treatment planning studies. Radiother Oncol 2020;153:67–78. [PubMed: 32976873]

[74]. Joye I, Lambrecht M, Jegou D, Hortobágyi E, Scalliet P, Haustermans K. Does a central review platform improve the quality of radiotherapy for rectal cancer? Results of a national quality assurance project. Radiother Oncol 2014;111:400–5. [PubMed: 24746578]

[75]. Chang ATY, Tan LT, Duke S, Ng WT. Challenges for quality assurance of target volume delineation in clinical trials. Front Oncol 2017;7:221. [PubMed: 28993798]


D1
D2
D3
Dtolerance
Dose
Volume
1
2
3
Acceptable
Minor Deviation
Major Deviation

Geometric Methods	Overlap	Boundary Distance	Path Length	Surface Agreement
Key examples	Volumetric DSC, JSC	Average and Maximum (Hausdorff) Distances	Added Path Length	Surface DSC
Strengths	Easy to compute	Sensitive to point positions	Better correlation with time spent contouring	Better correlation with time spent contouring
Weaknesses	Low sensitivity for complex boundaries	Does not account for proportion of contour requiring edits	May be less appropriate with "brush" contouring	Requires prespecified tolerance threshold

Other Methods	Dosimetric	Qualitative Scoring	Time
Strengths	Allows calculation of relevant validated parameters (e.g. parotid mean, lung V20)	Validated to be predict outcomes (clinical trials)	Reflects impact on clinical workflow
Weaknesses	Requires treatment planning (variable)	Subjective, review can be time consuming	Speed may not reflect quality

Fig. 1.

Overview of metrics used for contour evaluation.

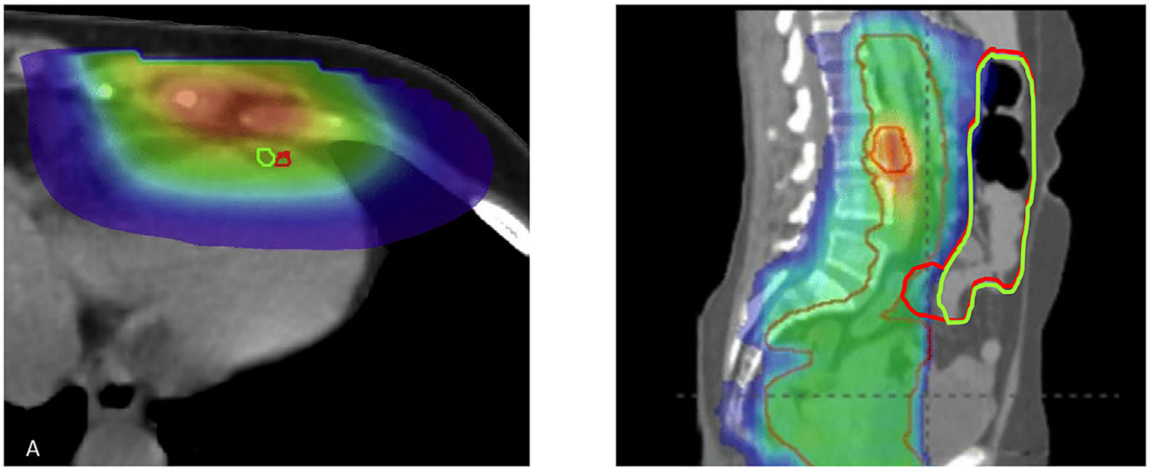


Fig. 2.

Examples of Geometric-Dosimetric discordance. On the left, two contours of the left anterior descending artery have almost no overlap but both structures receive a nearly identical dose (Figure reprinted with permission from reference [52]). On the right, two small bowel contours have excellent geometric agreement but disagreement within a high dose gradient region would result in a higher Dmax for the red contour.