

# Deep learning vs. atlas-based models for fast auto-segmentation of the masticatory muscles on head and neck CT images

**Wen Chen**

University of California Davis Comprehensive Cancer Center

**Brandon A Dyer**

University of Washington

**Xue Feng**

carina medical LLC

**Yimin Li**

The First Affiliated Hospital of Xiamen University

**Shyam Rao**

University of California Davis Comprehensive Cancer Center

**Stanley H. Benedict**

University of California Davis Comprehensive Cancer Center

**Quan Chen**

University of Kentucky

**Yi Rong** (✉ [yrong@UCDAVIS.EDU](mailto:yrong@UCDAVIS.EDU))

University of California Davis Comprehensive Cancer Center <https://orcid.org/0000-0002-2620-1893>

---

## Research

**Keywords:** deep learning model, masticatory muscles, auto-segmentation

**Posted Date:** April 27th, 2020

**DOI:** <https://doi.org/10.21203/rs.3.rs-23941/v1>

**License:**   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

---

**Version of Record:** A version of this preprint was published on July 20th, 2020. See the published version at <https://doi.org/10.1186/s13014-020-01617-0>.

# Abstract

**Background:** Trismus is caused by impaired function of masticatory muscles. Routine delineation of these muscles during planning may improve dose tracking and facilitate dose reduction resulting in decreased radiation-related trismus. This study aimed to compare a deep learning model vs. a commercial atlas-based model for fast auto-segmentation of the masticatory muscles on head and neck computed tomography (CT) images.

**Material and methods:** Paired masseter (M), temporalis (T), medial and lateral pterygoid (MP, LP) muscles were manually segmented on 56 CT images. CT images were randomly divided into training (n=27) and validation (n=29) cohorts. Two methods were used for automatic delineation of masticatory muscles (MMs): Deep learning auto-segmentation (DLAS) and atlas-based auto-segmentation (ABAS). Quantitative assessment of automatic versus manually segmented contours were performed using Dice similarity coefficient (DSC), recall, precision, Hausdorff distance (HD), HD95, and mean surface distance (MSD). The interobserver variability in manual segmentation of MMs was also evaluated. Differences in dose ( $\Delta$ Dose) to MMs for DLAS and ABAS segmentations were assessed. A paired t-test was used to compare the geometric and dosimetric difference between DLAS and ABAS methods.

**Results:** DLAS outperformed ABAS in delineating all MMs ( $p < 0.05$ ). The DLAS mean DSC for M, T, MP, and LP ranged between  $0.83 \pm 0.03$  to  $0.89 \pm 0.02$ , the ABAS mean DSC ranged between  $0.79 \pm 0.05$  to  $0.85 \pm 0.04$ . The mean value for recall, precision, HD, HD95, MSD also improved with DLAS for auto-segmentation and were close to the mean interobserver variation. With few exceptions,  $\Delta$ D99%,  $\Delta$ D95%,  $\Delta$ D50%, and  $\Delta$ D1% for all structures were below 10% for DLAS and ABAS and had no detectable statistical difference ( $P > 0.05$ ). DLAS based contours have dose endpoints more closely matched with that of the manually segmented when compared with ABAS.

**Conclusions:** DLAS auto-segmentation of masticatory muscles for the head and neck radiotherapy had improved segmentation accuracy compared with ABAS with no qualitative difference in dosimetric endpoints compared to manually segmented contours.

## Introduction

Advances in radiotherapy techniques, such as intensity modulated radiotherapy, have improved dose conformity to radiation targets, resulting in decreased dose to adjacent organs at risk (OARs)<sup>1,2</sup>. This has resulted in improved locoregional tumor control, as well as reduced incidence of late normal tissue side effects. As a result of these technologic advancements accurate and consistent delineation of tumor and OAR structures is imperative for optimal radiation planning, but such planning is time and labor intensive. Furthermore, given the complexity of head and neck cancer (HNC) anatomy intra- and inter-observer variations in manual segmentations are common and due to the substantial time required, some OARs may not be routinely contoured<sup>3-6</sup>.

The development of computational tools to automatically generate OAR contours can reduce the time and effort required for HNC contouring and plan development, as well as inter-observer contour variations. Specifically, organ auto-segmentation has been extensively studied<sup>7-10</sup> using both CT and MR image datasets<sup>11,12</sup>. One approach, atlas-based auto-segmentation (ABAS)<sup>13,14</sup>, is a traditional method for organ contouring and various factors can affect segmentation performance. These include the size of dataset used to create the atlas, approaches for image registration, and approaches for label fusion. Because the atlas size is fixed, the main limitation for ABAS is the ability to overcome variations in patient anatomy. In recent years, deep learning-based methods<sup>15,16</sup> have shown great success for biomedical image segmentation and have been introduced to the field of head and neck anatomy segmentation. However, the literature is limited in assessing masticatory muscle (MM) auto-segmentation<sup>17,18</sup>, which may be due to the lack of delineation guidelines for MMs.

Trismus, pain or difficulty with opening the mouth, is caused by impaired function of the MMs<sup>19</sup>. It may also result in poor dental hygiene, impaired chewing, malnutrition and psychological difficulties which will eventually lead to impacts on patients' health-related quality of life<sup>20,21</sup>. For HNC, trismus has been reported with an incidence of 35–55% in advanced oropharyngeal cancers<sup>22,23</sup>. Mouth opening is a complex action controlled by the synergistic actions of the paired MMs consisting of the masseter (M), temporalis (T), medial pterygoid (MP) and lateral pterygoid (LP) muscles. When the MMs are within the field of radiation, fibrosis may lead to trismus, reducing the range of movement. Therefore, in order to reduce HNC toxicities and improve quality-of-life, it is necessary to optimize radiation dose to the target and sparing the MMs. Several dosimetric studies<sup>22,24,25</sup> investigated the relationship between radiotherapy dose to MMs and trismus. Rao *et al* found that limiting the high dose volume of the ipsilateral MP to V68Gy < 10 cm<sup>3</sup> reduced swallowing toxicity<sup>22</sup>. However, no standardized MM OAR definition exists, or dose threshold for the MMs.

Previous studies evaluated the use of auto-segmentation to improve interobserver variability in contouring MMs<sup>18</sup>. However, to the best of our knowledge, this is the first paper to evaluate using a deep learning model for auto-segmentation of MMs. The aim of this study was to evaluate the feasibility and performance of deep learning auto-segmentation (DLAS) for paired MM segmentation. Auto segmentation performance using the atlas-based algorithm on a commercial software was also assessed and compared in terms of geometry and dosimetry accuracy. Furthermore, interobserver variability for MM segmentation are evaluated and compared with DLAS.

## Materials And Methods

### Imaging Data

In this study, 56 HNC patients between 2016 and 2018 were retrospectively selected under institutional review board approval. A variety of primary head and neck disease sites for patients receiving definitive

and adjuvant were included. Patients characteristics are shown in Table 1. All patients were staged according to the 8th AJCC staging system<sup>26</sup>.

Table 1  
Patients characteristics

Characteristics	Training group (n = 27)	Validation group(n = 29)
Primary site		
Oropharynx	16(59.3%)	20(69.0%)
Larynx	2(7.4%)	4(13.8%)
Nasopharynx and Sinonasal	4(14.8%)	2(6.9%)
Other sites	5(18.5%)	3(10.3%)
Stage		
I	3(11.1%)	2(6.9%)
II	3(11.1%)	3(10.3%)
III	5(18.5%)	6(20.7%)
IV	16(59.3%)	17(58.6%)
N/X	0(0%)	1(3.5%)
Primary Tumor Surgery		
Yes	15	17
No	12	12

The four paired masticatory muscles, masseter (M), temporalis (T) and medial/lateral pterygoids (MP, LP) muscles were contoured on a simulation CT scan. The volumes were then confirmed by a senior HNC radiation oncologist. All muscles were delineated using the soft tissue window and in accord with previous publication by Rao *et al*<sup>22</sup>. The CT images and segmented contours were extracted as DICOM files and uploaded to the in-house software and the commercial software available on the RayStation Treatment Planning (RaySearch Laboratory AB, Stockholm, Sweden) for further testing.

## Deep Learning Model For Image Segmentation

The deep learning based contouring software (INTContour, Carina Medical LLC, Lexington, KY) employs 3D U-Net structure<sup>27</sup> for organ segmentation. The algorithm has achieved good performance in 2017 AAPM thoracic challenge<sup>28</sup> and 2019 RT-MAC challenge. The original CT was resampled to have the same spatial resolution, matrix size and field of view. Two 3D U-Nets with and without dilated convolutions were trained and the output from both networks was averaged. Training and testing

augmentations such as random translation, rotation, scaling and left-right flipping were used to improve the model performance. The summation of the weighted cross entropy and soft Dice loss was used as the loss function. A detailed description of the segmentation method was previously published<sup>29,30</sup>. From the initial dataset of 56 patients, 27 were randomly selected for training and validation during the training process. After the model was trained, the remaining 29 patients were used for testing the performance. No model re-tuning and re-testing was performed.

## **Multi-atlas-based Auto Segmentation**

Datasets were imported in RayStation treatment planning system version 9A. Multi-atlas-based auto-segmentation algorithm (ABAS)<sup>31</sup> was used to generate contours. CT images in the training cohort (n = 27) with manual contours were used to build the atlas. For new imaging dataset, multiple atlas contours were first rigid registered to the new image to identify the best matching, which was then deformed and registered to the new CT image as the new automatic generated segmentation set. An ANAtomically Constrained Deformation Algorithm (ANACONDA) was used for image deformation in the process of ABAS in Raystation<sup>32</sup>. This algorithm uses both intensity-based and anatomic information-based approaches to calculate deformation vectors to achieve best match between images. In this study, the rest of 29 datasets was used for ABAS validation.

## **Interobserver Variability**

Five head and neck CT image sets were randomly selected for MM OAR segmentation by three physicians according to the aforementioned MM contouring guidelines. Paired MMs were segmented and interobserver variability was assessed by pairwise comparison of MM manual contours.

## **Evaluation Of Geometric Accuracy**

(See Supplemental Files)

## **Evaluation Of Dosimetric Impact Of Variation In Contouring**

(See Supplemental Files)

## **Statistical Analysis**

Analysis was performed using GraphPad Prism version 6 (Graph pad software) and SPSS software version 24.0 (SPSS Inc, Chicago, IL, USA). A paired t-test was used to compare the difference value of

DSC, recall, precision, HD, HD95%, MSD and  $\Delta$ dose between DLAS and ABAS. Statistical significance was defined as  $p < 0.05$ .

## Results

### Variation in Contouring

In all cases, both DLAS and ABAS can segment the muscles with an overall good representation. Figure 1 shows an example of the DLAS, ABAS, and manual contours. Contour variability was greatest for MP structures.

Figure 2 shows metrics of geometric and spatial similarity for all the structures manually delineated by the three clinicians. Overall, both T and MP were associated with lower values for DSC, recall, and precision compared with M and LP. Higher values for MSD and HD95/HD were seen for T and MP. Among all structures, T had the highest HD95/HD. More specifically, the mean value of DSC for M, T, LP, MP ranged between  $0.82 \pm 0.06$  to  $0.90 \pm 0.02$ , with an overall mean of  $0.86 \pm 0.05$ . The mean value ranges of HD and HD95 were  $0.42 \pm 0.08$  to  $1.46 \pm 0.85$  and  $0.20 \pm 0.03$  to  $0.40 \pm 0.17$ , respectively, with overall means of 0.82 and 0.31 (unit: cm). The mean values of MSD ranged from  $0.05 \pm 0.01$  to  $0.11 \pm 0.05$ , with an overall mean of  $0.08 \pm 0.05$  (unit: cm). The overall means of six metrics are shown in each sub-figure, which were used as the reference values for calculating scores.

Table 2 summarizes DLAS and ABAS geometrics indices for MM segmentations. DLAS was superior to ABAS for all quantitative metrics. More specifically, DSCs was  $0.86 \pm 0.03$  and  $0.83 \pm 0.04$  for DLAS and ABAS, respectively, as compared to the inter-observer variation reference of  $0.86 \pm 0.05$ . HD95 was  $0.30 \pm 0.09$  for DLAS and  $0.37 \pm 0.13$  for ABAS, as compared to the reference  $0.31 \pm 0.13$ . MSD was  $0.08 \pm 0.02$ ,  $0.11 \pm 0.03$ ,  $0.08 \pm 0.04$  for DLAS, ABAS, and reference, respectively. Overall, DLAS achieved equivalent performance compared to the mean interobserver variation for quantitative metrics, with smaller standard deviation (SD), except for precision. This indicates that DLAS is more geometrically accurate and reproducible compared to ABAS.

Table 2  
Mean values and standard deviation (Mean  $\pm$  SD) for the 6 metrics across all organs contoured.

Metrics	DLAS	ABAS	Physician Variation
DSC	0.86 $\pm$ 0.03	0.83 $\pm$ 0.04	0.86 $\pm$ 0.05
Recall	0.86 $\pm$ 0.05	0.81 $\pm$ 0.07	0.81 $\pm$ 0.07
Precision	0.85 $\pm$ 0.05	0.85 $\pm$ 0.07	0.92 $\pm$ 0.04
HD95	0.30 $\pm$ 0.09	0.37 $\pm$ 0.13	0.31 $\pm$ 0.13
HD	0.74 $\pm$ 0.39	0.83 $\pm$ 0.37	0.82 $\pm$ 0.53
MSD	0.08 $\pm$ 0.02	0.11 $\pm$ 0.03	0.08 $\pm$ 0.04

Figure 3 shows overall improvement in geometrics metrics for each pair MM when using DLAS, as compared to ABAS. Mean DSC for MM structures ranged from 0.79  $\pm$  0.05 to 0.85  $\pm$  0.04 for ABAS, and 0.83  $\pm$  0.03 to 0.89  $\pm$  0.02 for DLAS. For MM auto-segmentation structures, MP had the lowest DSC and recall value compared with other structures, and LP shows the lowest MSD value. However, T had a larger HD/HD95 value compared with other structures. This can be explained by the larger volume of T muscles. With few exceptions, paired t-test indicated that DLAS performed significantly better than ABAS for all the metrics ( $p < 0.05$ ).

The overall scores achieved by the two methods for every muscle is summarized in Fig. 4. For most muscle pairs, DLAS-generated structures had mean scores above 50 while ABAS was less than 50, indicating ABAS was inferior to the reference established base on the inter-observer variation.

Table 3 shows the percentages (%) of cases where auto-segmentation had worse performance than manual segmentation for each muscle based on mean DSC. The percentages of cases that perform worse than manual segmentation ranged from 20.7–65.5% for DLAS, and from 41.4–96.6% for ABAS. These results indicate that DLAS performance was superior compared to ABAS and that ABAS segmentations required more contour revision to achieve equipoise. Among all MM, T segmentations with either DLAS or ABAS had the fewest number of cases performing worse than that of manual segmentations.

Table 3

The percentages (%) of cases for each muscle auto-segmented by DLAS and ABAS which were worse than that achieved by physicians (mean DSC was used to compare the results)

	M_R	M_L	T_R	T_L	LP_R	LP_L	MP_R	MP_L
DLAS	62.1%	51.7%	20.7%	24.1%	65.5%	65.5%	44.8%	37.9%
	(18/29)	(15/29)	(6/29)	(7/29)	(19/29)	(19/29)	(13/29)	(11/29)
ABAS	96.6%	89.7%	48.3%	41.4%	96.6%	82.8%	79.3%	69.0%
	(28/29)	(26/29)	(14/29)	(12/29)	(28/29)	(24/29)	(23/29)	(20/29)

## Dosimetric Impact Of Variation In Contouring

Figure 5 shows dosimetric endpoints for DLAS and ABAS segmentations for paired MMs. Box plots show  $\Delta$ dose of D99%, D95%, D50%, and D1% of each muscle for DLAS and ABAS. The mean  $\Delta$ D99%,  $\Delta$ D95%,  $\Delta$ D50%, and  $\Delta$ D1% for most of the structures was less than 10%. However,  $\Delta$ D99% and  $\Delta$ D95% were large (up to 100%) in some cases, i.e. ipsilateral MMs showed a large degree of dose variation compared with the contralateral muscles. This indicates that, for the organs in a steep dose gradient and/or small organ volume, segmentation variability of several millimeters may drastically change MM dosimetric endpoints. Comparison of  $\Delta$ dose for DLAS and ABAS revealed generally similar results, and dose to MMs with DLAS more closely matched manual segmentations than did ABAS.

## Discussion

This is the first study to assess the feasibility of a deep learning method for contouring masticatory muscles in head and neck radiotherapy. Results indicate that our in-house DLAS as compared to the commercial ABAS tool provides accurate, consistent, reproducible MM contours without the need of any manual correction or user interference. Dosimetric comparison of MMs for DLAS and ABAS showed that the dose difference from that of manual contours has a minimal clinical impact with less variation and improved consistency.

To compare the interobserver variation in segmentation, structures were manual delineated by three physicians. Our results showed there is contour variation between observers. For T and MP structures the results showed more variation indicating that T and MP are more difficult to define anatomically. Prior to clinical implementation it is important to determine if these automated segmentation results fall within the variability seen with manual segmentation. Comparison of quantitative geometric indices showed that DLAS of MMs is more reproducible (less variable) than manual segmentations and dosimetric outcomes are similar.

Several studies<sup>7-9, 17,18,33</sup> previously evaluated the performance of different methods of auto-segmentation for head and neck radiotherapy. Hague *et al*<sup>18</sup> developed a new contouring atlas to

evaluate the reduction in interobserver variability for MP, LP, M, and T muscles. The authors found that an atlas reduced interobserver variability for all muscles and the mean DTA improved when the trainees used the atlas. Furthermore, they found that T had the largest reduction in variability ( $4.3 \pm 7.1$  v  $1.2 \pm 0.4$  mm,  $p = 0.06$ ), and for MP and T the distance between the center of mass (COM) and interobserver variability reduced in all directions. Our results also showed that DLAS is associated with smaller contour variation (smaller standard deviation in metrics) for all muscles compared to ABAS.

Comparison of MM dosimetric metrics by segmentation strategy (DLAS, ABAS, manual segmentation) showed small dose differences for most cases. However, some cases had large dosimetric differences in ipsilateral MM structures despite a small difference in contouring variation. This indicates that segmentation accuracy in areas with high dose and steep dose gradients is important. Furthermore, note that for well lateralized oral cavity or oropharyngeal tumors, the ipsilateral MMs may potentially receive high doses and should be given consideration during the planning process.

Previous studies<sup>17,18</sup> mainly consider geometric indices when they evaluated the performance of automated segmentation. A study by Kaderka *et al*<sup>18</sup> found that some small structures (such as anterior descending artery) may have low DSC indicating low degree of geometric similarity, yet are associated with small dosimetric impact. Compared to geometric similarity, dosimetric similarity provides more clinically relevant information in treatment plan quality evaluation, thus they are also included in our study. Similarly, it is noted that small structures, even when presented with a large variation between two contours, may not result in meaningful dosimetric differences.

There are some potential limitations in our study. Inter-observer variations were assessed based on five head and neck CT datasets and three physicians, given the time required for contouring the structures on four muscle pairs for all cases was substantial. While a larger dataset, or more observers, may help improve the power of analysis, the current study provides clear proof of principle that interobserver variability can be reduced by using DLAS for MMs. Another limitation is the definition of the reference contours which we considered as the “ground truth”. Since there are no consensus guidelines or standards for delineation of MMs, all physicians delineated the structures in accordance with the contouring guideline of the published study by Rao *et al*<sup>22</sup> which included 421 cases.

## Conclusion

In summary, to the best of our knowledge, this is the first paper that uses a deep learning model for the automatic segmentation of the mastication muscles for the purpose of optimizing radiation therapy treatment planning. We have developed a deep learning model to contour the masticatory muscles during head and neck radiotherapy planning and this strategy can be generally employed on commercial planning systems. This new method has been shown to significantly improve consistency in contouring of all masticatory muscles compared with our previous ABAS method, or manual segmentation. Use of DLAS for MM segmentation can help to reduce interobserver variability and standardize treatment volumes. It is important to note that this study identifies as critical that auto-segmentation of contours for

well lateralized oral cavity or oropharyngeal tumors during the planning phases, in which segmentation variability of several millimeters may drastically change ipsilateral MMs dosimetric endpoints.

## List Of Abbreviations

CT: computed tomography

M: masseter

T: temporalis

MP: medial pterygoid

LP: lateral pterygoid

MMs: masticatory muscles

DLAS: Deep learning auto-segmentation

ABAS: atlas-based auto-segmentation

DSC: Dice similarity coefficient

HD: Hausdorff distance

MSD: mean surface distance

OARs: organs at risk

HNC: head and neck cancer

## Declarations

### **Ethics approval and consent to participate**

Not applicable

### **Consent for publication**

The research was performed under institutional review board (IRB) approval. The images are entirely unidentifiable and there are no details on individuals reported within the manuscript.

### **Availability of data and materials**

The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

## Competing interests

XF and QC are co-founders of the Carina Medical, LLC.

## Funding

XF and QC are partially supported by NIH Grant R43EB027523 and Varian Research grant.

## Authors' contributions

YR and QC conceived and designed the study. WC, YML collected the patient's information, WC, BD, YML, SR, QC, XF performed the experiments. WC, QC and XF were responsible for the data analysis. WC drafted this manuscript. YR, BD and QC reviewed and edited the manuscript. SB and SR offered constructive suggestions for this study. All authors read and approved the final manuscript.

## Acknowledgements

Not applicable

## References

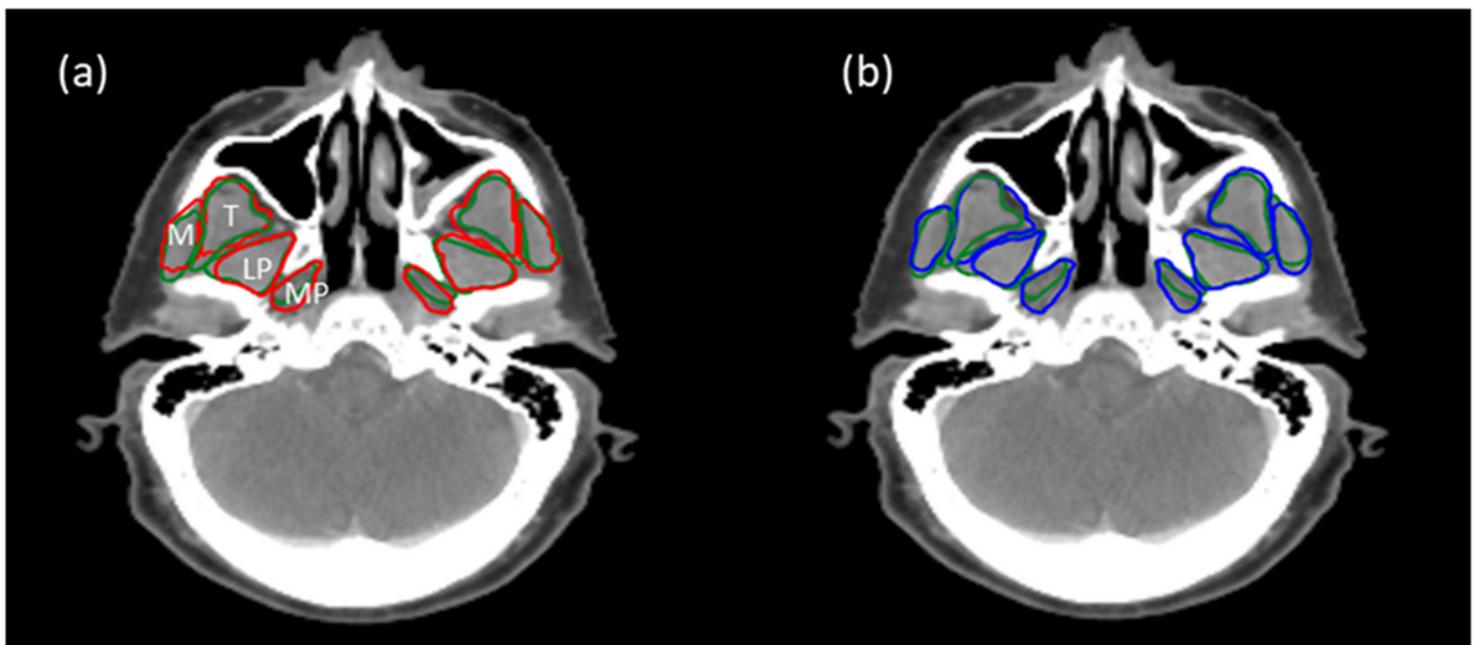
1. Mackie TR, Kapatoes J, Ruchala K, et al: Image guidance for precise conformal radiotherapy. *Int J Radiat Oncol Biol Phys* 56:89-105, 2003
2. Gomez-Millan J, Fernandez JR, Medina Carmona JA: Current status of IMRT in head and neck cancer. *Rep Pract Oncol Radiother* 18:371-5, 2013
3. Brouwer CL, Steenbakkers RJ, van den Heuvel E, et al: 3D Variation in delineation of head and neck organs at risk. *Radiat Oncol* 7:32, 2012
4. Peng YL, Chen L, Shen GZ, et al: Interobserver variations in the delineation of target volumes and organs at risk and their impact on dose distribution in intensity-modulated radiation therapy for nasopharyngeal carcinoma. *Oral Oncol* 82:1-7, 2018
5. Moore A: Observer variation in the delineation of organs at risk for head and neck radiation therapy treatment planning: a systematic review protocol. *JBIR Database System Rev Implement Rep* 16:50-56, 2018
6. Nelms BE, Tome WA, Robinson G, et al: Variations in the contouring of organs at risk: test case from a patient with oropharyngeal cancer. *Int J Radiat Oncol Biol Phys* 82:368-78, 2012
7. Daisne JF, Blumhofer A: Atlas-based automatic segmentation of head and neck organs at risk and nodal target volumes: a clinical validation. *Radiat Oncol* 8:154, 2013
8. Yang J, Beadle BM, Garden AS, et al: Auto-segmentation of low-risk clinical target volume for head and neck radiation therapy. *Pract Radiat Oncol* 4:e31-7, 2014
9. Qazi AA, Pekar V, Kim J, et al: Auto-segmentation of normal and target structures in head and neck CT images: a feature-driven model-based approach. *Med Phys* 38:6160-70, 2011

10. Dean JA, Welsh LC, McQuaid D, et al: Assessment of fully-automated atlas-based segmentation of novel oral mucosal surface organ-at-risk. *Radiother Oncol* 119:166-71, 2016
11. Kieselmann JP, Kamerling CP, Burgos N, et al: Geometric and dosimetric evaluations of atlas-based segmentation methods of MR images in the head and neck region. *Phys Med Biol* 63:145007, 2018
12. Lin L, Dou Q, Jin YM, et al: Deep Learning for Automated Contouring of Primary Tumor Volumes by MRI for Nasopharyngeal Carcinoma. *Radiology* 291:677-686, 2019
13. Isambert A, Dhermain F, Bidault F, et al: Evaluation of an atlas-based automatic segmentation software for the delineation of brain organs at risk in a radiation therapy clinical context. *Radiother Oncol* 87:93-9, 2008
14. Hoang Duc AK, Eminowicz G, Mendes R, et al: Validation of clinical acceptability of an atlas-based segmentation algorithm for the delineation of organs at risk in head and neck cancer. *Med Phys* 42:5027-34, 2015
15. Zhu W, Huang Y, Zeng L, et al: AnatomyNet: Deep learning for fast and fully automated whole-volume segmentation of head and neck anatomy. *Med Phys* 46:576-589, 2019
16. Ibragimov B, Xing L: Segmentation of organs-at-risks in head and neck CT images using convolutional neural networks. *Med Phys* 44:547-557, 2017
17. Teguh DN, Levendag PC, Voet PW, 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* 81:950-7, 2011
18. Hague C, Beasley W, Dixon L, et al: Use of a novel atlas for muscles of mastication to reduce inter observer variability in head and neck radiotherapy contouring. *Radiother Oncol* 130:56-61, 2019
19. Weber C, Dommerich S, Pau HW, et al: Limited mouth opening after primary therapy of head and neck cancer. *Oral Maxillofac Surg* 14:169-73, 2010
20. Scott B, Butterworth C, Lowe D, et al: Factors associated with restricted mouth opening and its relationship to health-related quality of life in patients attending a Maxillofacial Oncology clinic. *Oral Oncol* 44:430-8, 2008
21. Louise Kent M, Brennan MT, Noll JL, et al: Radiation-induced trismus in head and neck cancer patients. *Support Care Cancer* 16:305-9, 2008
22. Rao SD, Saleh ZH, Setton J, et al: Dose-volume factors correlating with trismus following chemoradiation for head and neck cancer. *Acta Oncol* 55:99-104, 2016
23. Pauli N, Johnson J, Finizia C, et al: The incidence of trismus and long-term impact on health-related quality of life in patients with head and neck cancer. *Acta Oncol* 52:1137-45, 2013
24. Gebre-Medhin M, Haghanegi M, Robert L, et al: Dose-volume analysis of radiation-induced trismus in head and neck cancer patients. *Acta Oncol* 55:1313-1317, 2016
25. van der Molen L, Heemsbergen WD, de Jong R, et al: Dysphagia and trismus after concomitant chemo-Intensity-Modulated Radiation Therapy (chemo-IMRT) in advanced head and neck cancer;

dose-effect relationships for swallowing and mastication structures. *Radiother Oncol* 106:364-9, 2013

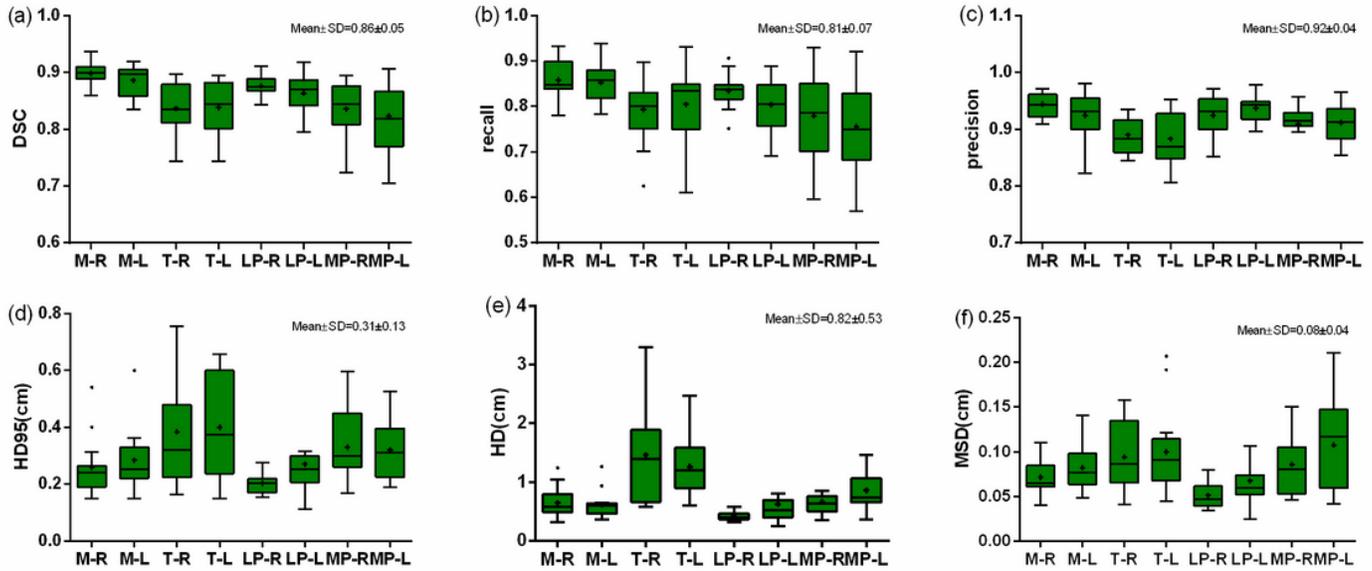
26. Jatin P, Shah PHM: New AJCC/UICC staging system for head and neck, and thyroid cancer. *REV. MED. CLIN. CONDES* - 2018; 29(4) 397-404
27. Çiçek Ö, Abdulkadir A, Lienkamp SS, et al: 3D U-Net: learning dense volumetric segmentation from sparse annotation, *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2016, pp 424-432
28. Yang J, Veeraraghavan H, Armato III SG, et al: Autosegmentation for thoracic radiation treatment planning: A grand challenge at AAPM 2017. *Medical physics* 45:4568-4581, 2018
29. Feng X, Bernard ME, Hunter T, et al: Improving accuracy and robustness of deep convolutional neural network based thoracic OAR segmentation. *Physics in Medicine & Biology*, 2020
30. Feng X, Qing K, Tustison NJ, et al: Deep convolutional neural network for segmentation of thoracic organs-at-risk using cropped 3D images. *Medical physics*, 2019
31. Delpon G, Escande A, Ruef T, et al: Comparison of Automated Atlas-Based Segmentation Software for Postoperative Prostate Cancer Radiotherapy. *Front Oncol* 6:178, 2016
32. Weistrand O, Svensson S: The ANACONDA algorithm for deformable image registration in radiotherapy. *Med Phys* 42:40-53, 2015
33. Lee H, Lee E, Kim N, et al: Clinical Evaluation of Commercial Atlas-Based Auto-Segmentation in the Head and Neck Region. *Front Oncol* 9:239, 2019
34. Kaderka R, Gillespie EF, Mundt RC, et al: Geometric and dosimetric evaluation of atlas based auto-segmentation of cardiac structures in breast cancer patients. *Radiother Oncol* 131:215-220, 2019

## Figures



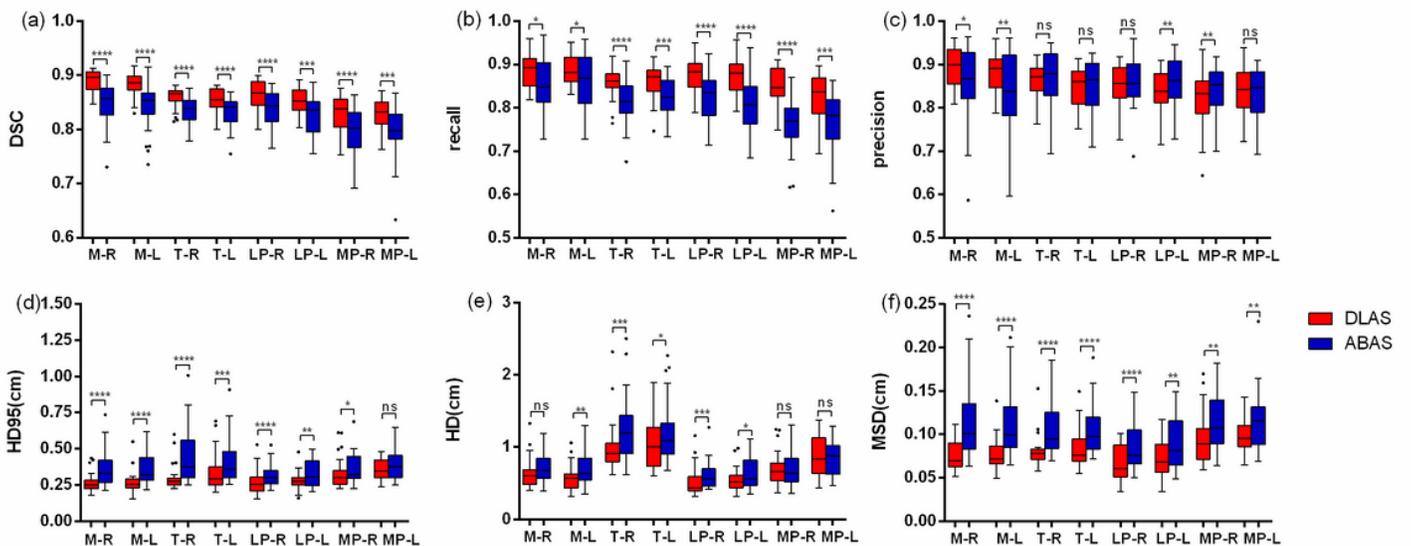
**Figure 1**

Transverse view of different contours for one representative patient. (a) manual contours (green lines, reference standard) vs. DLAS (red lines), (b) manual contours (green lines) vs. ABAS (blue lines).



**Figure 2**

Comparisons of the mean value of six metrics for the three individual clinician manual contours. In each box, the central mark is the median and edges are the 25 and 75th percentiles. and the upper and lower whiskers represents the highest and lowest values. The overall values (mean±SD) for every metrics were presented on the right upper corner for each subfigure. "+" in the box represents the mean values.



**Figure 3**

Comparison DLAS and ABAS performance. The performance was evaluated with (a) DSC, (b) recall, (c) precision, (d) HD95, (e) HD, (f) MSD. In each box, the central mark is the median and edges are the 25 and 75th percentiles. and the upper and lower whiskers represents the highest and lowest values. Two-tail t test was used for analysis. \*P<0.05, \*\* P<0.01, \*\*\* P<0.001, \*\*\*\* P<0.0001, ns, no significance.

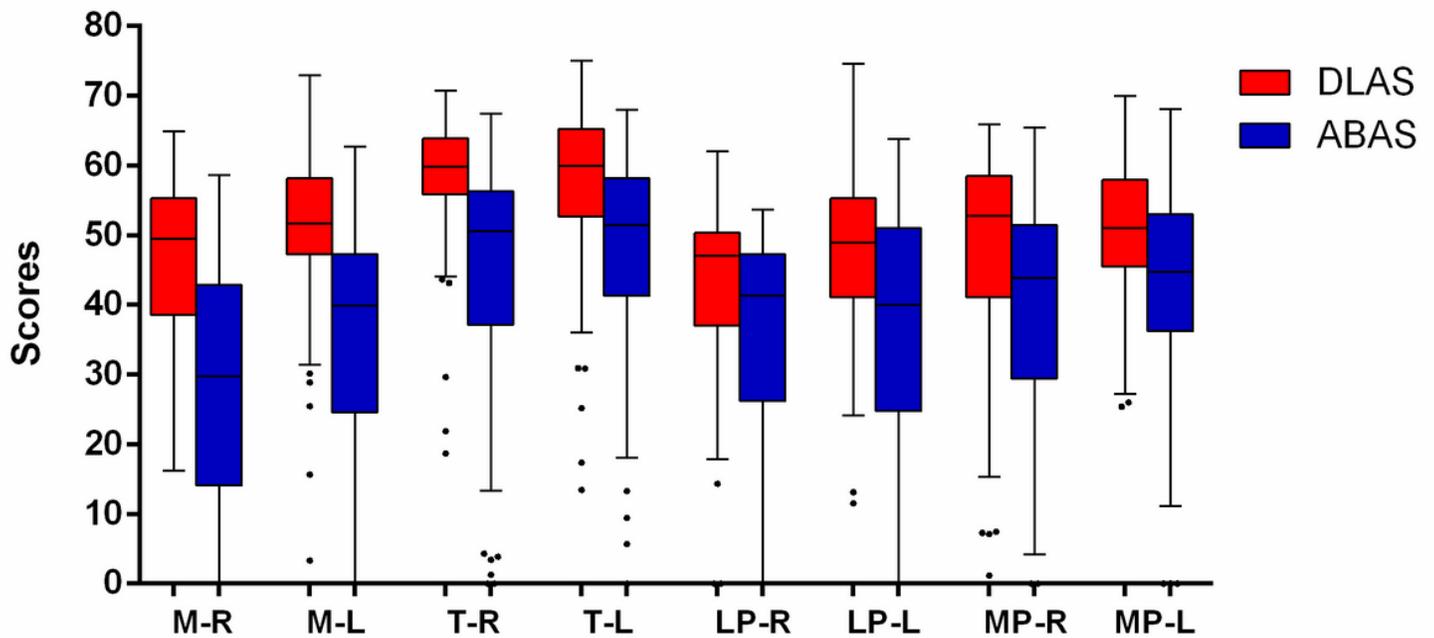


Figure 4

The overall scores achieved by the two methods for all pairs of muscles. In each box, the central mark is the median and edges are the 25 and 75th percentiles. and the upper and lower whiskers represents the highest and lowest values.

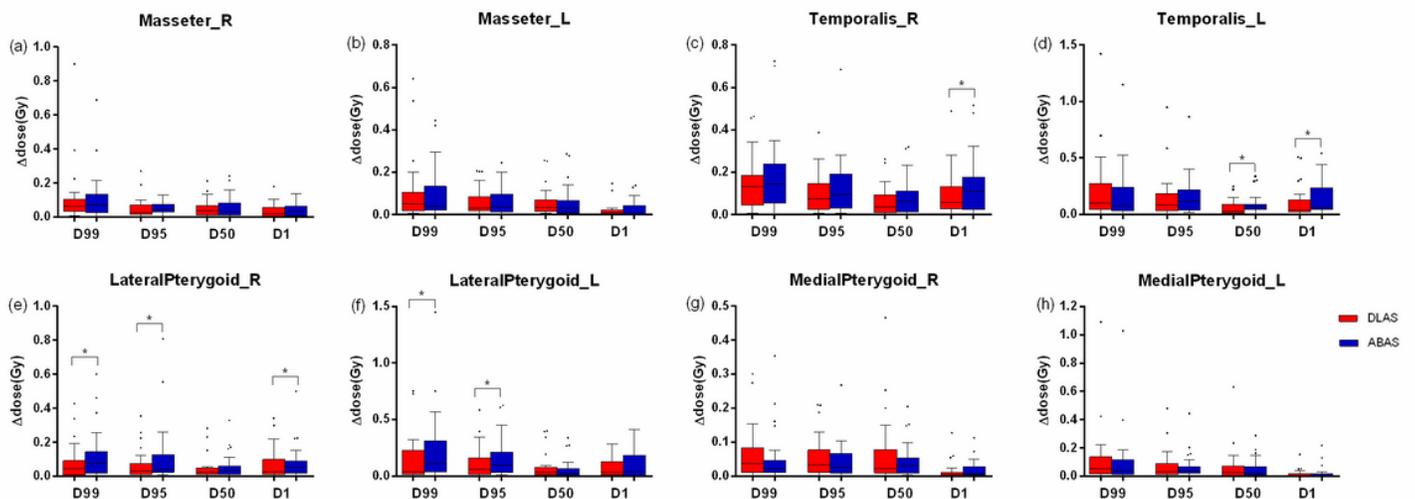


Figure 5

Comparisons of  $\Delta$ dose of DLAS vs ABAS. Two-tail t test was used for analysis. \*P<0.05. In each box, the central mark is the median and edges are the 25 and 75th percentiles. and the upper and lower whiskers represents the highest and lowest values.