

Original Article

Cite this article: Sillanpaa J, Sood A, and Reynolds M. (2025) A geometric and dosimetric comparison of three AI-based autocontouring packages in the head and neck region. *Journal of Radiotherapy in Practice*. 24(e33), 1–8. doi: [10.1017/S1460396925100204](https://doi.org/10.1017/S1460396925100204)

Received: 13 May 2025
Revised: 20 June 2025
Accepted: 26 June 2025


Keywords:

AI; autocontouring; artificial intelligence; efficiency and quality; head and neck cancer

Corresponding author:

Jussi Sillanpaa; Email: silla032@umn.edu

A geometric and dosimetric comparison of three AI-based autocontouring packages in the head and neck region

Jussi Sillanpaa , Amit Sood and Margaget Reynolds

Dept. of Radiation Oncology, University of Minnesota, 420 Delaware St SE, Minneapolis, Minnesota, MN USA

Abstract

Introduction: AI-based autocontouring products claim to be able to segment organs with accuracy comparable to humans. We compare the geometric and dosimetric performance of three AI-based autocontouring packages (Autocontour 2.5.6, (“RF”); Annotate 2.3.1, (“TP”) and RT-Mind_AI 1.0, (“MM”)) in the head and neck region.

Methods: We generated 14 organ at risk (OAR) autocontours on 13 computed tomography (CT) image sets. They were compared with clinical (human-generated) contours. The geometric differences were quantified by calculating Dice coefficients and Hausdorff distances. The autocontours were compared visually with the clinical contours by an expert physician. The autocontour sets were also ranked for accuracy by two physicians. The dosimetric effects were evaluated by recalculating treatment plans on the autocontoured CT sets.

Results: RF and TP slightly outperformed MM in geometric metrics (the percentage of OARs having mean Dice coefficients > 0.7 was RF 57.1 %, TP 64.3 % and MM 50.0%). The physician judged RF and TP contours to be more anatomically accurate, on average, than the manual contours (manual contour mean accuracy score 2.49, RF 2.28, MM 3.24, TP 1.93). The mean scores given to the autocontours by the two physicians were better for RF and TP, compared to MM (RF 1.86, MM 2.36, TP 1.77). The dosimetric differences were similar for all three programs and were not strongly correlated with the geometric differences.

Conclusions: The performance of the three autocontouring packages in the head and neck region is similar, with TP and RF slightly outperforming MM. The correlation between geometric and dosimetric metrics is not strong, and dosimetric evaluation is therefore recommended before clinical use of autocontouring software.

Introduction

Contouring organs at risk (OARs) is a crucial task in radiation therapy – the treatment plan and the dose–volume histogram (DVH) are only as good as the contours and generating the contours is a major time commitment.^{1,2} Atlas-based autocontouring has been available for more than two decades, but its usefulness outside the brain is limited.^{3,4} Recently, artificial intelligence (AI)-based autocontouring tools have become commercially available.^{5–7} They promise improved accuracy, greatly reduced variation and significant efficiency gains.^{8–10} Although AI-generated contours will always have to be reviewed and, if necessary, adjusted by humans,¹¹ they may enable considerable time savings.¹² This is especially true for the head and neck region, for which a large number of OARs are often contoured¹³ and a delay in the start of radiotherapy is associated with an increased risk of local recurrence.^{14,15} Adaptive radiotherapy, in particular, would benefit greatly from fast OAR contour generation.^{16,17}

We study the performance of three commercially available AI-based autocontouring packages (Autocontour 2.5.6, RADformation Inc. (“RF”), New York, NY, USA; Annotate 2.3.1, Therapanacea (“TP”), Paris, France; RT-Mind_AI 1.0, MedMind Inc. (“MM”), Delaware, USA) in the head and neck region. Head and neck contouring is a useful test case for several reasons. The large number of OARs translates to a high potential for time savings, the patients have often had prior surgery, meaning their anatomy may be distorted and organs may have been fully or partially removed and the presence of metallic dental work results in artifacts that make contouring challenging. We compare the autocontours with those generated manually by experienced dosimetrists and quantify the geometric and dosimetric effects of using autocontouring and rank the autocontours for anatomical accuracy.

Methods

The computed tomography (CT) treatment planning image sets of 13 head and neck cancer patients (5 oropharynx, 2 sinonasal, 2 oral cavity, 2 orbit, 1 buccal and 1 frontal face) consenting for the use of their patient information in research were randomly selected (according to a

Table 1. The mean dice similarity coefficient and standard deviation of the autocontours (N = number of patients with the OAR contoured), the best value for each OAR in bold

Organ	N	RF Ave	RF StDev	MM Ave	MM StDev	TP Ave	TP StDev
Brainstem	13	0.863	0.022	0.857	0.031	0.881	0.025
Parotid L	13	0.789	0.092	0.786	0.095	0.801	0.094
Parotid R	12	0.815	0.067	0.770	0.053	0.815	0.050
Chiasm	12	0.227	0.163	0.197	0.157	0.337	0.171
Esophagus	7	0.760	0.103	0.653	0.126	0.739	0.131
OpN L	12	0.561	0.097	0.526	0.064	0.612	0.083
OpN R	12	0.600	0.046	0.547	0.056	0.633	0.047
Mandible	11	0.851	0.028	0.811	0.025	0.875	0.021
OralCav	10	0.769	0.095	0.617	0.060	0.785	0.054
Cochlea L	11	0.412	0.134	N/A	N/A	0.509	0.215
Cochlea R	11	0.466	0.133	N/A	N/A	0.530	0.161
SubmandR	6	0.817	0.043	0.748	0.103	0.785	0.098
SubmandL	6	0.773	0.124	0.662	0.167	0.778	0.111
Spinal cord	9	0.655	0.163	0.767	0.097	0.773	0.112
Average over all OARs		0.668		0.662		0.704	

retrospective research protocol, approved by the institutional review board). The slice thickness of the CT scans was 2 mm; the dose calculation grid resolution was 2–3 mm, depending on the patient. The contours for 14 OARs (brainstem, L parotid, R parotid, chiasm, L optic nerve, R optic nerve, esophagus, mandible, oral cavity, L cochlea, R cochlea, L submandibular gland, R submandibular gland, spinal cord) were generated by the three autocontouring packages and compared with the clinical (human-generated) contours, generated by experienced dosimetrists and reviewed by physicians. The clinical set did not include every OAR for every patient, and the autocontours were only evaluated if a corresponding clinical contour existed.

The RF and TP autocontouring packages are not trainable, while MM can be adjusted to mimic a particular physician's contouring. We used MM with the default settings, to keep the comparison fair and because one of the common aims of using autocontouring is to enforce uniformity in OAR structures across an institution.

Geometric performance was quantified by calculating Dice similarity coefficients (DSCs) and Hausdorff distances (HDs) between the clinical contours and the autocontours, using the 3D Slicer software. The DSC between structures A and B is defined as¹⁸

$$\text{DSC} = 2 \times \text{volume}(A \cap B) / (\text{volume}(A) + \text{volume}(B))$$

DSC has a value between 0 and 1, with 1 indicating perfect overlap and 0 no overlap. Values of approximately 0.7 are generally considered indicating good overlap, but this will depend on the size of the structure – the overlapping interiors of large structures will result in a high DSC, even if the boundaries do not match well.

The two directional, 3-dimensional HD between structures A and B is defined as the maximum of the minimum distances of points a on structure A and b on structure B,¹⁹ $\text{HD}(A, B)$

$$= \max \{ \sup_{a \in A} [d(a, B)], \sup_{b \in B} [d(A, b)] \}$$

HD has a unit of length and a non-negative value, with 0 mm indicating perfect agreement. We calculated HD95 (95% of the points on the boundaries of the structures are within HD95 of each other). In contrast to DSC, it is easier for small structures to get good (small) HD values, even if they do not overlap at all (a longer contour makes finding a really bad point more likely). HD is determined by the part of the contour with the worst agreement, whereas DSC is affected by all areas that are non-overlapping.

A good DSC or HD indicates good agreement between the autocontours and clinical contours, but does not by itself guarantee anatomical accuracy – there is considerable interobserver variation in clinical contours. This is especially true if the OAR is not expected to get a significant dose that would justify spending a lot of time contouring it manually. Therefore, the anatomical accuracy of the OAR contours was also ranked subjectively by physicians experienced in treating head and neck cancer. A physician not involved in the creation of the manual contours compared their accuracy with the autocontours. The four contour sets were ranked on a four point Likert scale from most (1) to least accurate (4) for each patient and organ, and the scores averaged. Two physicians did a similar ranking for the autocontours only, which were ranked most (1) to least (3) accurate and the scores averaged.

We also quantified the dosimetric performance of the autocontouring packages. Intensity Modulated Radiotherapy (IMRT) treatment plans generated on the clinical contours were recalculated (without reoptimising) on the autocontoured structures and the change in the DVH quantified. The treatment planning system employed was Philips Pinnacle 16.2.1 (Philips Medical Systems, Gainesville, FL, USA). We did not generate new treatment plans based on the autocontoured structures, as this would have added an uncontrolled variable (whether the change in the DVH is due to a change in the OAR contour or the quality of optimisation in the new plan).

Table 2. The mean HD95 and standard deviation [mm] of the autocontours (N = number of patients with the OAR contoured), the best value for each organ in bold

Organ	N	RF Ave	RF StDev	MM Ave	MM StDev	TP Ave	TP StDev
Brainstem	13	4.0	1.2	3.8	1.5	3.9	1.3
Parotid L	13	6.4	2.1	5.9	2.2	5.8	2.9
Parotid R	12	5.8	2.6	6.2	1.7	5.4	2.0
Chiasm	12	5.6	1.6	5.7	1.9	4.6	1.2
Esophagus	7	7.7	8.6	9.2	7.9	8.0	8.4
OpN L	12	3.2	1.1	5.9	2.2	2.6	0.7
OpN R	12	2.6	0.5	6.4	2.6	2.5	0.4
Mandible	11	3.7	3.2	5.7	3.3	3.9	3.9
OralCav	10	8.9	4.1	18.5	2.9	10.7	2.7
Cochlea L	11	2.6	0.7	N/A	N/A	2.9	1.0
Cochlea R	11	2.5	0.6	N/A	N/A	2.6	0.7
SubmandR	6	3.2	0.8	3.8	1.3	3.7	1.5
SubmandL	6	3.7	1.2	6.0	2.2	3.8	1.0
Spinal cord	9	5.2	1.3	3.7	1.8	3.7	1.9
Average over all OARs		4.6		6.7		4.6	

Results

The DSCs are presented in Table 1 and HD95 distances in Table 2. The physician-generated anatomical accuracy scores are listed in Table 3A (manual contours compared with the autocontours) and Table 3B (mean of autocontour scores from two physicians), and the DVH metrics in Table 4. The DVH metrics selected corresponds to those used at our institution for evaluating clinical plans (D_MAX for the brainstem, spinal cord, optic nerves and chiasm, D50 for the parotid and submandibular glands). Figure 1 shows a comparison between clinical and automatically generated parotid and spine contours. Figure 2 shows a comparison of HD95 and DSC for selected OARs for RF, and a comparison of the DSCs of all the autocontouring packages for the same OARs. Figure 3 shows the dosimetric change for the spinal cord and left parotid as a function of HD95 and DSC.

All three autocontouring packages posted similar DSC results (the mean DSC of all OARs is RF 0.668, TP 0.704, MM 0.662). 23/40 (57.5%) of the DSCs were above 0.7 and 35/40 (87.5%) above 0.5, the exceptions being the chiasm and the cochlea. These are small structures, so a low DSC is not surprising.

RF had the best DSC for 2 OARs, TP for 11 and for 1, RF and TP were equally good (MM does not generate cochlear contours on a CT scan). The percentage of OARs having mean Dice coefficients > 0.7 was RF 57.1 %, TP 64.3 %, MM 50.0%). With the exception of the chiasm, the DSC values for RF and TP were similar to those reported in the literature for earlier versions of these programs.^{20–22}

RF and TP had slightly lower HD95 values than MM (mean HD95 of all OARs is RF 4.6 mm, TP 4.6 mm, MM 6.7 mm). RF had the best HD95 for 7 OARs, TP for 5, MM for 1 and for 1, TP and MM were equally good (MM does not generate cochlear contours on a CT scan). In particular, the oral cavity and the optical structures generated by RF and TP matched the clinical contours better than the MM. The structures with the highest mean HD95 were the oral cavity and esophagus, with the parotids and chiasm also having relatively large HDs, despite their smaller size.

The physician-generated anatomical accuracy scores, averaged over all OARs, were clinical contours 2.49, RF 2.28, TP 1.93 and

Table 3A. Anatomical accuracy of the contours, compared to clinical contours (a lower number denotes higher accuracy, the best value for each organ in bold)

Organ	N	Clinical	RF	MM	TP
Brainstem	13	1.77	2.38	3.08	2.77
Parotid L	13	2.62	2.46	3.08	1.62
Parotid R	12	3.08	1.85	3.23	1.85
Chiasm	12	3.08	2.38	2.54	2.00
Esophagus	7	1.75	2.67	3.44	1.89
OpN L	12	2.46	2.62	3.54	1.38
OpN R	12	2.31	2.38	3.62	1.69
Mandible	11	2.33	2.46	3.00	2.00
OralCav	10	2.73	1.36	3.91	2.00
Cochlea L	11	1.67	1.83	N/A	2.50
Cochlea R	11	2.38	1.77	N/A	1.85
SubmandR	6	2.86	1.57	3.43	2.14
SubmandL	6	3.00	1.29	3.71	2.00
Spinal cord	9	1.91	3.91	2.36	1.82
Average		2.49	2.28	3.24	1.93

MM 3.24. Similarly to HD95, RF and TP outperformed MM and were in fact judged slightly more anatomically accurate than the clinical contours. Similar results of AI-based autocontouring being more accurate than clinical contours have been reported by other authors for various body sites.²⁰ The clinical contours were judged most accurate for 3 OARs, RF for 4, TP for 6 and for one OAR, RF and TP were equally accurate. When the two physicians compared the autocontouring sets only, the results were very similar (RF 1.86, TP 1.77, MM 2.36), with RF most accurate for 7 OARs, TP for 6 and MM for 1.

Table 3B. Anatomical accuracy of the autocontours (mean of scores from two physicians). The best value for each organ in bold

Organ	N	RF	MM	TP
Brainstem	13	1.51	2.28	2.13
Parotid L	13	2.12	2.14	1.70
Parotid R	12	1.93	2.30	1.77
Chiasm	12	1.96	2.12	1.93
Esophagus	7	1.65	2.71	1.64
OpN L	12	2.37	2.22	1.41
OpN R	12	2.21	2.29	1.49
Mandible	11	1.64	2.57	1.79
OralCav	10	1.41	3.00	1.59
Cochlea L	11	1.08	N/A	1.92
Cochlea R	11	1.23	N/A	1.77
SubmandR	6	1.31	2.68	2.01
SubmandL	6	1.24	2.67	2.10
Spinal cord	9	2.95	1.32	1.73
Average		1.86	2.36	1.77

Discussion

Generally, a high DSC corresponds to a low HD95, but the relationship between the two metrics is not very strong, for reasons noted in the Methods section. This is illustrated for RF in Figure 2A. Figures 2B and 2C compare the DSC and HD95 values of the three autocontouring packages; this time the correlations are evident, all three have similar DSC and HD95 for the same patient.

The autocontouring packages will always attempt to generate a contour, even for OARs that have been surgically removed (e.g., some patients in our set had their submandibular glands removed prior to the simulation CT scan). The person reviewing the autocontours should be cognizant of this and remove the contours for OARs that are not present, lest target coverage be compromised in an attempt to spare a non-existent OAR.

A possible reason for RF and TP outperforming the clinical contours is that if the OAR is far from the target and not expected to get a significant dose, a person manually contouring it may not want to spend a lot of time maximising the anatomical accuracy. The accuracy of OAR contours is important even in these cases, for example if the patient requires reirradiation at a location closer to the OAR or a retrospective dose response study is carried out at a later date.

It is known that contouring is subject to a degree of interobserver variability; separate physicians may draw the same OAR quite differently. This is particularly relevant to the physician-generated anatomical accuracy scores. A detailed study of the effects of interobserver variability would require blinded comparisons of clinical and AI generated contours by a large number of physicians and is unfortunately outside the scope of this article.

Although several authors have studied the performance of AI-based autocontouring algorithms, they have usually quantified only the geometric differences, not the dosimetric ones.^{13,20,23–27} The relationship between the two, however, is not as straightforward as one might suppose. As shown in Figure 3, a high DSC or a low HD95 do not always correspond to a small dosimetric effect –

if the OAR is in a high dose gradient, even good geometric agreement can result in a large dosimetric effect and vice versa. If the user is interested in the magnitude of the dosimetric effects, a dose calculation with the autocontours should always be performed when testing autocontouring packages.

The mean difference in the D50 values does not depend strongly on the exact shape of the contour and are typically on the order of a few percent. For individual patients, the change in parotid D50 can be big, as indicated by the large standard deviation listed in Table 4. The changes in mean DMAX values are typically bigger since they are determined by the voxel receiving the largest dose.

The MM autocontouring package is trainable, whereas the other two are not. Had we trained MM on our institution's prior patients, its geometric agreement with the clinical contours most likely would have improved.

Chiasm and optic nerves

While the clinical chiasm contours were X-shaped in the axial plane, the autocontours were more elliptical. The chiasm and optic nerves in the clinical contours always overlapped, and the autocontoured ones did not always do so and the transition point from chiasm to optic nerve varied. This resulted in worse geometric metrics, even if the optic pathway as a whole was well contoured. The DSC values of the chiasm were the lowest of any OAR, due partly to its small size but also to the fact that all the autocontouring packages posted better anatomical accuracy scores than the clinical contours for this structure. Other investigators have also reported low DSCs for autocontoured chiasms, possibly due to its poor visualisation on a CT scan.^{23,24} We had the clinical contours for the chiasm retrospectively reviewed by a group of physicians; they agreed that the anatomical accuracy of the contours was not as good as for the other OARs. The HD95 values for nerves were low for RF and TP, but over 5 mm for MM.

The DMAX values for the autocontoured optic nerves and chiasms are, on average, smaller than for the clinical contour. The location of the hotspot in the nerve was usually at the chiasm end; since the clinical nerve contours always overlapped with chiasm, the DMAX for the clinical contours was, on average, higher.

Brainstem and spinal cord

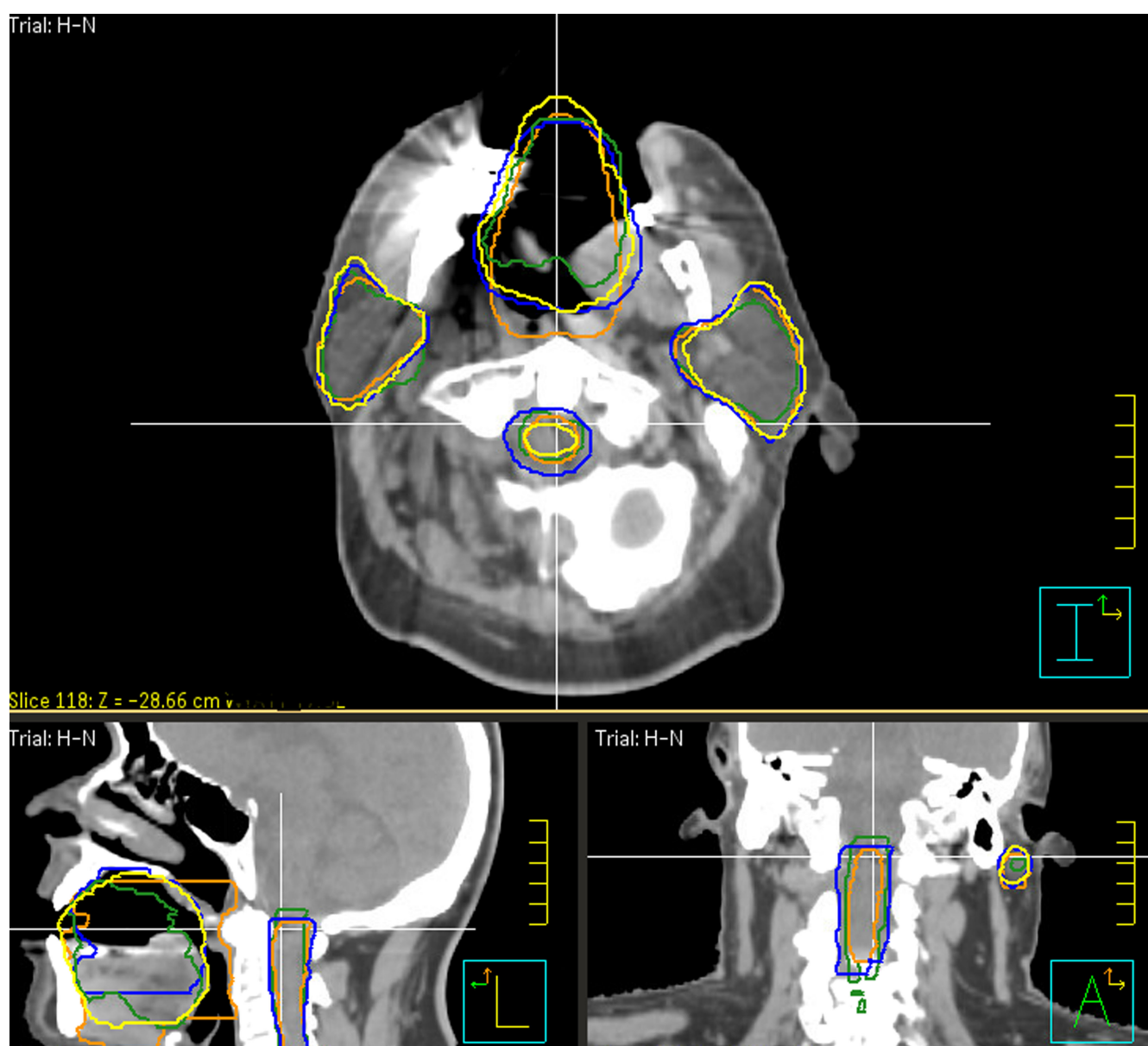
The brainstem DSC and HD95 values were very good for all the autocontouring packages. For the spine, the RF contours are bigger than the others and often approximate the spinal canal, rather than the true spinal cord. This results in the DMAX values being systematically higher for the RF contours. It should be noted that clinical contouring of spinal cord varies from clinic to clinic, a lot of institutions will intentionally contour the whole canal or add a margin, while still calling the structure spinal cord.

Parotids and submandibular glands

The parotid and submandibular gland DSC values were very good. The HD95 values of the submandibular glands were lower than for the parotids, partly due to their smaller size. These structures are not very easy to delineate on a CT set and are often affected by metal artifacts due to dental work. RF and TP were judged to be more anatomically accurate than the clinical contours.

Table 4. Mean changes in DVH metrics for clinical treatment plans recalculated on autocontour sets. The smallest absolute mean change is printed in bold

Organ	N	Metric	RF Mean Diff [%]	RF StDev [%]	MM Mean Diff [%]	MM StDev [%]	TP Mean Diff [%]	TP StDev [%]
Brainstem	13	D_MAX	3.3	9.7	3.1	17.1	7.7	14.5
Parotid L	13	D50	3.5	15.3	-0.1	11.9	0.5	13.1
Parotid R	12	D50	-1.8	12.9	-7.1	7.9	-3.9	8.1
Chiasm	12	D_MAX	-6.0	12.2	-4.5	8.5	-6.0	10.0
OpN L	12	D_MAX	0.1	8.4	-6.8	11.8	-3.5	7.3
OpN R	12	D_MAX	-10.4	19.5	-12.1	13.3	-10.1	14.9
SubmandR	6	D50	-1.9	2.8	-4.5	6.0	-2.0	3.2
SubmandL	6	D50	-0.2	3.2	-0.2	3.8	-0.6	2.0
Spinal cord	9	D_MAX	12.7	9.1	2.5	7.0	-1.8	8.2

**Figure 1.** Parotid, oral cavity and spinal cord contours for a sample patient. Clinical contours: green, RF: blue, TP: yellow, MM: orange.

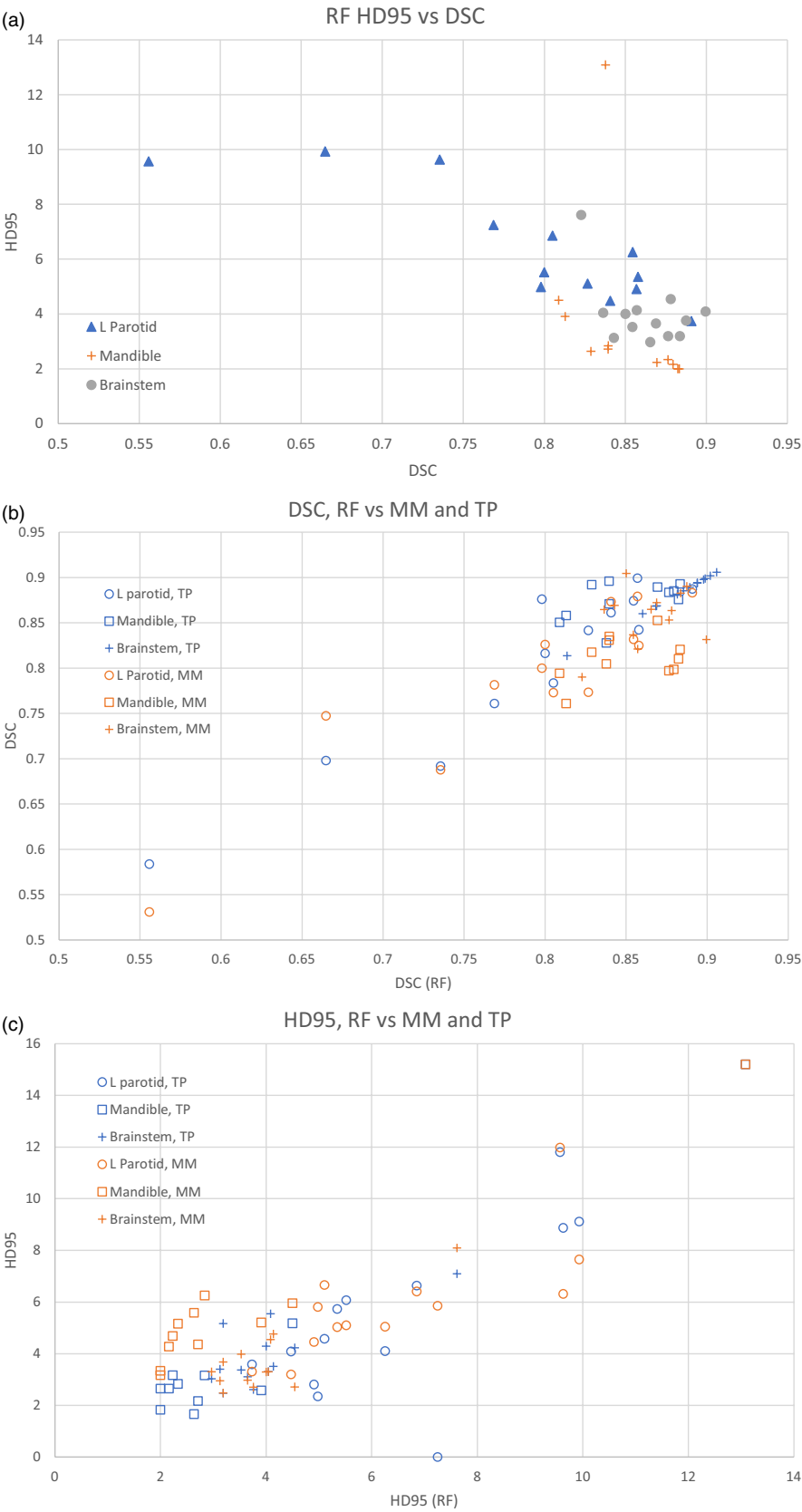


Figure 2. Comparison of DSC and HD95. Top: HD95 and DSC for RF; Middle: DSCs of MM and TP, compared to RF; Bottom: HD95s of MM and TP, compared to RF.

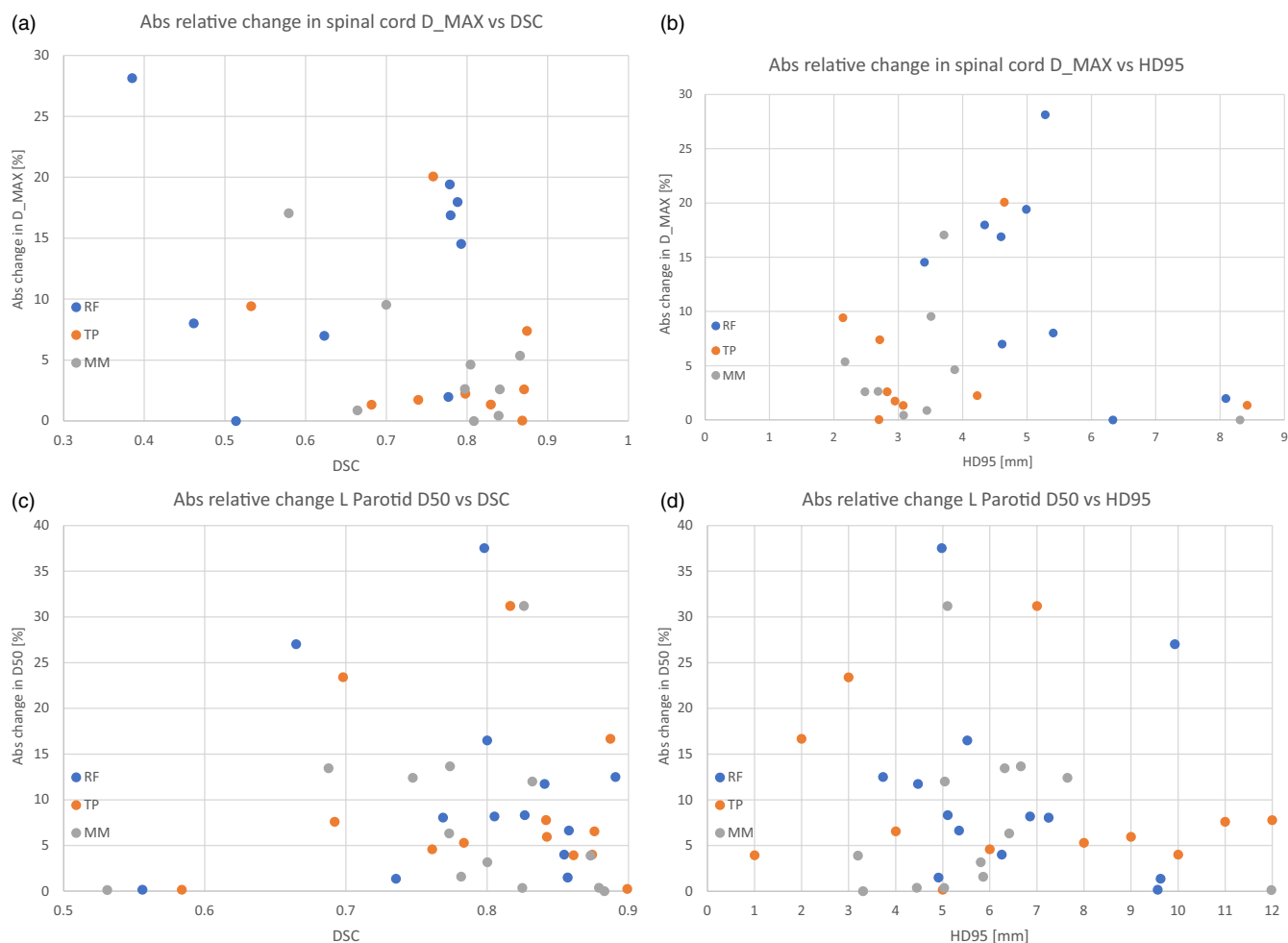


Figure 3. Comparison of absolute relative changes in dosimetry with DSC and HD95. Top: Change in spinal cord D_MAX, compared to DSC; second from Top: Change in spinal cord D_MAX, compared to HD95; Third from Top: Change in left parotid D50, compared to DSC; Bottom: Change in left parotid D50, compared to HD95.

Oral cavity, mandible and esophagus

The DSC values for these structures were very good. The HD95 for the oral cavity and esophagus were relatively high, partly due to their large size. RF and TP outperformed MM for these OARs. The MM oral cavity contours had larger volumes and extended further in the posterior and inferior directions. The caveats that were noted for clinical spinal cord contours also apply to oral cavity (a lot of institutions intentionally err on the side of a generous contour).

Conclusions

This study evaluated the geometric and dosimetric performance of three AI-based autocontouring packages in the head and neck region. The geometric agreement between the clinical contours and RF and TP was slightly better than with MM. The mean anatomical accuracy of the two (RF and TP) of the three autocontouring packages was judged to be better than the original clinical contours; the dosimetric performance of all three was very similar. Had MM been trained on previous contours of the physicians at our institution, its performance would most likely have improved.

The dosimetric effects depend on both the quality of auto contours and the dose gradients in the plan, thus the correlation between geometric and dosimetric metrics was not strong. All

three autocontouring packages can be used to generate OAR contours that can be used clinically with a modest amount of human editing, resulting in treatment planning time savings and uniformity of contouring. All autocontouring packages should be evaluated against the current contouring practice of the institution and checked for systematic differences (e.g., spinal canal vs. spinal cord, the superior/inferior level at which the cord contour ends) before being put into clinical use; dosimetric comparisons are also recommended.

Acknowledgements. None.

Financial support. None.

Competing interests. None.

References

1. 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: 401–410.
2. 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: 254–625.

3. Greenham S, Dean J, Fu CK, et al. Evaluation of atlas-based auto-segmentation software in prostate cancer patients. *J Med Radiat Sci* 2014; 61 (3): 151–158.
4. Vrtovec T, Mocnik D, Strojani P, Pernus F, Ibragimov B. Auto-segmentation of organs at risk for head and neck radiotherapy planning: From atlas-based to deep learning methods. *MedPhys* 2020; 47: E929–E950.
5. Cardenas CE, Yang J, Anderson BM, Court LE, Brock KB. Advances in autosegmentation. *Semin Radiat Oncol* 2019; 29: 185–197.
6. Lei Y, Fu Y, Wang T, et al. Deep Learning Architecture Design for Multi-Organ Segmentation. *Auto-Segmentation for Radiation Oncology*. CRC Press; 2021.
7. Fu Y, Lei Y, Wang T, Curran WJ, Liu T, Yang X. A review of deep learning based methods for medical image multi-organ segmentation. *Phys Med* 2021; 85: 107–122.
8. Valentini V, Boldrini L, Damiani A, Muren LP. Recommendations on how to establish evidence from auto-segmentation software in radiotherapy. *Radiother Oncol* 2014; 112 (3): 317–320.
9. Tao CJ, Yi JL, Chen NY, et al. Multi-subject atlas-based autosegmentation reduces interobserver variation and improves dosimetric parameter consistency for organs at risk in nasopharyngeal carcinoma: a multi-institution clinical study. *Radiother Oncol* 2015; 115: 407–411.
10. Cardenas C, Yang J, Anderson BM, Court LE, Brock KB. Advances in auto-segmentation. *Semin Radiat Oncol* 2019; 29: 185–197.
11. Claessens M, Oria CS, Brouwer C, et al. Quality assurance for AI-based applications in radiation therapy. *Semin Radiat Oncol* 2022; 32: 421–431.
12. 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: 943–947.
13. van der Veen J, Willems S, Dechuyner S, et al. Benefits of deep learning for delineation of organs at risk in head and neck cancer. *Radiother Oncol* 2019; 138: 68–74.
14. Huang J, Barbera L, Brouwers M, Browman G, Mackillop J. Does delay in starting treatment affect the outcomes of radiotherapy? A systematic review. *J Clin Oncol* 2003; 21: 555–563.
15. Chen Z, King W, Pearcey R, Kerba M, Mackillop W. The relationship between waiting time for radiotherapy and clinical outcomes: a systematic review of the literature. *Radiother Oncol* 2008; 87: 3–16.
16. Glide-Hurst CK, Lee P, Yock AD, et al. Adaptive radiation therapy (ART) strategies and technical considerations: a state of the ART review from NRG oncology. *Int J Radiat Oncol Biol Phys* 2021; 109: 1054–1075.
17. Lim-Reinders S, Keller BM, Al-Ward S, Sahgal A, Kim A. Online adaptive radiation therapy. *Int J Radiat Oncol Biol Phys* 2017; 15 (4): 994–1003.
18. Dice LR. Measures of the amount of ecologic association between species. *Ecology* 1945; 26 (3): 297–302.
19. Vaassen F, Hazelaar C, Vaniqui A, et al. Evaluation of measures for assessing time-saving of automatic organ-at-risk segmentation in radiotherapy. *Phys Imaging Radiat Oncol* 2020; 13: 1–6.
20. Goddard L, Velten C, Tang J, et al. Evaluation of multiple-vendor AI autocontouring solutions. *Radiat Oncol* 2024; 19 (1): 69.
21. Kim Y, Biggs S, Mackonis E. Investigation on performance of multiple AI-based auto-contouring systems in organs at risks (OARs) delineation. *Phys Eng Sci Med*. 2024; 47: 1123–1140.
22. Doolan P, Charalambous S, Roussakis Y, et al. A clinical evaluation of the performance of five commercial artificial intelligence contouring systems for radiotherapy. *Front Oncol* 2023; 13: 1213068.
23. Kim N, Chun J, Chang JS, Lee CG, Keum KiC, Kim JS. Feasibility of continual deep learning-based segmentation for personalized adaptive radiation therapy in head and neck area. *Cancers* 2021; 13: 1–195.
24. Ibragimov B, Xing L. Segmentation of organs-at-risks in head and neck CT images using convolutional neural networks. *Med Phys* 2017; 44: 547–554.
25. Hu Y, Nguyen H, Smith C, et al. Clinical assessment of a novel machine-learning automated contouring tool for radiotherapy planning. *J Appl Clin Med Phys* 2023; 24: e13949.
26. Bustos L, Sarkar A, Doyle L, et al. Feasibility evaluation of novel AI-based deep-learning contouring algorithm for radiotherapy. *J Appl Clin Med Phys* 2023; 24: e14090.
27. Marschner S, Datar M, Gaasch A, et al. A deep image-to-image network organ segmentation algorithm for radiation treatment planning: principles and evaluation. *Radiat Oncol* 2022; 17: 129.