Deep learning-based auto-segmentation of swallowing and chewing structures in CT

Aditi Iyer, Maria Thor, Rabia Haq, Joseph O. Deasy, Aditya P. Apte*

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

Abstract:

Delineating swallowing and chewing structures in Head and Neck (H&N) CT scans is necessary during radiotherapy treatment (RT) planning to limit the incidence of speech dysfunction, dysphagia, and trismus. Automating this process is desirable so as to reduce the amount of manual labor required and to generate reproducible contours. However, this is a challenging problem due to the low soft tissue contrast and morphological complexity of the structures involved.

In this work, we developed deep learning-based models using 194 H&N CT scans from our institution to automate the delineation of the masseters (left and right), medial pterygoids (left and right), larynx, and constrictor muscles using the DeepLabV3+ architecture. An ensemble of three models was developed for each group of structures using axial, coronal, and sagittal images to provide useful contextual information from the three different views. Each of these models was trained in 2.5D by populating the channels with three consecutive slices. The resulting probability maps were averaged, and voxels were assigned labels corresponding to the class with the highest combined probability. This method offers the advantage of providing increased redundancy in case of unintended failures of the individual models.

A hold-out dataset of 24 H&N CT scans was used to evaluate the resulting models against expert contours in terms of Dice Similarity Coefficient (DSC) and 95th percentile Hausdorff distance. Additionally, clinically relevant dose-volume histogram-based metrics were extracted from manual and auto-generated contours and compared to assess clinical utility. The models yielded DSC exceeding or comparable to the performance of previously reported auto-segmentation methods. Differences in DVH metrics compared to ground-truth were not found to be statistically significant for all structures except the constrictor muscles. The resulting auto-segmentation models are distributed through the open-source platform CERR, accessible at https://github.com/cerr/CERR.

*Corresponding author: iyera@mskcc.org

INTRODUCTION

Delineating organs at risk (OAR) is a key component of radiotherapy treatment (RT) planning to limit post-treatment complications. Manual delineation is time-consuming, subjective, and prone to errors due to factors such as complexity and level of experience [1]. Therefore, there has been a great deal of interest in automating this process, to produce accurate and reproducible segmentations in a time-efficient manner.

In head and neck (H&N) cancer treatment, isolating the larynx and pharyngeal constrictor muscles is of particular interest to limit speech dysfunction and dysphagia following RT [2] [3]. Additionally, the masseters and medial pterygoids have been identified [4] as critical structures in developing radiation-induced trismus. However, delineating these structures is challenging due to their complex morphology and low soft tissue soft contrast in CT images. In this work, we present an auto-segmentation method to delineate these structures and examine its suitability for clinical use.

Few semi-automatic and automatic methods have been previously developed to segment OARs in the head and neck. Traditional multi-atlas-based auto-segmentation (MABAS) methods involve propagating and combining manually-segmented OARs from a library of CT scans by registering them with the CT scan to be segmented as in [5] and [6]. Other approaches offer strategies to further refine MABAS using organ-specific intensity [7] and texture [8] features or shape representation models [9] [10]. However, MABAS is sensitive to inter-subject anatomical variations as well as image artifacts, and image registration is computationally intensive, requiring several minutes even with highly efficient implementations [11].

Convolutional neural networks (CNNs) have recently been applied successfully to various medical image segmentation applications. Ibragimov et al. [12] trained 13 CNNs, applied in sliding-window fashion to segment H&N OARs including the larynx and pharynx, which were further refined using MRF-based post-processing. In [13], Ward van Rooij et al. employed the popular 3D U-net [14] architecture to segment H&N OARs including the pharyngeal constrictor muscle with encouraging results. Zhu et all. [11] extended the U-Net model by incorporating squeeze-and excitation (SE) residual blocks and a modified loss function to improve segmentation of smaller structures such as the chiasm and optic nerves. In [15], Men et al. segmented nasopharyngeal tumor volumes in H&N CT using a modified version of the VGG-16 [16] architecture, replacing fully-connected layers with fully-convolutional layers and introducing improved decoder networks to rebuild high-resolution feature maps. Tong et al [17] trained a fully convolutional neural net (FCNN), incorporating prior information by training a shape

representation model to regularize shape characteristics of 9 H&N OARs. The FocusNet [18] developed by Gao et al. utilizes multiple CNNs to segment H&N OARs including the larynx, first segmenting large structures, then identifying the centers of smaller structures within the surrounding region.

In this work, we present ensemble models developed from three different views (axial, sagittal, and coronal) to further improve segmentation accuracy of the OARs identified above and evaluate the clinical applicability of our approach.

MATERIALS AND METHODS

A. Dataset

CT scans of 243 H&N cancer patients from our institution were retrospectively collected under IRB 16-142 and accessed under IRB 16-1488 for the development of auto segmentation models in this work. This dataset was randomly partitioned into training (80%), validation (10%) and test (10%) sets. The validation set (10%) was used to estimate model performance while tuning hyperparameters, and the test set (10%) was used for unbiased evaluation of the final model. The dataset comprised CT images showing considerable variation in shape and appearance around the structures of interest, including slices with dental artifacts due to metallic dental implants. Training was performed using 2D slices originally of size 512x512 voxels. Additional data characteristics are listed in Table 1. Expert-delineated contours of the masseters (left, right), medial pterygoids (left, right) larynx, and constrictor muscles, generated using in-house tool Top Module, served as ground truth. Ground-truth contours of the masseters and medial pterygoids were absent in CT scans of 96 patients, 8 of which also lacked contours of the constrictor muscles. These scans were accordingly excluded from training and validation of respective models.

Table 1. Data characteristics

Attribute	Median	Min	Max	5th percentile	95th percentile
No. slices	80	48	279	65	103
Voxel spacing (mm)	0.78	0.51	1.37	0.61	1.17

Slice thickness (mm)	3.00	1.25	6.00	2.50	3.27

B. CNN segmentation model

The DeepLabV3+ CNN [19] was selected due to its impressive performance on the PASCAL VOC 2012 and Cityscapes datasets. It has also been successfully applied to medical image segmentation, e.g. Elguindi et al. [26] for segmenting the prostate in MR and Haq et al. [27] to segment cardio-pulmonary substructures in CT [ref]. This model combines the advantages of an encoder network comprising multiple dilated convolution layers, applied in parallel at different rates to capture contextual information at different scales, with a decoder network capable of effectively recovering object boundaries.

In this work, training was performed using ResNet-101 [20] as the backbone of the encoder network. A publicly-distributed implementation [21] of DeepLabV3+ using the Pytorch framework was utilized, and a soft-max layer was appended to obtain voxel-wise probabilities. On attempting to train a single multi-class model, it was noted that the imbalance in class labels due to variation in the volumes of the OARs led to poor performance for smaller structures. Modifying the loss based on class frequencies did not produce significant improvement as previously observed in [18]. Consequently, three distinct models were developed- one each for the masticatory structures, the larynx, and the pharyngeal constrictors. A sequential framework was utilized in which each segmented OAR was used to improve localization of subsequently segmented structures. Additionally, an ensemble of three independently-trained models was developed per OAR group using 2D axial, sagittal, and coronal slices. As compared to 3D convolutional neural networks (CNNs) which are highly memory-intensive, training in 2D enabled us to employ a more complex CNN while still providing contextual information and increased redundancy from the three different orientations.

A high-performance GPU cluster with four NVIDIA GeForce GTX 1080Ti GPUs with 11GB of memory each was used in training. Batch-normalization was applied using mini-batches of 8 images, resized to 320 x 320 voxels. Input images were standardized through scaling the intensities to [0,1] and normalization to zero-mean and unit-variance. Data augmentation was applied to each 2D image through scaling, cropping, and rotations. The channels of the

networks were populated with three consecutive slices (axial, sagittal, and coronal, respectively, for the three models), and the cross-entropy loss was employed in training. Data preprocessing and export to HDF5 [28] format was performed using the Computational Environment for Radiological Research (CERR) [23]. Details of the individual models are presented in the following sections, and Table 2 lists the hyperparameters used.

B.1 Masticatory structures

A single model was trained to segment the left and right masseters (MML, MMR) and medial pterygoids (PML, PMR). H&N CT scans were automatically cropped from nose to shoulders where required. The index of the first row of the patient's outline was recorded on each axial slice and the gradient was calculated. The slice with the largest gradient was identified as the start of the nose. To detect the shoulders, the maximum width of the patient's outline was computed on axial slices below the nose, and the slice corresponding to a large, abrupt increase in width was selected. 2D axial, sagittal, and coronal images and corresponding masks were then extracted within a bounding box around the cropped outline for training. Anterior and posterior limits of the bounding box were determined from the first slice.

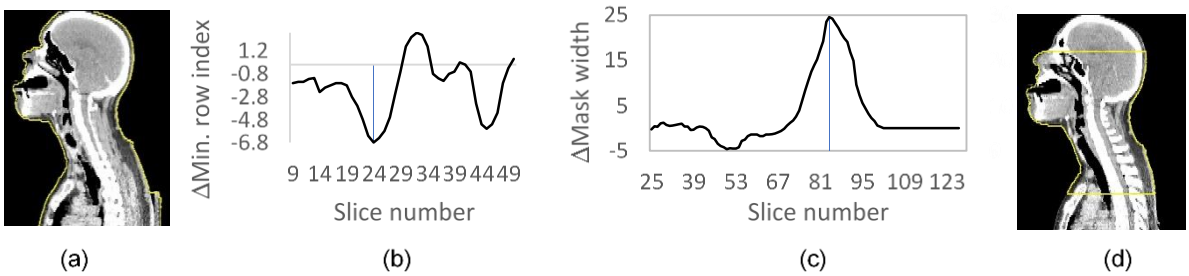


Fig 1. Automatic scheme for cropping axial H&N CT images from nose to shoulders. (a) Patient outline (b) Automatic identification of slice corresponding to start of nose (c) Automatic identification of shoulder slice (d) Cropped outline.

Optimization was performed using Stochastic Gradient Descent (SGD) with hyperparameters listed in Table 2 and learning rate was decayed following the polynomial scheduling policy. An early stopping strategy was employed to avoid overfitting if there was no improvement in the validation loss.

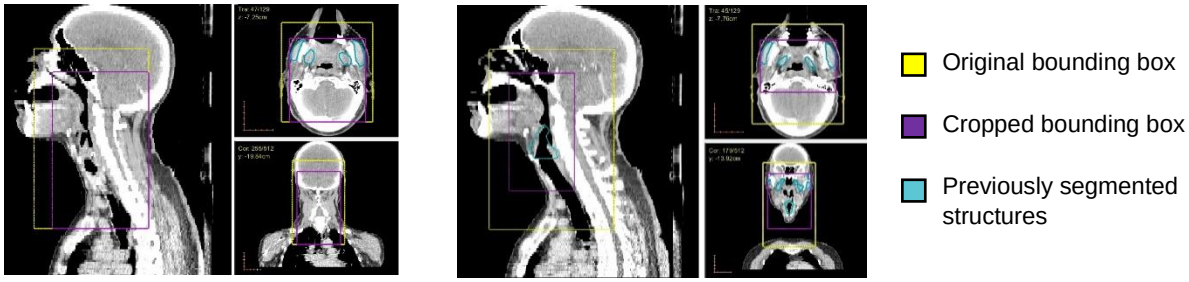
B.2. Swallowing structures

B.2.1 Larynx

CT scans were cropped further, with the anterior, left, right, and superior limits defined by the corresponding extents of the union of the masticatory structures. 2D axial, sagittal and coronal images and associated masks were extracted within the resulting bounding box and training was performed using SGD.

B.2.2 Constrictor muscles

To better localize the constrictor muscles (CM), CT scans were cropped with the anterior, left, right, and superior limits defined by the corresponding extents of the union of the masticatory structures. The posterior and superior limits were defined by corresponding extents of the larynx, with sufficient padding. 2D axial, sagittal, and coronal images were extracted within the resulting bounding box. Optimization was performed using adaptive moment estimation (Adam) [22].



(a) (b)

Fig 2. Automatic scheme for cropping patient outline to localize (a) the larynx using previously segmented masticatory structures and (b) the constrictor muscles using the larynx and masticatory structures.

For each of the above OARs, probability maps returned by the three models (axial, sagittal, coronal) were averaged and voxels were assigned to the class with the highest resulting probability, thereby producing stable consensus segmentations and providing robustness against occasional failures of the individual models. The masks were further post-processed to remove isolated voxels by discarding all but the largest connected component. Morphological processing was performed to fill holes and obtain smooth contours.

Table 2. Summary of hyperparameters for training.

Model	Structure	Learning rate	Optimizer	Momentum	Weight Decay
1	Masticatory structures (MML,MMR)	Axial: 0.002 Sagittal: 0.003 Coronal: 0.002	SGD	0.9	0.0001
2	Larynx	Axial: 0.0003 Sagittal: 0.0003 Coronal: 0.0003	SGD	0.9	0.0002
3	Constrictor muscles	Axial: $1x10^{-6}$ Sagittal: $8x10^{-7}$ Coronal : $2x10^{-6}$	Adam	0.9	0.0001

C. RESULTS

The performance of the proposed framework on the test set of 24 H&N cancer patients was compared against manual segmentation using three evaluation metrics. The Dice Similarity Coefficient (DSC) was used to assess the degree of overlap between manual (A) and automated (B) segmentations, computed as:

$$DSC = 2\ \frac{|A \cap B|}{|A| + |B|}$$

The 95th percentile Hausdorff distance (HD_{95}), i.e., maximum distance between boundary points of A and B, was computed to capture the impact of a small number of large segmentation errors on the overall segmentation quality. Additionally, the differences in clinically relevant [29-30] dose-volume histogram (DVH) metrics were computed between deep learning-based and manual contours, and the Wilcoxon signed-rank test was performed to test for statistical disparities. The differences in DVH metrics for all structures except the constrictor muscles were not found to be statistically significant.

Table 3. DSC and HD_{95} for segmentations resulting from the proposed framework. Of the 24 test patients, the no. with ground truth contours available for comparison is noted in the third column.

Structure	**DSC (mean ± std. deviation)**	**HD_{95}mm (mean ± std. deviation)**	**No. test patients**
MML	0.8693 ± 0.0206	3.6324 ± 1.2590	11
MMR	0.8755 ± 0.0231	3.4292 ± 0.6232	11
PML	0.8139 ± 0.0258	3.9558 ± 0.8048	11
PMR	0.8010 ± 0.0321	4.4646 ± 1.0772	11
Larynx	0.8335 ± 0.0405	4.7215 ± 2.0408	24
CM	0.6741±0.0703	4.9937± 1.8191	22

Table 4. Comparison of relevant DVH metrics extracted using manual and auto-generated contours. Median, first and third quartiles of percentage differences are presented.

Structure	**Metric**	**Difference (%)**	**p-value**	**No. test patients**
Masseters	Ipsilateral mean dose	0.8136(-0.1657, 2.2830)	0.8864	11
Medial pterygoids	Ipsilateral mean dose	0.2387(-0.9486, 1.1062)	0.1475	11
Larynx	Mean dose	0.0667(-2.0534, 2.3050)	0.6377	24
Constrictor muscle	Mean dose	-0.7090(-1.5109, 0.2682)	**0.0363**	22

Figure 2 shows examples of segmentations resulting from our algorithm for qualitative comparison.

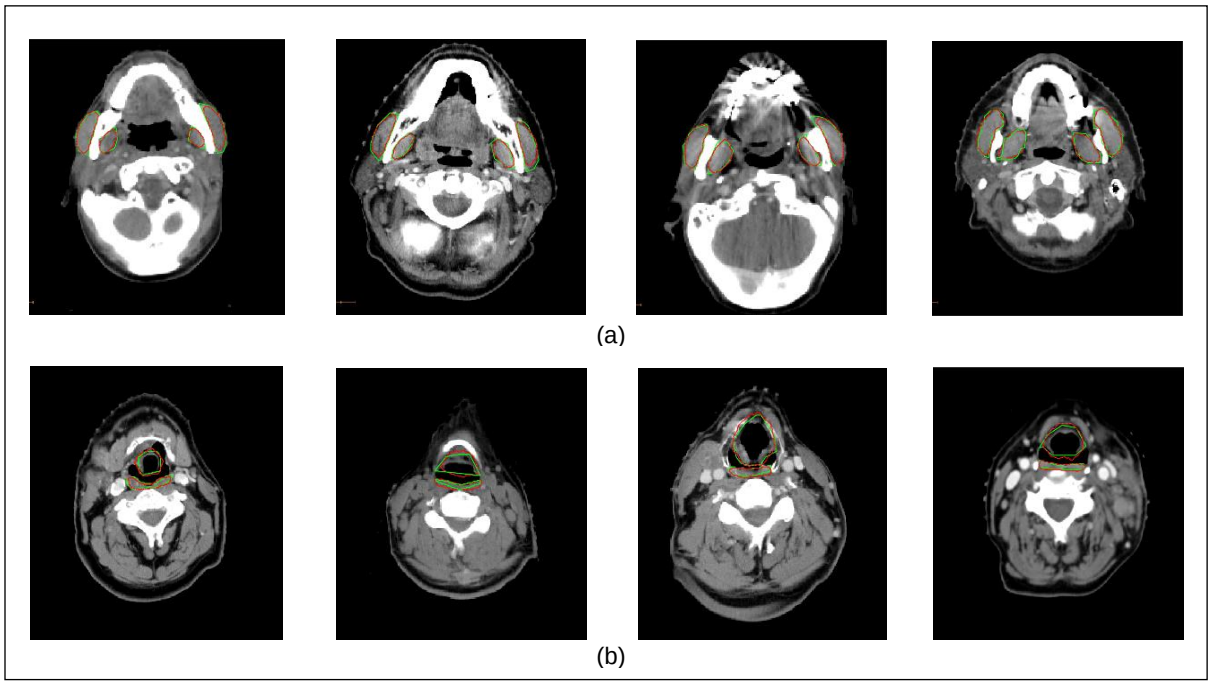


Fig 2. Segmentation results for (a) chewing and (b) swallowing structures, shown in four axial cross-sections. Ground truth contours are depicted in green and deep-learning-based segmentations in red.

D. Discussion

The proposed method exceeds or is comparable to the performance of previously reported auto-segmentation methods [5,12,13,18,24,16]. Table 5 shows the performance of our algorithm alongside previously reported results.

Table 5. DSC (mean ± std. deviation) for swallowing and chewing structures for the proposed framework and those reported in previously published works.

Structure	Proposed method	Tao et al. [5]	Ibragimov et al. [12]	van Rooij et al. [13]	Gao et al. [18]	Thomson et al. [24]	Han et al. [26]

							(*leave-one-out validation)
Masseters (MML,MMR)	0.87 ± 0.02	-	-	-	-	-	0.83
Medial Pterygoids (PML,PMR)	0. 81 ± 0.03	-	-	-	-	-	0.84
Larynx	0.834 ± 0.04	0.73 ± 0.04,	0.856 ± 0.04	0.78 ± 0.05	0.66±0.29	0.58	-
CM	0.67±0.07	0.65 ± 0.06	0.69 ± 0.06	0.68 ± 0.09	-	0.5	-

E. CONCLUSION

We developed a fully automatic method for segmenting swallowing and chewing structures on CT images using Deep Learning. We validated our approach against manual delineation by an expert and demonstrated its potential for use in RT treatment planning

REFERENCES

1. Harari, Paul M., Shiyu Song, and Wolfgang A. Tomé. Emphasizing conformal avoidance versus target definition for IMRT planning in head-and-neck cancer. *International Journal of Radiation Oncology* Biology* Physics* 77.3 (2010): 950-958.
2. Choi, Mehee, et al. Development of a standardized method for contouring the larynx and its substructures. *Radiation Oncology* 9.1 (2014): 285.
3. Levendag, P. C., 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. *Radiotherapy and Oncology* 85.1 (2007): 64-73.
4. Kraaijenga, Sophie A., 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* 41.5 (2019): 1387-1394.
5. Tao, Chang-Juan, et al. Multi-subject atlas-based auto-segmentation reduces interobserver variation and improves dosimetric parameter consistency for organs at risk in

nasopharyngeal carcinoma: a multi-institution clinical study. *Radiotherapy and Oncology* 115.3 (2015): 407-411.
6. Levendag, P. C., et al. Atlas based auto-segmentation of ct images: clinical evaluation of using auto-contouring in high-dose, high-precision radiotherapy of cancer in the head and neck. *International Journal of Radiation Oncology• Biology• Physics* 72.1 (2008): S401.
7. Fortunati, Valerio, et al. Tissue segmentation of head and neck CT images for treatment planning: a multiatlas approach combined with intensity modeling. *Medical physics* 40.7 (2013): 071905.
8. Qazi, Arish A., et al. Auto-segmentation of normal and target structures in head and neck CT images: A feature-driven model-based approach. *Medical physics* 38.11 (2011): 6160-6170.
9. Wang, Zhensong, et al. Hierarchical vertex regression-based segmentation of head and neck CT images for radiotherapy planning. *IEEE Transactions on Image Processing* 27.2 (2017): 923-937.
10. Fritscher, Karl D., et al. Automatic segmentation of head and neck CT images for radiotherapy treatment planning using multiple atlases, statistical appearance models, and geodesic active contours. *Medical physics* 41.5 (2014): 051910.
11. Zhu, Wentao, et al. AnatomyNet: Deep learning for fast and fully automated whole-volume segmentation of head and neck anatomy. *Medical physics* 46.2 (2019): 576-589.
12. Ibragimov, Bulat, and Lei Xing. Segmentation of organs-at-risks in head and neck CT images using convolutional neural networks. *Medical physics* 44.2 (2017): 547-557.
13. van Rooij, Ward, et al. Deep Learning-Based Delineation of Head and Neck Organs at Risk: Geometric and Dosimetric Evaluation. *International Journal of Radiation Oncology* Biology* Physics* 104.3 (2019): 677-684.
14. Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *International Conference on Medical image computing and computer-assisted intervention*. Springer, Cham, 2015.
15. Men, Kuo, et al. Deep deconvolutional neural network for target segmentation of nasopharyngeal cancer in planning computed tomography images. *Frontiers in oncology* 7 (2017): 315.
16. Simonyan, Karen, and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).
17. Tong, Nuo, et al. Fully automatic multi-organ segmentation for head and neck cancer radiotherapy using shape representation model constrained fully convolutional neural networks. *Medical physics* 45.10 (2018): 4558-4567.
18. Gao, Yunhe, et al. FocusNet: Imbalanced Large and Small Organ Segmentation with an End-to-End Deep Neural Network for Head and Neck CT Images. *arXiv preprint arXiv:1907.12056* (2019).
19. Chen, Liang-Chieh, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation. *Proceedings of the European conference on computer vision (ECCV)*. 2018.
20. He, Kaiming, et al. Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.
21. https://github.com/jfzhang95/pytorch-deeplab-xception.git
22. Kingma, Diederik P., and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*(2014).
23. Deasy, Joseph O., Angel I. Blanco, and Vanessa H. Clark. CERR: a computational environment for radiotherapy research. *Medical physics* 30.5 (2003): 979-985.

24. Thomson, David, et al. Evaluation of an automatic segmentation algorithm for definition of head and neck organs at risk. *Radiation Oncology* 9.1 (2014): 173.
25. Han, Xiao, et al. Atlas-based auto-segmentation of head and neck CT images. *International Conference on Medical Image Computing and Computer-assisted Intervention*. Springer, Berlin, Heidelberg, 2008.
26. Elguindi, Sharif et al. Deep learning-based Auto-Segmentation of Targets and OARs for MR-Only planning of Prostate Radiotherapy. *Radiotherapy and Oncology*. Manuscript under review.
27. Haq, Rabia, et al. Cardio-pulmonary Substructure Segmentation of CT images using Convolutional Neural Networks for Clinical Outcome Analysis*. International Conference on Medical Imaging with Deep Learning -- Extended Abstract Track*. 2019.
28. Folk, Mike, Albert Cheng, and Kim Yates. HDF5: A file format and I/O library for high performance computing applications. *Proceedings of supercomputing*. Vol. 99. 1999.
29. Tsai, Chiaojung Jillian, et al. Modeling dose response for late dysphagia in patients with head and neck cancer in the modern era of definitive chemoradiation. *JCO clinical cancer informatics* 1 (2017): 1-7.
30. Rao, Shyam D., et al. Dose-volume factors correlating with trismus following chemoradiation for head and neck cancer. *Acta Oncologica* 55.1 (2016): 99-104.