# UC Davis
## UC Davis Previously Published Works

**Title**

Evaluating Automatic Segmentation for Swallowing-Related Organs for Head and Neck Cancer

**Permalink**

https://escholarship.org/uc/item/0263s2jq

**Authors**

Li, Yimin
Rao, Shyam
Chen, Wen
et al.

**Publication Date**

2022

**DOI**

10.1177/15330338221105724

Peer reviewed

# Evaluating Automatic Segmentation for Swallowing-Related Organs for Head and Neck Cancer

Yimin Li, MD[1,2], Shyam Rao, MD[2] (iD), Wen Chen, MD[3],
Soheila F. Azghadi, MD[2], Ky Nam Bao Nguyen, MD[2],
Angel Moran, MD[2], Brittni M Usera, MD[2], Brandon A Dyer, MD[4],
Lu Shang, PhD[2], Quan Chen, PhD[5] (iD) and Yi Rong, PhD[2,6] (iD)

## Abstract

**Purpose:** To evaluate the accuracy of deep-learning-based auto-segmentation of the superior constrictor, middle constrictor, inferior constrictor, and larynx in comparison with a traditional multi-atlas-based method. **Methods and Materials:** One hundred and five computed tomography image datasets from 83 head and neck cancer patients were retrospectively collected and the superior constrictor, middle constrictor, inferior constrictor, and larynx were analyzed for deep-learning versus multi-atlas-based segmentation. Eighty-three computed tomography images (40 diagnostic computed tomography and 43 planning computed tomography) were used for training the convolutional neural network, and for atlas-based model training. The remaining 22 computed tomography datasets were used for validation of the atlas-based auto-segmentation versus deep-learning-based auto-segmentation contours, both of which were compared with the corresponding manual contours. Quantitative measures included Dice similarity coefficient, recall, precision, Hausdorff distance, 95th percentile of Hausdorff distance, and mean surface distance. Dosimetric differences between the auto-generated contours and manual contours were evaluated. Subjective evaluation was obtained from 3 clinical observers to blindly score the autosegmented structures based on the percentage of slices that require manual modification. **Results:** The deep-learning-based auto-segmentation versus atlas-based auto-segmentation results were compared for the superior constrictor, middle constrictor, inferior constrictor, and larynx. The mean Dice similarity coefficient values for the 4 structures were 0.67, 0.60, 0.65, and 0.84 for deep-learning-based auto-segmentation, whereas atlas-based auto-segmentation has Dice similarity coefficient results at 0.45, 0.36, 0.50, and 0.70, respectively. The mean 95th percentile of Hausdorff distance (cm) for the 4 structures were 0.41, 0.57, 0.59, and 0.54 for deep-learning-based auto-segmentation, but 0.78, 0.95, 0.96, and 1.23 for atlas-based auto-segmentation results, respectively. Similar mean dose differences were obtained from the 2 sets of autosegmented contours compared to manual contours. The dose–volume discrepancies and the average modification rates were higher with the atlas-based auto-segmentation contours. **Conclusion:** Swallowing-related structures are more accurately generated with DL-based versus atlas-based segmentation when compared with manual contours.

[1] Department of Radiation Oncology, Xiamen Radiotherapy Quality Control Center, Xiamen Cancer Center, Xiamen Key Laboratory of Radiation Oncology, The First Affiliated Hospital of Xiamen University, The Third Clinical Medical College, Fujian Medical University, Xiamen, Fujian, China
[2] Department of Radiation Oncology, University of California Davis Medical Center, Sacramento, CA, USA
[3] Department of Radiation Oncology, Xiangya Hospital, Central South University, Changsha, China
[4] Department of Radiation Oncology, Legacy Health, Portland, OR, USA
[5] Department of Radiation Oncology, City of Hope comprehensive Cancer Center, Duarte, CA, USA
[6] Department of Radiation Oncology, Mayo Clinic Arizona, Phoenix, AZ, USA

**Corresponding Authors:**
Yi Rong, PhD, Department of Radiation Oncology, Mayo Clinic Arizona, 5777 E. Mayo Blvd, Phoenix, AZ 85054, USA.
Email: rong.yi@mayo.edu

Quan Chen, PhD, Department of Radiation Oncology, City of Hope comprehensive cancer center, Duarte, CA 91010..
Email: quanchen@gmail.com

## Introduction

Incorporation of multimodal imaging and intensity-modulated radiotherapy improved local tumor control for head and neck cancer (HNC) patients and decreased toxicity to adjacent organs at risk (OAR).[1] Because of improved tumor control and patient outcome, shifting focus to radiation-induced, swallow-related toxicities have been a major concern for HNC quality of life management. In order to determine the radiation dose–volume effects,[2,3] delineation of the normal organs is a prerequisite to study the radiation dose effect of different pharyngeal muscles on dysphagia.[4] Traditionally, the pharyngeal muscles are contoured manually but can be further segmented into upper, middle, and lower pharyngeal muscles. However, manual segmentation of these muscles is time-consuming[5] and is associated with interobserver variability.[2,6] Furthermore, due to the anatomic complexity, accurate segmentation of these structures requires extensive HN expertise.

The HN region is considered to be the most challenging anatomic site for radiation therapy due to the large number, small size, and geometric complexity of organs to be contoured. Automatic segmentation has been used to facilitate manual segmentation with reduced interobserver variations.[7–9] Atlas-based auto-segmentation (ABAS) has been tested for various cancer sites,[10] and refinement for clinical application continues.[7,11–14] Separately, deep-learning-based models trained with convolutional neural networks (CNNs) have demonstrated great accuracy in delineating various organs.[15–19] DL-based CNNs have shown promising accuracy in recent years[16,20,21]; however, research specifically using CNNs to outline the subdivisions of the pharyngeal contractors and larynx for swallowing-related functions has not been thoroughly investigated. In this study, we aim to evaluate the accuracy of deep-learning-based auto-segmentation (DLAS) of the swallowing-related OARs compared with ABAS and the gold-standard manual segmentation. The performance of these 2 methods was comprehensively evaluated using quantitative geometric and dosimetric measures, along with subjective scores.

## Methods and Materials

### Dataset

The study has received approval from Institutional Review Board (IRB #1313551-1). A total of 105 computed tomography (CT) image sets in 83 HN cancer patients (including diagnosis and simulation CTs) were retrospectively identified from the database from November 2014 and April 2017. Four swallowing OARs were manually segmented by an HNC expert radiation oncologist (YL), following the published consensus guideline.[22] All manual contours were then reviewed and modified if needed by another radiation oncologist specialized in HNC (SR). The 4 OARs were named CONSSUP, CONSMID, CONSINF, and LARYNX, corresponding to the superior constrictor muscle, mid constrictor muscle, inferior constrictor muscle, and larynx, respectively. Among the 105 data sets of CT and manual contours, 50 were used for training and 33 for validation. After model training and validation, the remaining 22 were used for testing and performance benchmarking. Patient characteristics with their corresponding treatment plans are listed in Supplemental Table S1. Exclusion criteria include patients with prior radiation therapy or surgery in the head and neck region. No model retuning or retesting was performed. The majority of HN radiotherapy plans had a prescription up to 70 Gy to the primary target in 30 to 35 fractions.

### Auto-Segmentation Model Creation

A commercial software (INTContour, Carina Medical LLC) was used to create deep-learning models for swallowing OARs and has been previously described.[23,24] The software employs 3D U-Net structure[25] for automatic organ segmentation. Briefly, CT datasets were resampled to consistent spatial resolution and matrix size. Two separate 3D U-Nets were trained, one with dilated convolutions and one without. The outputs from both networks were averaged. Augmentations such as random translation, rotation, scaling, and left-right flipping were used for both training and model deployment. The loss function used includes the weighted cross entropy and soft Dice loss. The model training was packaged as an "incremental learning" feature in the software which allows the user to train new or update existing organ segmentation models. The hyperparameter tuning and network training schemes were automated, and no human intervention was performed.

The ABAS model (MIM Version 6.9.6, MIM Software Inc) was built using a representative set of 25 CT datasets taken from the training set. The cases selected for training the atlas model were carefully evaluated to represent a wide spectrum of clinical scenarios, including patients with dental implants and

different head tilt angles. The same 22 datasets that were used for testing the DLAS model were used for the ABAS model testing and evaluation.

## Quantitative Evaluation

*Geometric comparisons.* The performance of both DLAS and ABAS was evaluated by comparing differences between the automatically generated and manual contours. Dice similarity coefficient (DSC), precision, recall, Hausdorff distance (HD), 95th percentile of Hausdorff distance (HD95), measure surface distance (MSD), and mean dose were used as the evaluation metrics for all 22 patients in the testing cohort. The calculation formula of each parameter is as follows:

1. The DSC,[26] precision, and recall are parameters to measure the degree of overlap between 2 volumes (Vx and Vy). The calculation formula is as follows:

$$DSC = \frac{2|Vx \cap Vy|}{|Vx| + |Vy|}$$

$$Precision = \frac{|Vx \cap Vy|}{|Vy|}$$

$$Recall = \frac{|Vx \cap Vy|}{|Vx|}$$

For DSC, Precision, and Recall are ratio or percentage, the range of the above 3 metrics are [0, 1], with 1 being the best value, and 0 being the worst. The HD determines the maximum distance from one point of a contour to the closest pair-wise point of another contour. The HD95 is the 95th percentile of the distance of the corresponding point between the 2 structures.[27] The specific points on contour X and contour Y are described as "x" and "y." HD and HD95 define as:

$$d_{HD} = \max\left(\min_{x \in X} d(x), \min_{y \in Y} d(y)\right)$$

$$\vec{d}_{HD95}(X, Y) = k_{95}\left(\min_{y \in Y} d(x, y)\right), d_{HD95}(X, Y)$$

$$= \frac{\vec{d}_{HD95}(X, Y) + \vec{d}_{HD95}(Y, X)}{2}$$

The directed average surface distance is the mean distance from the point in profile X to the closest point in profile Y, which describes as:

$$\vec{d}_{avg}(X, Y) = \frac{1}{|x|} \sum_{x \in X} \min_{y \in Y} d(x, y)$$

The MSD is the average of the pairwise surface distances and is calculated as follows:

$$d_{avg}(X, Y) = \frac{\vec{d}_{avg}(X, Y) + \vec{d}_{avg}(Y, X)}{2}$$

The unit is cm in this study and 0 represents the most ideal value.

*Dosimetric comparisons.* Dose statistics were computed for DLAS and ABAS contours and manual contours (reference) using the dose distribution of the original, delivered clinical treatment plans. Dosimetric metrics of manual contours and autosegmented contours for 4 studied OARs were assessed.

## Qualitative Evaluation

DLAS and ABAS generated contours were randomly assigned to 3 radiation oncologists for independent evaluation of the 4 swallowing structures following expert consensus guidelines.[28] The number of slices that needs manual modification in each of the structure was recorded. The average modification rate was calculated for each structure from all 3 evaluators.

## Statistical Analysis

SPSS (SPSS Inc, version 22) and GraphPad Prism version 6 (Graph pad software) were used to analyze the data. We used paired *t* test to calculate the difference of DSC, recall, precision, HD, HD95%, MSD, and mean absolute dose between DLAS and ABAS. A *P* value <.05 was regarded as statistically significant.

# Results

## Quantitative Evaluation

For all 4 studied structures, an improvement in a quantitative evaluation of overlapping areas is observed in DLAS contours, as compared to ABAS (Figure 1). Similarly, an improvement in distance measures for all 4 studied structures was observed in DLAS contours, compared to ABAS (Figure 2). Most of the distance related parameters indicated that DLAS is greater than ABAS, and the comparison shows statistical significance (*P* < .05) (Figure 2).

The greatest improvement was observed in CONSMID, details shown in Table 1. The DSC increased from 0.36 ± 0.18 (ABAS) to 0.60 ± 0.19 (DLAS) and the HD95 decreased from 0.95 cm ± 0.40 cm (ABAS) to 0.57 cm ± 0.57 cm (DLAS). The DSC of LARYNX increased on average from 0.70 ± 0.13 (ABAS) to 0.84 ± 0.05 (DLAS) and the HD95 decreased on average from 1.90 cm ± 0.75 cm (ABAS) to 1.00 cm ± 0.33 cm (DLAS). When comparing the CONSMID and LARYNX, the improvement of DSC and HD95 in CONSSUP and CONSINF was relatively smaller.

The mean absolute dose obtained from the manual and auto-contours is shown in Figure 3. There is no statistical significance
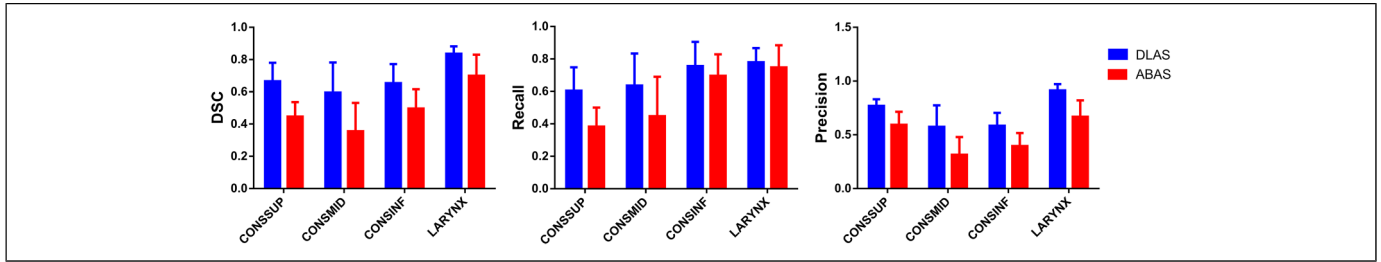
**Figure 1.** Average and 95% confidence interval of DSC, Recall, Precision of DLAS (blue) and ABAS (red). All others were significant with a *P* value <.001 of the paired test, except for the Recall difference of CONSINF (*P* = .038) and LARYX (*P* = .301). Abbreviations: CONSSUP, Superior pharynx constrictor muscle; CONSMID, Middle pharynx constrictor muscle; CONSINF, Inferior pharynx constrictor muscle; LARYNX, larynx; ABAS, atlas-based auto-segmentation; DLAS, deep-learning-based auto-segmentation; DSC, Dice similarity coefficient.
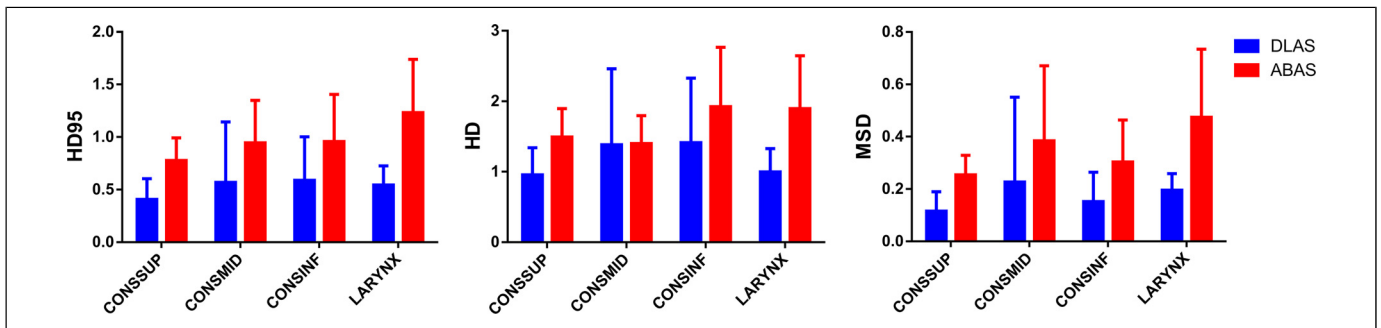


**Figure 2.** Average and 95% confidence interval of HD 95, HD and MSD of DLAS (blue) and ABAS (red). All others were significant with a *P* value <.001 of the paired test, except for the HD of CONSMID (*P* = .945), CONSINF (*P* = .011), and MSD of the CONSMID (*P* = .067). For OAR abbreviations refer to Figure 1. ABAS, atlas-based auto-segmentation; DLAS, deep-learning-based auto-segmentation; HD, Hausdorff distance; MSD, mean surface distance.

**Table 1.** Statistical Data Corresponding to Individual ROIs for DLAS and ABAS.

|                       | DSC             | Recall          | Precision       | HD95            | HD              | MSD             |
|-----------------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| Superior constrictor  |                 |                 |                 |                 |                 |                 |
| DLAS                  | $0.67 \pm 0.11$ | $0.60 \pm 0.14$ | $0.77 \pm 0.06$ | $0.41 \pm 0.20$ | $0.96 \pm 0.38$ | $0.12 \pm 0.07$ |
| ABAS                  | $0.45 \pm 0.09$ | $0.38 \pm 0.12$ | $0.59 \pm 0.12$ | $0.78 \pm 0.21$ | $1.49 \pm 0.40$ | $0.25 \pm 0.07$ |
| Middle constrictor    |                 |                 |                 |                 |                 |                 |
| DLAS                  | $0.60 \pm 0.19$ | $0.64 \pm 0.20$ | $0.57 \pm 0.20$ | $0.57 \pm 0.57$ | $1.38 \pm 1.08$ | $0.23 \pm 0.32$ |
| ABAS                  | $0.36 \pm 0.18$ | $0.45 \pm 0.24$ | $0.31 \pm 0.16$ | $0.95 \pm 0.40$ | $1.40 \pm 0.40$ | $0.38 \pm 0.29$ |
| Inferior constrictor  |                 |                 |                 |                 |                 |                 |
| DLAS                  | $0.65 \pm 0.12$ | $0.76 \pm 0.15$ | $0.58 \pm 0.12$ | $0.59 \pm 0.41$ | $1.41 \pm 0.91$ | $0.15 \pm 0.11$ |
| ABAS                  | $0.50 \pm 0.12$ | $0.70 \pm 0.13$ | $0.40 \pm 0.12$ | $0.96 \pm 0.45$ | $1.92 \pm 0.84$ | $0.30 \pm 0.16$ |
| Larynx                |                 |                 |                 |                 |                 |                 |
| DLAS                  | $0.84 \pm 0.05$ | $0.78 \pm 0.09$ | $0.91 \pm 0.06$ | $0.54 \pm 0.18$ | $1.00 \pm 0.33$ | $0.20 \pm 0.06$ |
| ABAS                  | $0.70 \pm 0.13$ | $0.75 \pm 0.14$ | $0.67 \pm 0.15$ | $1.23 \pm 0.51$ | $1.90 \pm 0.75$ | $0.47 \pm 0.26$ |

Abbreviations: ABAS, atlas-based auto-segmentation; DBAS, deep-learning-based auto-segmentation; DSC, Dice similarity coefficient; HD, Hausdorff distance; HD95, 95th percentile of Hausdorff distance; MSD, mean surface distance.

for all studied structures (*P* > .05). The ratios of the OAR volume obtained by DLAS and ABAS, to the reference volume obtained from those manual contours, are $V_{DLAS}$/$V_{Manual}$, and $V_{ABAS}$/$V_{Manual}$, respectively. The mean values of $V_{DLAS}$/$V_{Manual}$ and $V_{ABAS}$/$V_{Manual}$ for CONSSUP, CONSMID, CONSINF, and LARYNX are $0.79 \pm 0.20$ VS $0.70 \pm 0.34$, $1.19 \pm 0.30$ VS $1.52 \pm 0.69$, $1.32 \pm 0.25$ VS $1.88 \pm 0.49$, $0.86 \pm 0.14$ VS $1.15 \pm 0.22$. There is statistical significance in differences for all studied structures, except for CONSSUP (*P* > .05).

The time for building an ABAS model library took about 2 h, and the model training time for DLAS took about 6 h.
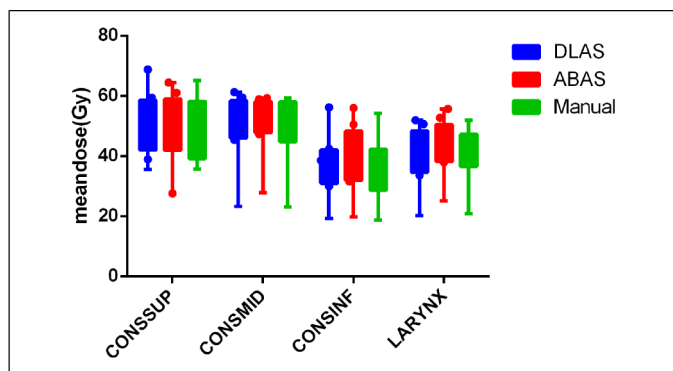
**Figure 3.** Mean doses of DLAS (blue), ABAS (red), and manual (green) contours. For OAR abbreviations refer to Figure 1. ABAS, atlas-based auto-segmentation; DLAS, deep-learning-based auto-segmentation.
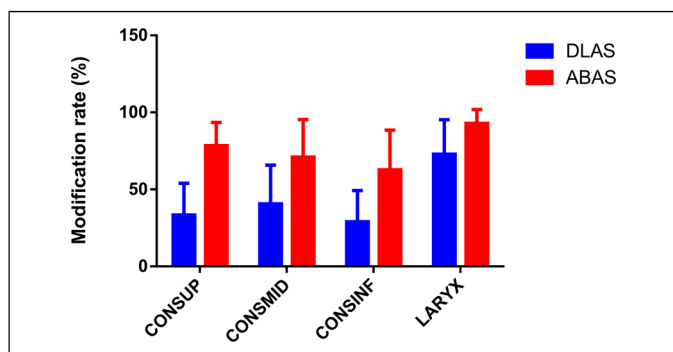


**Figure 4.** Modification rate of the 4 OARs. For OAR abbreviations refer to Figure 1.

The average time for each patient to complete an ABAS and DLAS cycle is about 10 and 2 min, respectively.

## Qualitative Evaluation

Overall, the average modification rate of ABAS was higher than that of DLAS for all studied structures, as shown in Figure 4. The modification rate of CONSSUP, CONSMID, and CONSINF in DLAS was lower than 50%, while those of ABAS exceeded 60%. CONSSUP had the largest improvement, from $0.78 \pm 0.15$ (ABAS) to $0.33 \pm 0.21$ (DLAS), while Larynx had the smallest, from $0.93 \pm 0.09$ (ABAS) to $0.73 \pm 0.22$ (DLAS). . The independent sample $t$ test indicated that difference was statistically significant for all 4 structures ($P < .05$).

DLAS performed better than ABAS when segmenting all 4 swallowing structures. The DSC and modification rate of CONSSUP are significantly better for DLAS than that of ABAS, and the performance of DLAS is significantly better than ABAS in judging the posterior and lateral border (Figure 5A). For CONSMID and CONSINF, DLAS also showed higher accuracy in determining the contour borders in 3-dimensions compared to ABAS. In the case with the lowest DSC (0.272), it was found that the

autogenerated contour is too forward compared to the manual one for CONSINF (Figure 5B). However, this particular patient had a tracheostomy and a cannula, which may have caused inaccuracy of automatic segmentation due to significant anatomy change.

LARYNX has the best DSC and HD95 performance compared to the pharyngeal constrictor muscles. Modification rate exceeded 0.5 in both DLAS ($0.73 \pm 0.22$) and ABAS ($0.93 \pm 0.09$) for LARYNX. Figure 6 shows reasonable overlap between DLAS and manual contours, with inferior overlap between ABAS and manual. Figure 7 shows a representative case of Larynx performance that the DLAS failed to maintain the triangle shape for larynx.

## Discussion

To our knowledge, this is the first data to evaluate the feasibility of automatic segmentation of swallowing function structure based on deep-learning CNN. In our study, DLAS and ABAS were comprehensively evaluated using quantitative methods of geometric overlap, distance measures, absolute dose difference, and qualitative modification rates. Compared with ABAS, the results indicate that DLAS creates more accurate, consistent, reproducible pharyngeal contraction muscles, and larynx contours with less manual correction. Although it takes more training time for DLAS than building an ABAS model, it takes less time for each patient to complete a DLAS cycle, which will greatly improve our clinical efficiency in eliminate manual contouring time.

Our results show similar DSC values to those 22 normal organs determined by van Dijk et al[18] using a training set of more than 500 HNC patients. In this study, the pharyngeal contraction muscle was automatically segmented and evaluated as a whole structure. We expanded upon van Dijk's paper by segmenting the pharyngeal contraction muscle into 3 parts: superior, middle, and inferior pharyngeal contraction muscle. In addition, we contour the larynx, which was analogous to "glottis area" in van Dijk's study.[18] The subdivision of the pharyngeal constrictor is meaningful in the evaluation of side effects of radiotherapy for dysphagia. Levendag's research has shown that the radiation dose of different parts of the pharyngeal constrictor causes different quality of life scores for patients with swallowing dysfunction.[4] In fact, the guidelines for delineation of organs related to swallowing in radiotherapy published by Miranda[22] in 2011 have already suggested a detailed delineation of the pharyngeal constrictor muscle. It is challenging to manually contour these substructures. Our aim was to evaluate the feasibility and accuracy of deep learning to automatically delineate the subdivided pharyngeal contraction muscle and larynx which may be beneficial for more elaborate research on dysphagia in the future. The difference between organ dose of the automatically segmented and the manual contour was analyzed, which serves as a prelude for future correlation analysis between radiation dose of swallowing related organs and dysphasia after radiotherapy.

Examples of comparing DLAS and ABAS are shown in Figure 5 with best and worst case scenarios. When large anatomy deviation exists (probably due to surgery), DLAS may
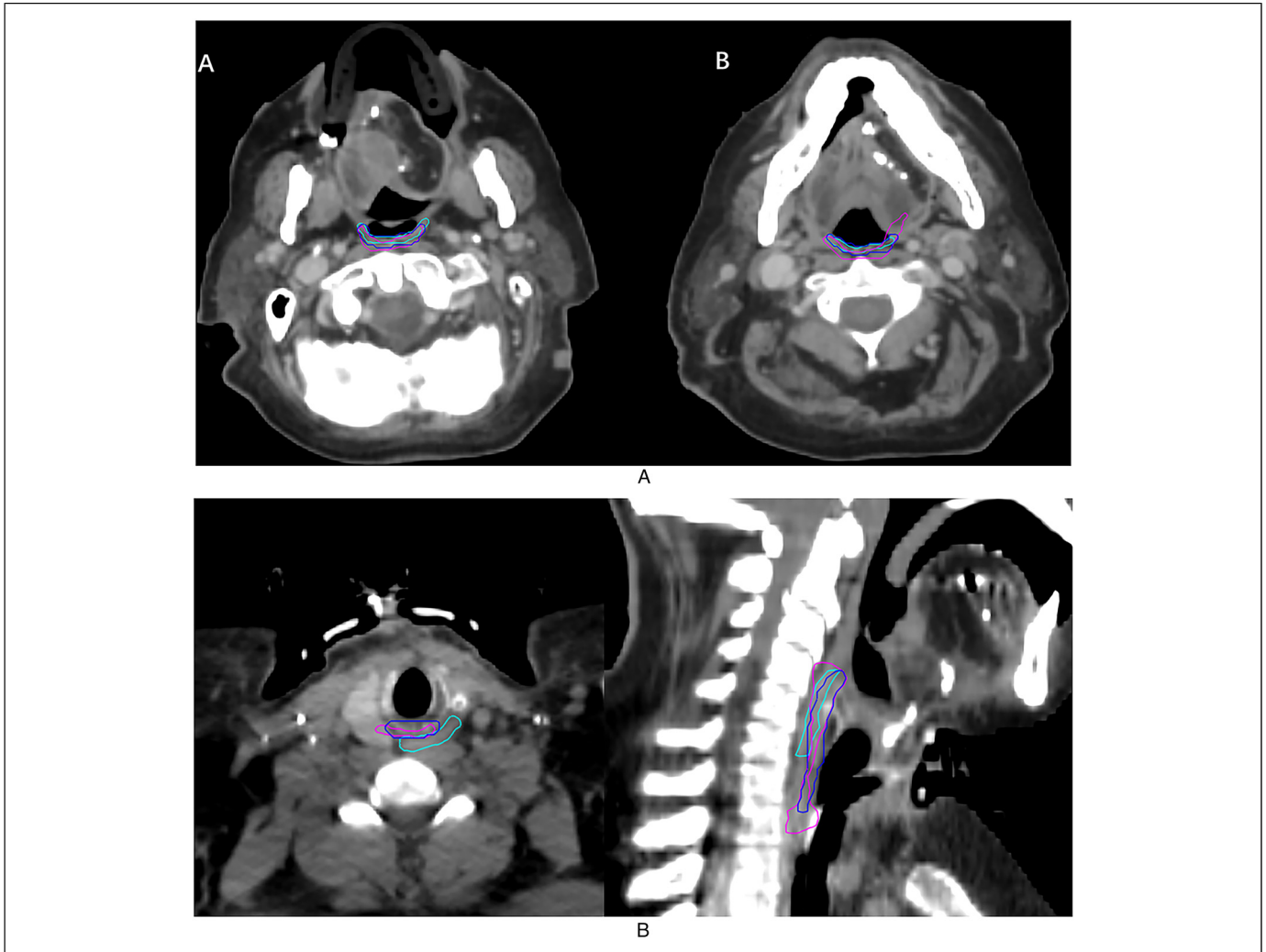
**Figure 5.** (A) Panel A shows the back border are more accurate in DLAS, and Panel B shows the posterior and lateral border are more accurate in DLAS. (B) A worst scenario case with a tracheotomy and a cannula, which may cause inaccuracy due to large anatomy change. Panel A shows the axial slice, and Panel B shows the sagittal slice. Cyan, magenta, and blue lines represent manual, ABAS, and DLAS, respectively. ABAS, atlas-based auto-segmentation; DLAS, deep-learning-based auto-segmentation.



**Figure 6.** A representative case with good overlap between DLAS and Manual, while ABAS is relatively poor, in axial, sagittal, and coronal views. Cyan, magenta, and blue lines represent manual outline, ABAS, and DLAS in turn. ABAS, atlas-based auto-segmentation; DBAS, deep-learning-based auto-segmentation.
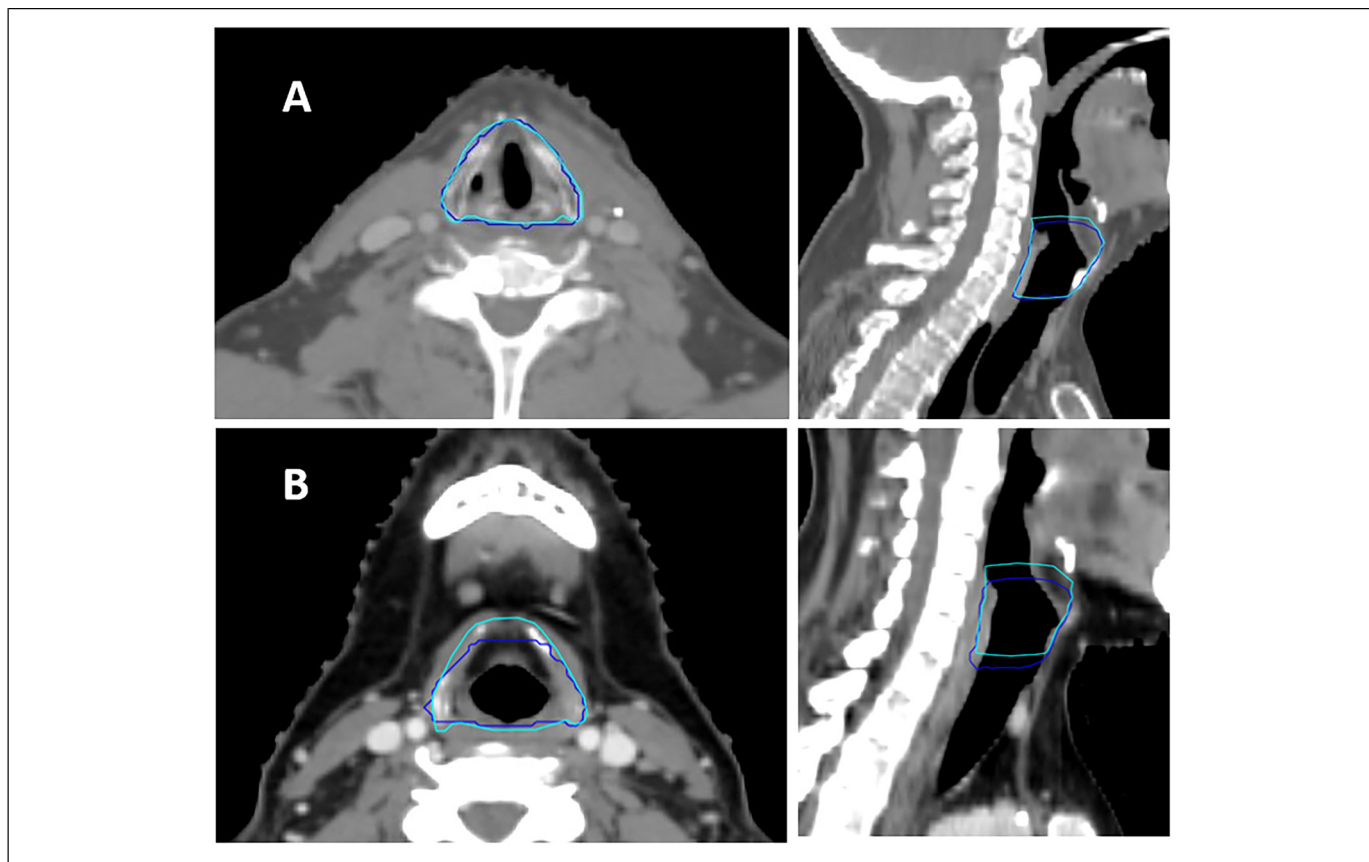
**Figure 7.** Panel A shows DLAS and manual overlap very well. Panel B shows the discrepancy between DLAS and manual locates at the cranial side and caudal side mostly. Cyan, magenta, and blue lines represent manual outline, ABAS, and DLAS in turn. ABAS, atlas-based auto-segmentation; DBAS, deep-learning-based auto-segmentation.

have higher inaccuracy when the special scenarios are not included in the training model (Figure 5B). This phenomenon is somewhat similar to Tong's[29] research, which indicated that the shape of the structure can be greatly impacted on the results of automatic segmentation for surgical patients. Although the modification rate of LARYNX is high, the results are still meaningful, as the Larynx actually has high Dice score, and similar HD95, HD, MSD as other structures. The discrepancy between DLAS and manual locates at the cranial side and caudal side mostly. It can be found clearly in sagittal plane (shown in Figure 6). In addition, due to the cartilage boundary of the larynx is relatively clear, the difference looks more eye-catching, which may also be the reason for the high proportion of unacceptable from the evaluators. The deep-learning method showed encouraging results for autosegmenting larynx, as shown in Figure 6; however for a few cases, as shown in Figure 7, the contours from the deep-learning model failed to maintain the triangle shape for the larynx. This is likely due to the variation in the intensities surrounding the bony structures and the model failed to utilize the high-level shape information to make the correct segmentation. Previous study also has shown that Intra-imaging uncertainty for segmented geometry and position. Larynx is highly variable during deglutition and could skew training during imaging the larynx and there is around an 8 mm total laryngeal setup error.[30]

Although we hypothesized that the model has learned both high-level and low-level information for segmentation, it may not yield the right balance for every case. A straightforward and effective solution for improvement is to use a larger training set. Our future study will explore the probabilistic output of the model to use additional post-processing steps to explicitly enforce shape constraints. In addition, researchers have proposed the concept of swallowing functional units[31,32] in recent years, which brought a more in-depth study on the anatomy and physiology of swallowing function compare to the classic constrictor muscles and larynx. This may provide another application in future research for the auto-segmentation concerning swallowing organs.

## Conclusion

The application of DLAS in subdivided swallowing-related organs (superior, middle, and inferior pharyngeal contractions and larynx) is feasible. In comparison with an automatic drawing based on Atlas, DLAS is significantly better than ABAS in the aspect of geometric overlap and distance parameters as well as the subjective evaluation and will consume less time. In the near future, DLAS may replace ABAS as a routine in the automatic segmentation of swallowing-related organs.

## ORCID iDs

Shyam Rao      https://orcid.org/0000-0002-8813-5879
Quan Chen      https://orcid.org/0000-0001-5570-2462
Yi Rong      https://orcid.org/0000-0002-2620-1893

## Supplemental Material

Supplemental material for this article is available online.

## References

1. Gregoire V, Langendijk JA, Nuyts S. Advances in radiotherapy for head and neck cancer. *J Clin Oncol*. 2015;33(29):3277-3284.
2. Geets X, Daisne JF, Arcangeli S, et al. Inter-observer variability in the delineation of pharyngo-laryngeal tumor, parotid glands and cervical spinal cord: comparison between CT-scan and MRI. *Radiother Oncol*. 2005;77(1):25-31.
3. Eisbruch A, Schwartz M, Rasch C, et al. Dysphagia and aspiration after chemoradiotherapy for head-and-neck cancer: which anatomic structures are affected and can they be spared by IMRT? *Int J Radiat Oncol Biol Phys*. 2004;60(5):1425-1439.
4. Levendag PC, Teguh DN, Voet P, et al. Dysphagia disorders in patients with cancer of the oropharynx are significantly affected by the radiation therapy dose to the superior and middle constrictor muscle: a dose-effect relationship. *Radiother Oncol*. 2007;85(1):64-73.
5. Vorwerk H, Zink K, Schiller R, et al. Protection of quality and innovation in radiation oncology: the prospective multicenter trial the German society of radiation oncology (DEGRO-QUIRO study). Evaluation of time, attendance of medical staff, and resources during radiotherapy with IMRT. *Strahlenther Onkol*. 2014;190(5):433-443.
6. Brouwer CL, Steenbakkers RJ, van den Heuvel E, et al. 3D variation in delineation of head and neck organs at risk. *Radiat Oncol*. 2012;7:32.
7. Sharp G, Fritscher KD, Pekar V, et al. Vision 20/20: perspectives on automated image segmentation for radiotherapy. *Med Phys*. 2014;41(5):050902.
8. Lim JY, Leech M. Use of auto-segmentation in the delineation of target volumes and organs at risk in head and neck. *Acta Oncol*. 2016;55(7):799-806.
9. 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.
10. Han X, Hoogeman MS, Levendag PC, et al. Atlas-based auto-segmentation of head and neck CT images. *Med Image Comput Comput Assist Interv*. 2008;11(Pt 2):434-441.
11. 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*. 2011;81(4):950-957.
12. Van de Velde J, Wouters J, Vercauteren T, et al. Optimal number of atlases and label fusion for automatic multi-atlas-based brachial plexus contouring in radiotherapy treatment planning. *Radiat Oncol*. 2016;11:1.
13. Yeo UJ, Supple JR, Taylor ML, et al. Performance of 12 DIR algorithms in low-contrast regions for mass and density conserving deformation. *Med Phys*. 2013;40(10):101701.
14. Zhong H, Kim J, Chetty IJ. Analysis of deformable image registration accuracy using computational modeling. *Med Phys*. 2010;37(3):970-979.
15. Meyer P, Noblet V, Mazzara C, Lallement A. Survey on deep learning for radiotherapy. *Comput Biol Med*. 2018(1);98:126-146.
16. Ibragimov B, Xing L. Segmentation of organs-at-risks in head and neck CT images using convolutional neural networks. *Med Phys*. 2017;44(2):547-557.
17. Liang S, Tang F, Huang X, et al. Deep-learning-based detection and segmentation of organs at risk in nasopharyngeal carcinoma computed tomographic images for radiotherapy planning. *Eur Radiol*. 2019;29(4):1961-1967.
18. van Dijk LV, Van den Bosch L, Aljabar P, et al. Improving automatic delineation for head and neck organs at risk by deep learning contouring. *Radiother Oncol*. 2020;142:115–123.
19. Chen W, Li Y, Dyer BA, et al. Deep learning vs. Atlas-based models for fast auto-segmentation of the masticatory muscles on head and neck CT images. *Radiat Oncol*. 2020;15(1):176.
20. Vrtovec T, Mocnik D, Strojan P, et al. Auto-segmentation of organs at risk for head and neck radiotherapy planning: from atlas-based to deep learning methods. *Med Phys*. 2020;47(9):e929-e950.
21. 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*. 2019;46(2):576-589.
22. Christianen ME, Langendijk JA, Westerlaan HE, et al. Delineation of organs at risk involved in swallowing for radiotherapy treatment planning. *Radiother Oncol*. 2011;101(3):394-402.
23. Feng X, Qing K, Tustison NJ, et al. Deep convolutional neural network for segmentation of thoracic organs-at-risk using cropped 3D images. *Med Phys*. 2019;46(5):2169-2180.

24. Feng X, Bernard ME, Hunter T, Chen Q. Improving accuracy and robustness of deep convolutional neural network based thoracic OAR segmentation. *Phys Med Biol*. 2020;65(7):07NT01.

25. 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; 424-432.

26. D LR. Measures of the amount of ecologic association between species. *Ecology*. 1945;26:297-302.

27. Heimann T, van Ginneken B, Styner MA, et al. Comparison and evaluation of methods for liver segmentation from CT datasets. *IEEE Trans Med Imaging*. 2009;28(8):1251-1265.

28. Brouwer CL, Steenbakkers RJ, Bourhis J, et al. CT-based delineation of organs at risk in the head and neck region: DAHANCA, EORTC, GORTEC, HKNPCSG, NCIC CTG, NCRI, NRG oncology and TROG consensus guidelines. *Radiother Oncol*. 2015;117(1):83-90.

29. Tong N, Gou S, Yang S, et al. Fully automatic multi-organ segmentation for head and neck cancer radiotherapy using shape representation model constrained fully convolutional neural networks. *Med Phys*. 2018;45(10):4558-4567.

30. Baron CA, Awan MJ, Mohamed AS, et al. Estimation of daily interfractional larynx residual setup error after isocentric alignment for head and neck radiotherapy: quality assurance implications for target volume and organs-at-risk margination using daily CT on- rails imaging. *J Appl Clin Med Phys*. 2014;16(1):5108.

31. Gawryszuk A, Bijl HP, Holwerda M, et al. Functional swallowing units (FSUs) as organs-at-risk for radiotherapy. PART 1: physiology and anatomy. *Radiother Oncol*. 2019;130:62-67.

32. Gawryszuk A, Bijl HP, Holwerda M, et al. Functional swallowing units (FSUs) as organs-at-risk for radiotherapy. PART 2: advanced delineation guidelines for FSUs. *Radiother Oncol*. 2019;130:68-74.