Physics and Imaging in Radiation Oncology 19 (2021) 96–101

Contents lists available at ScienceDirect

Physics and Imaging in Radiation Oncology

journal homepage: www.sciencedirect.com/journal/physics-and-imaging-in-radiation-oncology

Deep learning auto-segmentation and automated treatment planning for trismus risk reduction in head and neck cancer radiotherapy

Maria Thor *, Aditi Iyer, Jue Jiang, Aditya Apte, Harini Veeraraghavan, Natasha B. Allgood, Jennifer A. Kouri, Ying Zhou, Eve LoCastro, Sharif Elguindi, Linda Hong, Margie Hunt, Laura Cerviño, Michalis Aristophanous, Masoud Zarepisheh, Joseph O. Deasy

Department of Medical Physics, Memorial Sloan Kettering Cancer Center, USA

ARTICLE INFO

Keywords:
Cancer
Radiation
Head neck
Deep learning
Masseter
Medial pterygoid
Mastication
Chewing
Trismus

ABSTRACT

Background and Purpose: Reducing trismus in radiotherapy for head and neck cancer (HNC) is important. Automated deep learning (DL) segmentation and automated planning was used to introduce new and rarely segmented masticatory structures to study if trismus risk could be decreased.

Materials and Methods: Auto-segmentation was based on purpose-built DL, and automated planning used our in-house system, ECHO. Treatment plans for ten HNC patients, treated with 2 Gy × 35 fractions, were optimized ($ECHO_0$). Six manually segmented OARs were replaced with DL auto-segmentations and the plans re-optimized ($ECHO_1$). In a third set of plans, mean doses for auto-segmented ipsilateral masseter and medial pterygoid (MI_{Mean}, MPI_{Mean}), derived from a trismus risk model, were implemented as dose-volume objectives ($ECHO_2$). Clinical dose-volume criteria were compared between the two scenarios ($ECHO_0$ *vs.* $ECHO_1$; $ECHO_1$ *vs.* $ECHO_2$; Wilcoxon signed-rank test; significance: $p < 0.01$).

Results: Small systematic differences were observed between the doses to the six auto-segmented OARs and their manual counterparts (median: $ECHO_1 = 6.2$ (range: 0.4, 21) Gy *vs.* $ECHO_0 = 6.6$ (range: 0.3, 22) Gy; $p = 0.007$), and the $ECHO_1$ plans provided improved normal tissue sparing across a larger dose-volume range. Only in the $ECHO_2$ plans, all patients fulfilled both MI_{Mean} and MPI_{Mean} criteria. The population median MI_{Mean} and MPI_{Mean} were considerably lower than those suggested by the trismus model ($ECHO_0$: $MI_{Mean} = 13$ Gy *vs.* ≤42 Gy; $MPI_{Mean} = 29$ Gy *vs.* ≤68 Gy).

Conclusions: Automated treatment planning can efficiently incorporate new structures from DL auto-segmentation, which results in trismus risk sparing without deteriorating treatment plan quality. Auto-planning and deep learning auto-segmentation together provide a powerful platform to further improve treatment planning.

1. Introduction

The majority of squamous-cell head and neck cancer (HNC) patients present with locally advanced disease (LA-HNC) [1], which has a high relapse probability with local control rates of 15–40% and a poor prognosis with only 50% of patients surviving up to five years after completed treatment [2]. In addition to being curative, treatments should be driven by individualization preserving organ function and quality of life [1]. In radiotherapy (RT) for LA-HNC, this could be accomplished by assigning patient-specific dose-volume objectives for normal tissue complication probabilities (NTCPs) and tumor control probabilities (TCPs) for a wide range of organs at risk (OARs) and tumor volumes [3,4]. However, such an approach requires automation to allow for widespread adoption within a clinically feasible time frame.

While automation should concentrate on the entire RT workflow, a recent study has demonstrated that the pressing need in radiation oncology lies within segmentation and treatment planning [5]. Further, introducing new OARs with associated dose-volume criteria in HNC, a tumor site that already includes a large collection of OARs and target volumes, would add substantial load both in terms of manual contouring and planning as new dose-volume objectives require more parameters to tweak. State-of-the-art auto-segmentation algorithms for HNC OARs [6] and tumor volumes [6,7] have demonstrated physician level accuracy and considerable time savings as opposed to manual segmentation.

* Corresponding author at: Dept. of Medical Physics, Memorial Sloan Kettering Cancer Center, 485 Lexington Ave., New York, NY 10017, USA.
E-mail address: thorm@mskcc.org (M. Thor).

https://doi.org/10.1016/j.phro.2021.07.009
Received 8 April 2021; Received in revised form 15 July 2021; Accepted 16 July 2021
Available online 28 July 2021
2405-6316/© 2021 The Authors. Published by Elsevier B.V. on behalf of European Society of Radiotherapy & Oncology. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Auto-segmentation is more reproducible than manual segmentation, is likely to provide more consistent segmentations, and has the potential to improve outcomes of clinical trials [8]. Based on the RTOG 0617 trial data, Thor et al. [8] found the trial's manual heart segmentations to vary considerably in the superior-inferior direction compared to deep learning auto-segmented hearts, and the dose to the auto-segmented hearts was significantly higher. Similarly to auto-segmentation, automated treatment planning has proven to be efficient and accurate: In the prospective HNC study by Voet et al. [9], the treating physician chose automatically generated over manually optimized treatment plans in 97% of cases, and others have recently found automated plans to have a 67% superior success rate [10].

In this HNC-focused work, auto-segmentation was combined with automated treatment planning to study whether new masticatory OARs could be introduced and trismus risk decreased. Automated treatment planning was utilizing our in-house treatment planning system [11] and trimsus risk was assessed from a published model [12].

2. Material and methods

Data for ten patients on the Institutional Review Board Approved #16-422 study were included. These patients had previously been treated using three-phase intensity-modulated RT to 70 Gy in 2 Gy fractions for LA-HNC of the oropharynx, and were specifically selected for the current study to explore the usability of the trismus risk model.

2.1. Deep learning auto-segmentation

A total of eight OARs were segmented de novo using a combination of two deep learning auto-segmentation methods. Six of these OARs (brainstem (BS), mandible, parotid glands (PGs), and submandibular glands (SGs)) were generated using a local block-wise self-attention Unet (UnetSA) method developed in-house [13] while the two remaining OARs (ipsilateral masseter and ipsilateral medial pterygoid, M_i and MP_i) were generated using the DeepLabV3+ method with the resnet-101 backbone [14]. The reported performance of these two algorithms is of similar magnitude as assesseed in hold-out validation data, *e.g.*, the population average DSC was 0.84–0.85 ± 0.04–0.05 for the PGs [13], 0.87 ± 0.02 for the left and right masseters and 0.81 ± 0.03 for the left and right medial pterygoids [14]. The UnetSA OARs had previously been trained on manually segmented OARs defined by different HNC treatment planners in a total of 48 treatment planning CT scans [13] while the training OARs for the DeepLabV3+ M and MP had been segmented post-treatment by a single radiation oncology resident in another dataset consisting of 148 treatment planning CT scans [12]. The two segmentation pipelines are executed in batch mode and requires on average one to two minutes per patient.

2.2. Automated treatment planning

For the purpose of this study, all plans were re-optimized using our in-house developed automated treatment planning system, the Expedited Constrained Hierarchical Optimization (ECHO) [11], followed by dose calculation using Varian Eclipse v.15.0. Within ECHO, all critical clinical criteria are enforced as hard constraints, and the desirable goals are optimized sequentially by solving multiple constrained optimization problems. Briefly, in the initial step, tumor dose-volume criteria are emphasized, and then transformed into constraints for further steps. In the next step, constrained by the achieved tumor dose characteristics, OAR objectives are optimized. In a final step, subject to prior tumor and OAR dose characteristics, delivery-related criteria and dose characteristics outside the tumor region are optimized. In the current study, all plans used IMRT as the treatment modality with the original beam configuration and enabled jaw-tracking.

ECHO is currently implemented clinically for the treatment of oligometastases, paraspinal tumors [15], and prostate cancer, but is not yet fully available for HNC, and by the time of conducting this study only the last third phase (the tumor boost plan) had been integrated with ECHO as a guidance during the optimization process to inform planners regarding the remaining two phases. The included ten patients were previously used to develop this pipeline. Given that the prescribed dose to the concerned third phase was 20 Gy, all dose-volume criteria were scaled down with a factor of 3.5 (70 Gy prescription/20 Gy = 3.5). The dose-volume criteria considered are depicted in Table S1 both for the total prescription and the third phase scaled prescription.

In addition to the ECHO reference plan ($ECHO_0$) in which all manually OARs, previously defined by the case-specific HNC treatment planners, were used, two additional treatment plans were optimized via ECHO for each patient: in the first plan ($ECHO_1$), the manually segmented BS, mandible, PGs, and SGs were replaced with their UnetSA counterparts. In the second plan ($ECHO_2$), we introduced DeepLabV3+-based segmentations of M_i and MP_i along with objective functions for their mean doses. The criteria applied to M_i and MP_i mean doses resulted from NTCP modeling of trismus in 421 HNC internal patients and corresponded to a 10% mild trismus rate [12], and while these were proposed for internal use three years ago, they have not been widely adopted given the added manual segmentation load required prior to our auto-segmentation solution.

2.3. Comparisons

To assess the impact of substituting already existing manually segmented OARs with auto-segmented OARs for the six UnetSA OARs, which is the underlying motivation behind the current study, dose was compared between $ECHO_0$ and $ECHO_1$. To further study the suitability of the M_i and MP_i criteria since they are not being used for treatment planning in our clinic currently, the two DeepLabV3+ OARs, M_i and MP_i, $ECHO_1$ and $ECHO_2$ were instead compared considering $ECHO_1$ the reference treatment plan. It is worth emphasizing that the dose was re-optimized replacing the manual OARs with the auto-segmented OARs. In all dose comparisons, a Wilcoxon signed-rank test was used and significance was denoted at $p < 0.01$, which was Bonferroni-corrected for four comparisons, *i.e.* $ECHO_0$ *vs.* $ECHO_1$ and $ECHO_1$ *vs.* $ECHO_2$ with two separate comparisons for the auto-segmented OARs and the remaining manually segmented OARs and PTV. In all dose comparisons, the clinical criteria were compared (Table S1).

The six UnetSA OARs for which there was also manually segmented OARs were compared geometrically with these manual OARs using the volumetric Dice Similarity Coefficient (DSC_v), the 95th percentile of the Hausdorff distance ($Hausdorff_{95}$), and centroid distances in axial, coronal and sagittal planes ($CENT_{Axial}$, $CENT_{Sag}$ and $CENT_{Cor}$). The centroid distance is that between the centers of mass of two compared segmentations, and its inclusion here was motivated by $CENT_{Axial}$ that was recently shown to have the strongest correlation with dose differences between automated and manually segmented trial hearts among 18 investigated volume similarity metrics in the RTOG 0617 clinical trial data [8]. $CENT_{Sag}$ and $CENT_{Cor}$ were included to allow for comparison of centroid distances in all three planes. The UnetSA OARs and their manual counterparts in addition to M_i and MP_i were also qualitatively evaluated relative to the anatomy in which any deviations were recorded on a slice-by-slice basis.

Lastly, for the six UNET OARs in $ECHO_1$, the volume similarity metrics were associated with the observed dose differences between the manual and the UnetSA OARs ($|Dose_{Manual}\text{-}Dose_{UnetSA}|$) using linear regression in which significance was denoted at $p < 0.003$ (corrected for six structures*three volume similarity metrics). Similar to the comparison above, dose metrics used in defining the clinical dose-volume criteria were compared (Table S1).

3. Results

3.1. Treatment plans based on auto-segmented OARs emphasize overall normal tissue dose sparing

All $ECHO_0$ and $ECHO_1$ plans fulfilled the BS and PGs criteria while eight, two, and one $ECHO_0$ plans and eight, one, and zero $ECHO_1$ plans fulfilled the contralateral SG, mandible, and ipsilateral SG criteria, respectively (Table 1). At the clinical criteria, doses to the auto-segmented OARs in the $ECHO_1$ plans were systematically lower than doses to the manual OARs in the $ECHO_0$ plans, but the differences were small (median (range) for all six OARs combined: $ECHO_1$ = 6.2 (0.4, 21) Gy *vs.* $ECHO_0$ = 6.6 (0.3, 22) Gy; p = 0.007). Outside of the compared dose-volume criteria, normal tissue sparing was improved in the $ECHO_1$ plans in particular for BS, ipsilateral PG, and contralateral SG in addition to slightly more consistent OAR doses across all patients and for all six OARs (Fig. 1). The dose-volume criteria for the remaining structures, for which no auto-segmentations were available, were adhered to with a similar extent and, while tumor homogeneity was somewhat improved in the $ECHO_1$ plans, no statistically significant differences were observed (Table S2; Fig. S1).

All 30 optimized treatment plans except for one $ECHO_0$ plan and one $ECHO_1$ plan for the same patient fulfilled the M_i and MP_i mean dose criteria by a considerable margin (M_i = 3.6 (1.2–9.1) Gy *vs.* criterion = 12 Gy; MP_i = 8.8 (1.6–18.9) Gy *vs.* criterion = 19.4 Gy). However, only in the $ECHO_2$ plans all patients fulfilled the M_i and MP_i mean dose criteria: the MP_i mean dose was reduced from 20 Gy to 18.9 Gy in the patient not previously fulfilling this criterion. The $ECHO_2$ plans provided normal tissue sparing to a larger extent also outside of the specified M_i and MP_i mean dose criteria compared to the $ECHO_1$ plans as illustrated in the narrower DVH bounds in the lower panel of Fig. 1. No statistically significant dose differences were established between either the auto-segmented OARs or the remaining organs (p = 0.06, 0.49).

Table 1
Population median (range) doses for the eight studied OARs and all combined at the clinical max and mean dose-volume criteria. In $ECHO_0$, optimization was performed based on manual segmentations, and in $ECHO_1$, the manually segmented brainstem, mandible, parotid glands and submandibular glands were replaced with the corresponding UnetSA OARs, and in $ECHO_2$, the DeepLabV3+ OARs (ipsilateral masseter and medial pterygoid) were also included in addition to the UnetSA OARs. Note: In the second column, doses in parenthesis indicate acceptable levels. * Automated DeepLabV3+ OARs inserted after optimization.

Organ	Prescription: 20 Gy	$ECHO_0$	$ECHO_1$	$ECHO_2$
	Clinical criteria	*Median (Range)*	*Median (Range)*	*Median (Range)*
Brainstem	Max Dose ≤15.4 Gy (≤17.1 Gy)	4.5 (0.3, 13) Gy	5.6 (0.4, 15) Gy	5.7 (0.3, 14) Gy
Mandible	Max Dose ≤20 Gy	21 (15, 22) Gy	21 (15, 22) Gy	21 (14, 21) Gy
Masseter	Ipsilateral: Mean Dose ≤12 Gy	3.7 (1.2, 9.0) Gy*	3.6 (1.3, 9.2) Gy	3.4 (1.2, 9.0) Gy
Medial Pterygoid	Ipsilateral: Mean Dose ≤19.4 Gy	8.2 (1.6, 20) Gy*	9.4 (1.7, 20) Gy	9.6 (1.6, 19) Gy
Parotid gland	Contralateral: Mean Dose ≤2.9 Gy	1.2 (0.4, 2.2) Gy	1.3 (0.4, 2.1) Gy	1.3 (0.4, 2.8) Gy
	Ipsilateral: Mean Dose ≤5.7 Gy (≤7.4 Gy)	2.2 (1.4, 4.7) Gy	2.5 (1.7, 3.1) Gy	2.6 (1.7, 3.1) Gy
Submandibular gland	Contralateral: Mean Dose ≤7.4 Gy	6.6 (1.6, 19) Gy	6.2 (1.8, 18) Gy	6.3 (1.8, 18) Gy
	Ipsilateral: Mean Dose ≤11.1 Gy	18 (10, 20) Gy	19 (12, 20) Gy	19 (12, 20) Gy
All eight OARs	All criteria above	5.5 (0.3, 22) Gy	5.6 (0.4, 22) Gy	5.5 (0.3, 21) Gy

3.2. Auto-segmented and manually segmented OARs differ primarily in the superior-inferior direction

Even though the auto-segmented OARs were not post-processed, the similarity metrics indicated that they were comparable to the manually segmented OARs for the majority of patients and structures: the population median DSC_v was 0.85 (range: 0.32–0.96), and the population median $Hausdorff_{95}$ was 0.41 (0.10–3.61) cm (Fig. 2). Smaller structures had overall lower DSC_v compared to larger structures (SGs: DSC_v = 0.77 (0.47–0.79) *vs.* the five remaining OARs: DSC_v = 0.86 (0.32–0.96)), which is expected given that DSC_v is volume dependent and more forgiving for larger structures. The lowest DSC and highest $Hausdorff_{95}$ was due to one manually segmented brainstem that did not extend sufficiently in the superior-inferior direction.

Interestingly, the centroid distances indicated that differences between the auto-segmented and the manual OARs were primarily located in the axial plane as the $CENT_{Axial}$ distances were considerably larger than $CENT_{Sag}$ and $CENT_{Cor}$ (median (range): $CENT_{Axial}$ = 0.10 (0.0–2.8) cm; $CENT_{Sag}$ = 0.05 (0.0–0.3) cm and $CENT_{Cor}$ = 0.05 (0.0–0.6) cm). One extreme example was a BS, in which the manual version was cropped almost to 50% of its inferior-superior extension and in which $CENT_{Axial}$ was 2.8 cm compared to a $CENT_{Sag}$ of 0.01 cm and a $CENT_{Cor}$ of 0.64 cm. The auto-segmented BS, PGs, and SGs better adhered to the anatomy than did their manual counterparts in the vast majority of patients (8/10, 7/10, and 6/10 patients, respectively). The auto-segmented mandible captured the anatomy better or equally well as the manual mandible in 7/10 patients. All auto-segmented M_i and MP_i followed the anatomy, which was probably facilitated by the generous intensity gradients given the nearby bony anatomy including *e.g.*, the mandible, and only minor anterior/posterior extensions could have been performed in the border slices of two masseters and one medial pterygoid.

3.3. The axial centroid distance captures the dose difference between manually and deep learning auto-segmented OARs but correlations are weak

The correlation between the volume similarity metrics and $|Dose_{Manual}\text{-}Dose_{UnetSA}|$ for the six UnetSA OARs was overall weak and no significant linear association was established (significance: $p < 0.003$; p-value range: 0.02–0.64). Across all OARs, the stronger associations with $|Dose_{Manual}\text{-}Dose_{UnetSA}|$ were observed with $CENT_{Axial}$ with median R^2 = 0.15 (range: 0.09–0.24) and p = 0.15 (range: 0.09–0.20). The corresponding median across DSC, $Hausdorff_{95}$, $CENT_{Sag}$ and $CENT_{Cor}$ was R^2 = −0.06 (range: −0.09, 0.44) and p = 0.50 (0.02–0.64)).

4. Discussion

This automation proof-of-concept HNC planning study has demonstrated that automated treatment planning combined with auto-segmented OARs results in comparable level of normal tissue dose sparing as that of using manually segmented OARs, but with an additional emphasis on dose sparing outside the optimized and evaluated specified clinical dose-volume criteria. Further, introducing auto-segmentations for two novel OARs provides plans of similar quality, which is important in order to advance new science in terms of dose-volume criteria and associated OARs into clinical routine more rapidly compared to the pace of traditionally non-automated approaches.

At the dose-volume criteria, dose for the six auto-segmented OARs in the $ECHO_1$ plans were systematically different compared to the dose to their manually segmented counterparts in the $ECHO_0$ plans, these differences were small and importantly sizeable sparing was observed outside the criteria in the $ECHO_1$ plans compared to in the $ECHO_0$ plans (Fig. 1). For the remaining non-auto-segmented structures, criteria were adhered to with a similar extent, and while tumor dose homogeneity was somewhat improved in the $ECHO_1$ plans no systematic difference was

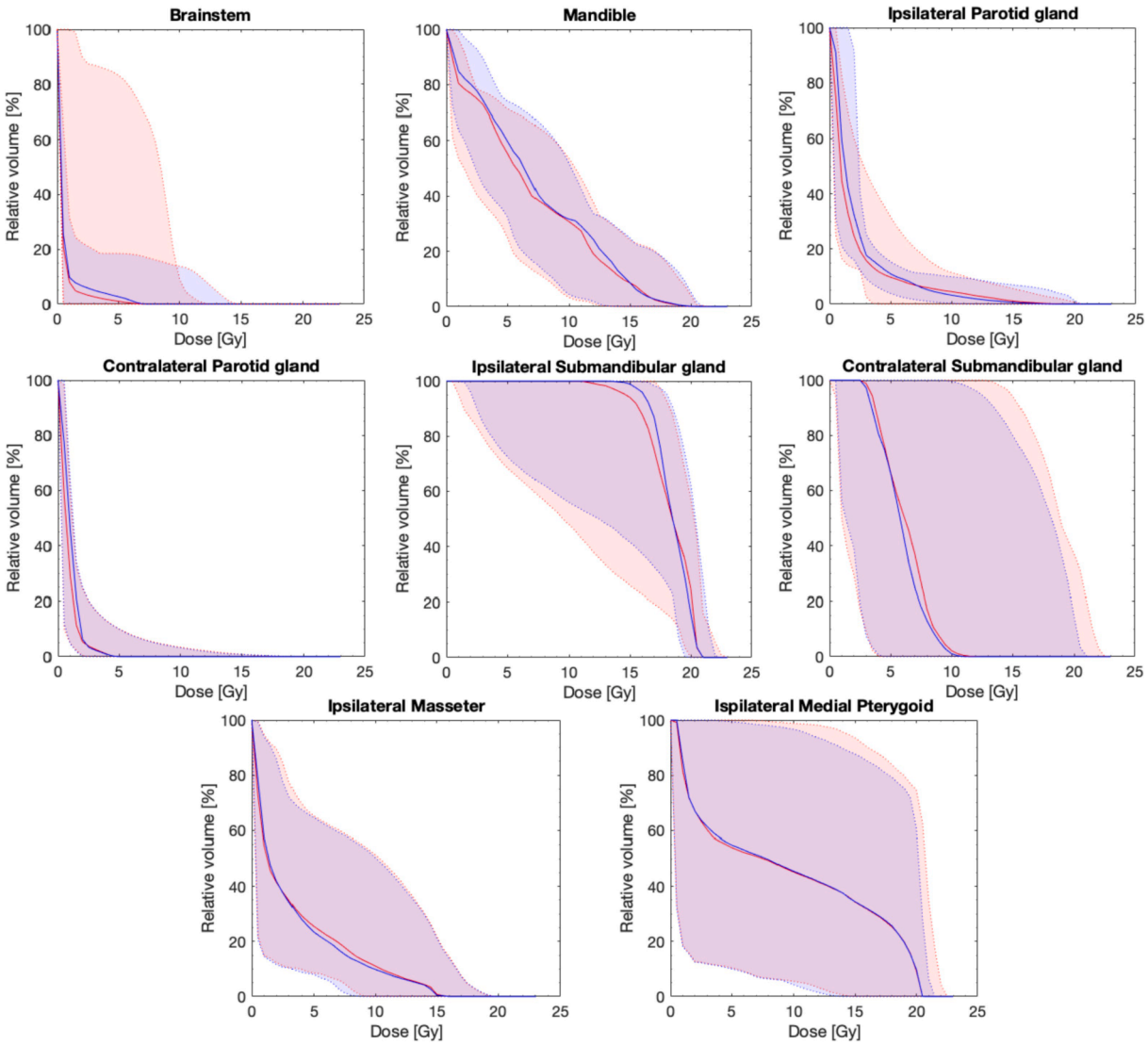

Fig. 1. Population median DVHs for the eight auto-segmented OARs (uncertainty bounds: population ranges). Two upper panels: DVHs for the manually segmented OARs in the $ECHO_0$ plans (red) and the UnetSA OARs in the $ECHO_1$ plans (blue). Lower panel: DVHs for the $ECHO_1$ (red) and $ECHO_2$ (blue) plans based on the DeepLab OARs. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

observed. In one of the few published HNC studies that have focused on dose differences between auto-segmented OARs and their manual correspondents and in which dose was, similarly as in this study, re-optimized and arose from automated treatment planning, van Rooij et al. [16] did not observe a similar dose sparing pattern as demonstrated here. Instead, they observed significantly higher constrictor and esophagus doses based on their auto-segmentations ($p = 0.005$, 0.002) [16]. It should be pointed out that the performance of our auto-segmentations is similar for the PGs (left, right: 0.85 ± 0.04, 0.85 ± 0.06 *vs.* 0.83 ± 0.03, 0.83 ± 0.02) and SGs (left, right: 0.78 ± 0.08, 0.76 ± 0.12 *vs.* 0.82 ± 0.07, 0.81 ± 0.13), but our BS DSCs are considerably higher (0.82 ± 0.18 *vs.* 0.64 ± 0.16). Another and likely important distinction is that knowledge-based planning (KBP) was used to generate their automated treatment plans. Since KBP is based on a library of existing treatment plans, the ultimate treatment plan quality will be determined based on the quality of the plans in the library with available range in anatomy, beams, image quality, *etc.* An analogy of the limitation with such a library approach can be made with atlas-based auto-segmentation [6]. Our automated treatment planning system, ECHO, does not rely on a library of plans but instead operates via constrained hierarchical optimization in which hard constraints are first fulfilled (typically tumor coverage and max dose OARs) followed by prioritizing desirable 'soft' clinical dose-volume criteria in a three-step approach [11].

Even without constraining the M_i mean dose $\leq$12 Gy and MP_i mean dose $\leq$19.4 Gy, the majority of treatment plans fulfilled these criteria and the population median doses were 3.7 Gy and 8.2 Gy (Fig. S2). The M_i mean dose $\leq$12 Gy and MP_i mean dose $\leq$19.4 Gy criteria were originally proposed to prevent mild trismus ($\geq$Grade 1) to exceed 10% in a previously treated cohort [12], but our findings indicate that they are not well calibrated, and more specifically they are too generous. Converted to the current fractionation scheme (assuming $\alpha/\beta = 3$ Gy) and scaled down to the 20 Gy prescription, the univariate linear regression relationship derived by Beasley et al. [17] suggests M_i mean dose $\leq$6.2 Gy (mouth opening $\geq$45 mm). At a 10% predicted trismus risk, the model by Lindblom et al. [18] proposes M_i mean dose $\leq$2.6 Gy or $\leq$5.1 Gy (mouth-opening $\leq$35 mm, patient-reported trismus) while at a same 10% risk level, the model by Kraijengaa et al. [19] (mouth-opening $\leq$35 mm; baseline mouth-opening $\leq$46 mm) indicates M_i mean dose $\leq$4.9 Gy. At a median, six of the ten patients included here met these four published M_i mean dose levels for which the median value was 5.0 Gy, which scaled up to the 70 Gy prescription corresponds to 18

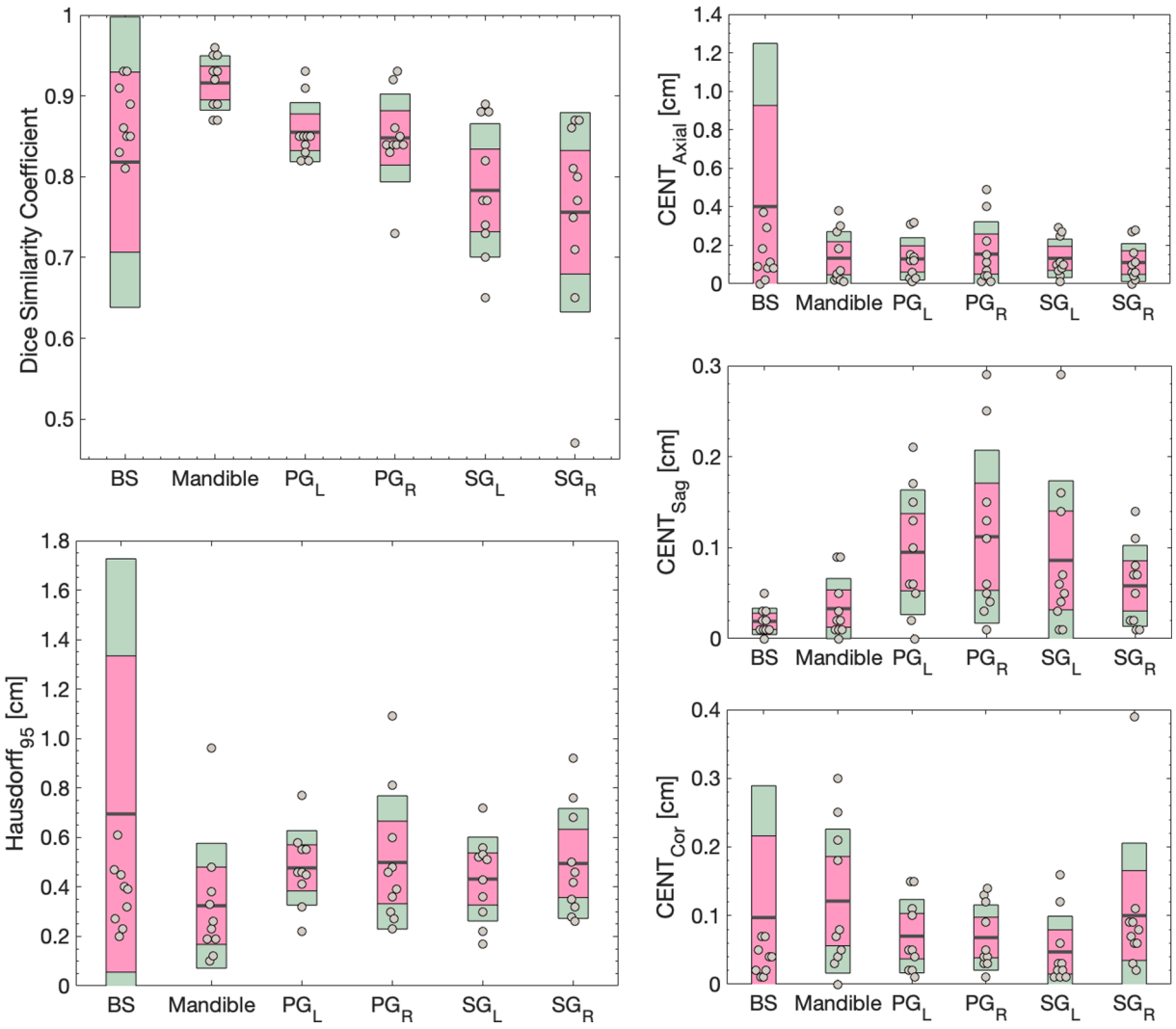

Fig. 2. Box plots for the volume similarity metrics between the six UnetSA OARs and their manual counterparts. Note: One brainstem data point has been excluded to improve visualization except for in the $CENT_{Sag}$ subfigure (excluded data point located at: $DSC_v = 0.32$; $Hausdorff_{95} = 3.61$; $CENT_{Axial} = 2.79$ cm; $CENT_{Cor} = 0.64$ cm). Abbreviations: BS: Brainstem; L: Left; PG: Parotid gland; R: Right; SG: Submandibular gland.

Gy, which can be compared to our current 42 Gy criterion. Kraijengaa et al. [19] further suggests MP_i mean dose ≤53 Gy (20 Gy prescription: 15.1 Gy), which was also fulfilled by six of our patients, and is considerably lower than our current 68 Gy MP_i mean dose. Taken together, we are currently updating our criteria to M_i mean dose <18 Gy and MP_i mean dose <53 Gy, but we will retain the previous criteria as upper bounds in situations where the new criteria cannot be met also given that the current study is limited to ten cases and includes only the tumor boost plan and not the elective nodal volumes, and we will closely monitor the level of adherence and revise accordingly.

Only a weak correlation was identified between the volume similarity metrics and $|Dose_{Manual}\text{-}Dose_{UnetSA}|$. The small sized dataset including only 10 patients may have been a limitation to illustrate a clear association. Again, dose was re-optimized for the new set of auto-segmentations. Despite the overall weak correlations, $CENT_{Axial}$ better explained the dose differences between paired structures than did DSC or $Hausdorff_{95}$. These results are similar to the findings in [8] in which 18 volume similarity metrics between clinical trial hearts and auto-segmented hearts were associated with three heart dose metrics differences and the strongest associations were observed using $CENT_{Axial}$ ($CENT_{Axial}$ *vs.* DSC_V and $Hausdorff_{95}$: $R^2 = 0.44$–0.51 *vs.* 0.32–0.51 and 0.33–0.57). In addition to provide the global similarity metrics such as DSC_V and $Hausdorff_{95}$, we suggest to also report the centroid distances in all three planes. This would provide a quick quantitative indicator to the direction of disagreements likely in need of further quality assurance and potentially post-processing.

In summary, we have demonstrated that automated treatment planning for HNC based on deep learning auto-segmented OARs is possible, and results in plans emphasizing overall normal tissue dose sparing also outside the specified clinical dose-volume criteria. In addition, incorporating new dose-volume criteria for novel auto-segmented masticatory OARs into this framework has proven feasible and provided a rapid evaluation and refinement of dose-volume criteria to reduce trismus risk. The two deep learning auto-segmentation algorithms used have been packaged using a singularity container [20] for HNC OAR segmentation, which is open-source and available via our CERR model library [21]. Internally, this HNC OAR auto-segmentation container is currently deployed with our treatment planning system to facilitate the planners with an OAR segmentation that only requires quality assurance and/or minor editing. Work is on-going to fully expand our automated ECHO treatment planning system for HNC.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence

the work reported in this paper.

Acknowledgment

This research was funded in part through the NIH/NCI Cancer Center Support Grant P30 CA008748.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.phro.2021.07.009.

References

[1] Longo DL, Chow LQM. Head and neck cancer. New Engl J Med 2020;382:60–72. https://doi.org/10.1056/NEJMra1715715.
[2] Braakhuis BJ, Brakenhoff RH, Leemans CR. Treatment choice for locally advanced head and neck cancers on the basis of risk factors: biological risk factors. Ann Oncol 2012;23(Suppl 10):173–7. https://doi.org/10.1093/annonc/mds299.
[3] Allen Li X, Alber M, Deasy JO, Jackson A, Ken Jee KW, Marks LB, et al. The use and QA of biologically related models for treatment planning: short report of the TG-166 of the therapy physics committee of the AAPM. Med Phys 2012;39:1386–409. https://doi.org/10.1118/1.3685447.
[4] Langendijk JA, Lambin P, De Ruysscher D, Widder J, Bos M, Verheij M. Selection of patients for radiotherapy with protons aiming at reduction of side effects: the model-based approach. Radiother Oncol 2013;107:267–73. https://doi.org/10.1016/j.radonc.2013.05.007.
[5] 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. https://doi.org/10.1016/j.phro.2020.11.002.
[6] 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. https://doi.org/10.1016/j.radonc.2019.03.004.
[7] Cardenas CE, Beadle BM, Garden AS, Skinner HD, Yang J, Joo Rhee D, et al. Generating high-quality lymph node clinical target volumes for head and neck cancer radiotherapy using a fully automated deep learning-based approach. Int J Radiat Oncol Biol Phys 2021;109:801–12. https://doi.org/10.1016/j.ijrobp.2020.10.005.
[8] 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. https://doi.org/10.1016/j.ijrobp.2020.11.011.
[9] Voet PWJ, Dirkx MLP, Breedveld S, Fransen D, Levendag PC, Heijmen BJM. Toward fully automated multicriterial plan generation: a prospective clinical study. Int J Radiat Oncol Biol Phys 2013;85:866–72. https://doi.org/10.1016/j.ijrobp.2012.04.015.
[10] Cornell M, Kaderka R, Hild SJ, Ray XJ, Murphy JD, Atwood TF, et al. Noninferiority study of automated knowledge-based planning versus human-driven optimization across multiple disease sites. Int J Radiat Oncol Biol Phys 2020;106:430–9. https://doi.org/10.1016/j.ijrobp.2019.10.036.
[11] Zarepisheh M, Hong L, Zhou Y, Oh JH, Mechalakos JG, Hunt MA, et al. Automated intensity modulated treatment planning: The expedited constrained hierarchical optimization (ECHO) system. Med Phys 2019;46:2944–54. https://doi.org/10.1002/mp.13572.
[12] Rao SD, Saleh ZH, Setton J, Tam M, McBride SM, Riaz N, et al. Dose-volume factors correlating with trismus following chemoradiation for head and neck cancer. Acta Oncol 2016;55:99–104. https://doi.org/10.3109/0284186X.2015.1037864.
[13] Berry S, Jiang J, Elguindi S, Hunt M, Deasy J, Veeraraghavan H. Self-attention based deep learning probabilistic parotid gland segmentation quality evaluation using dose volume histogram. Med Phys 2019;46:e383. https://doi.org/10.1002/mp.13589.
[14] Iyer A, Thor M, Haq R, Deasy JO, Apte AP. Deep learning-based auto-segmentation of swallowing and chewing structures. bioRxiv 2019. https://doi.org/10.1101/772178.
[15] Hong L, Zhou Y, Yang J, Mechalakos JG, Hunt MA, Mageras GS, et al. Clinical experience of automated SBRT paraspinal and other metastatic tumor planning with constrained hierarchical optimization. Adv Radiat Oncol 2020;5:1042–50. https://doi.org/10.1016/j.adro.2019.11.005.
[16] van Rooij W, Dahele M, Ribeiro Brandao H, Delaney AR, Slotman BJ, Verbakel WF. Deep learning-based delineation of head and neck organs at risk: geometric and dosimetric evaluation. Int J Radiat Oncol Biol Phys 2019;104:677–84. https://doi.org/10.1016/j.ijrobp.2019.02.040.
[17] Beasley W, Thor M, McWilliam A, Green A, Mackay R, Slevin N, et al. Image-based data mining to probe dosimetric correlates of radiation-induced trismus. Int J Radiat Oncol Biol Phys 2018;102:1330–8. https://doi.org/10.1016/j.ijrobp.2018.05.054.
[18] Lindblom U, Gärskog O, Kjellén E, Laurell G, Levring Jäghagen E, Wahlberg P, et al. Radiation-induced trismus in the ARTSCAN head and neck trial. Acta Oncol 2014;53:620–7. https://doi.org/10.3109/0284186X.2014.892209.
[19] Kraaijenga SA, Hamming-Vrieze O, Verheijen S, Lamers E, Molen L, Hilgers FJ, et al. Radiation dose to the masseter and medial pterygoid muscle in relation to trismus after chemoradiotherapy for advanced head and neck cancer. Head Neck 2019;41:1387–94. https://doi.org/10.1002/hed.v41.510.1002/hed.25573.
[20] Kurtzer GM, Sochat V, Bauer MW, Gursoy A. Singularity: Scientific containers for mobility of compute. PLoS ONE 2017;12:e0177459. https://doi.org/10.1371/journal.pone.0177459.
[21] Apte AP, Iyer A, Thor M, Pandya R, Haq R, Jiang J, et al. Library of deep-learning image segmentation and outcomes model-implementations. Phys Med 2020;73:190–6. https://doi.org/10.1016/j.ejmp.2020.04.011.