Appendix A - Data Processing

All patients received diagnostic contrast-enhanced computed tomography (CECT) imaging. To establish tumor location with respect to organs-at-risk (OARs), contours for GTVs and at least 41 OARs were contoured on individual patient's scans. GTVs were manually contoured, while OARs were automatically segmented using a previously validated approach [1]. These contours were then used to extract GTV and OAR volumes, as well as the minimum distances between the surface of each region of interest (GTV and OARs), allowing some distances to be negative if they overlapped due to concave-convex adjacency and partial volume effects. Mean dose values to each ROI were further extracted from each patient's radiation plans.

The minimum euclidean distances between the outer contour of each ROI and individual GTVs were denoised using a denoising autoencoder [2], which maps the original data to a lower-dimensional space and then attempting to reconstruct the original values, which serves a method of filtering noise in the data by looking at the values as a whole. The denoising autoencoder consisted of a neural net using a single hidden layer of 90 units and an output layer of 45 units that attempted to predict the original values given a corrupted version of the original inputs, which both used rectified linear unit (relu) activation functions. Gaussian dropout [3] with a dropout rate of .1, meaning 10% of the input values were randomly set to zero at each training step for the autoencoder, was applied to the input layer. Gaussian noise with a standard deviation of .5 was then applied to the hidden layer during training. By applying dropout and noise during training the encoder then learns to filter out noise in the input data using the other organ-tumor distances, given that these values are highly related. Training was done using 800 epochs with a batch size of 4. For training we used the Adam optimizer [4] with a learning rate of .0001. Once trained, all tumor-organ distances were passed through the autoencoder to create a denoised dataset. For patients with multiple GTVs, inter-organ distances were denoised separately before calculating a global minimum distance for each non-target ROI The denoising autoencoder was implemented in python with tensorflow [5].

Appendix B - Dose Prediction

Tumor prediction was done using a k-nearest-neighbors approach using T-ssim similarity as the distance function, as in [6]. Tumor-organ distances as well as volumes were used to calculate pairwise similarity between each patient. In the original paper, the weighted mean of the tumor distances were used in the case of a patient with 2 GTVs. Since our new dataset includes patients with up to 8 tumors, the minimum tumor-organ distances were used instead, as it was shown to be more effective in comparing patients with a large number of tumors. Additionally, when considering patient similarity, we included both the original patient cohort and a symmetric copy of the cohort, where we flipped the patients' data across the midline of the head, to account for lateral symmetry. No group archetypes were used, and patient matches were determined by including all patient matches with a similarity of .94 or more, or the 8 most similar patients, whichever was greater. These values were determined via linear search optimization. Mean values for the true doses, predicted doses, and organ-tumor distances used in the prediction for each ROI group are in Table B.1.

Table B.1. ROIs considered during dose prediction and clustering analysis. For bilateral organs, we report statistics for the mean values of each ROI pair. True planned dose and predicted dose both refer to the mean total dose delivered to an ROI over the entire treatment period. Organ-tumor distance represents the inter-contours distances between segmented ROIs. Negative values denote overlap between the GTV and ROI contour.

	True Planned Dose (Gy)		**Predicted Dose (Gy)**		**Organ-Tumor Distance (mm)**	
Organ	**Mean**	**Std**	**Mean**	**Std**	**Mean**	**Std**
Mandible	39.4	7.3	39.5	3.8	4.7	4.6
Extended Oral Cavity	51.7	7.1	52	3.7	-9.4	4.8
Medial Pterygoid Muscle Avg.	54	6.9	53.9	3.9	11.4	8.2
MPC	58.7	10.4	59.1	4.3	8.4	6.2
Esophagus	29.1	10	29.1	3.5	38.2	21.1
Spinal Cord	26	4.7	26.1	1.8	21	10.4
Cricopharyngeal Muscle	20.1	13.6	19.3	5.1	23.8	14.5
Cricoid Cartilage	24.7	12.7	24.3	4.6	21.6	13.1
IPC	33.3	16.8	33	7.7	13	8.7
Brainstem	13.2	4.5	13.1	1.7	29.4	16.2
Larynx	26.4	13.3	25.7	6.4	11.2	8
Thyroid Cartilage	38.1	11.6	38.1	6.2	7.9	7.2
Supraglottic Larynx	52.9	11.8	53.4	5.9	1.2	5.9
SPC	63.1	6.3	63.3	3.1	-4	3.6

Hyoid Bone	62.9	11.7	63.6	4	3	5.4
Soft Palate	58.1	9.4	58.2	5.1	3.1	9
Genioglossus Muscle	61.1	8.2	61.5	3.9	-3.4	5
Tongue	56.6	8.3	56.8	3.6	-2.9	5.5
Mylogeniohyoid Muscle	54.9	10.1	55.3	5	5.4	6.1
Hard Palate	25.3	10.1	24.9	4.4	28.6	18.4
Lower Lip	23.4	6.8	23.6	2.1	41.7	20.4
Upper Lip	16.6	6	16.5	2.2	46.9	23.4
Brachial Plexus Avg.	48.8	8.8	48.9	4.1	22.4	13
Thyroid Lobe Avg.	48.6	9.4	48.9	3.7	30.3	17.6
Sternocleidomastoid Muscle Avg.	56.1	8.8	56.2	4.1	11.9	10.5
Mastoid Avg.	40.8	9	40.8	3.8	36.4	19
Parotid Gland Avg.	28.5	6.9	28.5	3.5	15.8	10.3
Lateral Pterygoid Muscle Avg.	35.2	10.2	34.8	4.6	27.5	16.3
Masseter Muscle Avg.	29.7	6.6	29.5	3.2	22.4	12
Submandibular Gland Avg.	61.9	8.8	62.1	4.5	6.5	6.7
Anterior Digastric Muscle Avg.	54.1	9.5	54.4	4.8	9.9	7.7

Table B.2: Predicted dose mean and standard deviation by cluster. Dose values for bilateral organs report the mean values between both organs. Swallowing-related muscles are highlighted.

	Spatial Cluster 1		Spatial Cluster 2		Spatial Cluster 3		Spatial Cluster 4	
	Mean	**Std**	**Mean**	**Std**	**Mean**	**Std**	**Mean**	**Std**
Esophagus	31.4	0.5	29.7	0.8	24.5	5.7	30.9	2.6
Spinal Cord	26.9	0.3	26.0	0.4	24.1	2.9	27.6	1.3
Cricopharyngeal Muscle	21.9	2.0	18.3	1.8	14.7	4.4	25.0	6.0
Cricoid Cartilage	26.7	1.3	23.4	1.6	20.3	3.9	29.4	5.6
IPC	40.9	1.5	31.3	4.5	28.1	6.5	39.9	9.9
MPC	61.4	0.9	58.8	2.0	54.6	5.5	63.1	3.6
Brainstem	12.4	0.8	12.4	1.0	13.2	1.7	14.8	1.8

Larynx	30.3	0.5	24.0	2.8	21.3	4.0	32.8	8.3
Thyroid Cartilage	41.5	0.7	36.2	2.4	34.6	4.4	45.1	8.1
Supraglottic Larynx	58.6	1.0	52.0	3.4	49.9	6.9	58.7	6.4
SPC	63.4	1.4	62.7	1.1	60.0	3.8	66.9	2.1
Hyoid Bone	66.4	0.8	63.0	2.6	60.7	5.6	66.9	3.2
Soft Palate	53.8	2.3	56.7	4.0	55.1	3.4	64.4	3.3
Genioglossus Muscle	65.5	1.9	61.1	2.4	57.9	4.9	64.8	3.3
Tongue	58.6	2.5	56.0	1.4	53.6	4.1	61.1	2.8
Mylogeniohyoid Muscle	60.5	3.3	54.3	2.9	51.4	4.9	60.2	5.3
Extended Oral Cavity	52.0	2.0	50.8	1.4	48.7	3.2	57.1	2.9
Mandible	39.9	2.2	38.2	1.0	36.3	2.8	44.8	3.7
Hard Palate	20.6	1.3	23.2	3.1	24.2	3.0	29.9	4.0
Lower Lip	22.3	1.4	22.6	1.0	23.8	2.1	26.0	2.1
Upper Lip	15.7	1.3	15.6	1.3	16.2	1.9	18.7	2.6
Brachial Plexus (LR)	53.0	1.8	48.7	1.8	44.3	4.8	52.7	3.8
Thyroid Lobe (LR)	52.6	1.4	48.8	1.5	44.5	5.0	52.2	2.8
Sternocleidomastoid Muscle (LR)	59.1	0.4	56.3	1.5	50.9	5.7	59.5	2.9
Mastoid (LR)	40.7	1.3	39.7	1.8	37.7	3.4	45.6	3.2
Parotid Gland (LR)	28.8	2.1	27.5	1.1	25.5	2.2	33.2	3.6
Medial Pterygoid Muscle (LR)	53.6	0.7	53.1	1.2	49.6	3.8	59.0	2.9
Lateral Pterygoid Muscle (LR)	31.5	0.6	32.9	3.2	32.9	3.0	40.8	2.9
Masseter Muscle (LR)	28.9	0.6	28.8	1.5	25.9	2.9	33.6	2.0
Submandibular Gland (LR)	66.0	1.6	61.5	2.2	57.7	5.9	66.3	3.9

Anterior Digastric Muscle (LR)	59.7	4.1	53.3	2.8	51.0	4.6	59.3	5.1

Appendix C. - Covariate Selection

Our model uses Hierarchical clustering to segment the cohort, which is a standard data-mining technique for identifying patterns in the data. The most important consideration for this is the way in which we define similar. We postulated that the anatomical and spatial characteristics of the GTV and surrounding organs can be used as a source of similarity that isn't well captured in existing literature. Because the root cause of RAD is hard to attribute to individual organs, we considered the volume and minimum distance to GTV for all nearby ROIs in the head and neck to be potential covariates. Since treatment dose to non-target ROIs is considered to be the main driver of RAD, we further estimated these doses using the available covariates based on previously published methods. We then performed a search over these covariates to identify a representative subset of them to use in our final model.

Our set of available candidate features were 41 sets of tumor-organ distances, predicted mean treatment doses, and contour volumes. Of these 41 organs, 10 pairs of bilaterally symmetrical organs were combined using their euclidean norm, resulting in 31 composite ROIs for each covariate type. With these 93 candidate covariates, we performed a search to identify the most representative values to use in our final model. Because the results of clustering are dependent on the dataset used, we need to avoid overfitting the model that is only discriminative on the current cohort. To avoid this, we used bagging to estimate true distribution of clusters for each variable.

Concretely, we randomly sampled from the original dataset with replacement, such that the new dataset had the same number of patients as in the original dataset. Within this processed dataset, we performed agglomerative clustering using a weighted linkage function and 2-5 clusters on each individual candidate covariate. For each of the clustering results, the correlation between these clusters and RAD was measured using a two-tailed version of Fisher's exact test [7], and the inverse of the smallest p-value (among all options of 2-5 clusters) was used as the 'importance' for the candidate covariate. We chose Fisher's exact test over a standard Chi-squared test, because the exact test works well on small numbers of samples. Bagging was performed 500 times, and the mean importance for each variable was then used. The covariate with the strongest correlation, and therefore the most discriminative clusters, was then choses as a static covariate. This process was repeated, such that all non-static variables were combined individually with the static features, and a new ‘importance’ was measured. At each step, the most important variable was added to the static covariates until no covariates were found to improve correlation above the set of static covariates. These selected covariates were then used to produce the spatial clusters as described below.

Appendix D. - Univariate Feature Analysis

We performed univariate analysis of the spatial features to compare the 5 features' individual effect on baseline prediction. We report the change in cross-validation AUC score using logistic regression relative to the baseline features when including individual features, as in the previous analysis, both with or without performing clustering using these individual features. We additionally analyzed several other features identified as of-interest in recent literature [8,9], to provide a comparison: doses to the mandible or anterior-digastric muscle [10,11], as well as the mean volume of the sternocleidomastoids, masseter volume, and total gross tumor volume, which are shown in figure D.1. We further compared the AUC score increase from performing logistic regression with the addition of the tumor-organ distances for the 6 muscles related to swallowing: MPC, SPC, IPC, Cricopharyngeal Muscle, the Mylogeniohyoid Muscle, and the Supraglottic Larynx, as they are likely candidates for a causal model of dysphagia, which are reported in Figure D.2. For each feature set, we report the change mean difference from the baseline AUC score after including each feature.

While we identify features that allow for robust segregation of the patient cohort in an unsupervised manner, our approach does not capture individual features that may not cluster well, despite providing strong correlation with endpoints. Our univariable analysis on the effect of using individual organs in the dataset yielded low to no improvement when clustering was applied. The contribution of the clustered features was also not correlated with the improvement provided by unclustered features. Masseter volume produced a better AUC improvement than any of our individual proximity and dose features, consistent with other literature that found masseter dose-volume was related to swallowing dysfunction. However, masseter volume provided slightly worse results if clustering was applied. For our analysis of swallowing muscles, many of these muscles provide limited to no additional benefit to the predictive model, suggesting that they are likely fully encapsulated by other factors such as T-category.

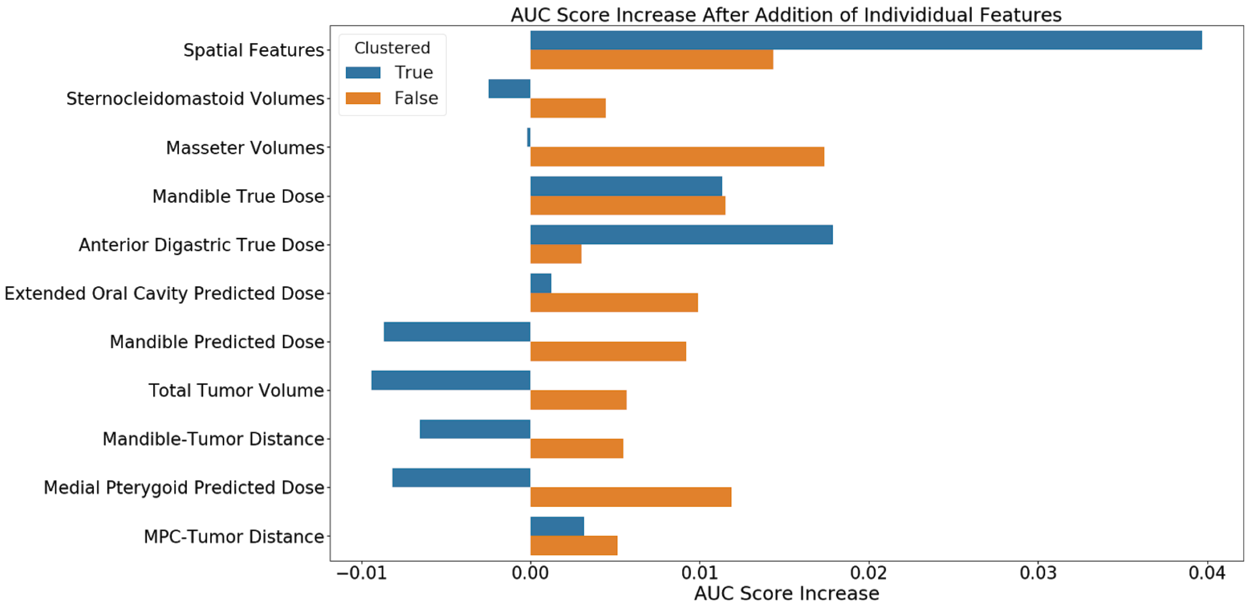

Figure D.1. Change in AUC score relative to baseline clinical features when introducing selected features with and without clustering. Features tested include the 6 spatial features together, and individually, as well as Masseter Volume, and true doses the the Anterior Digastric, and Mandible.

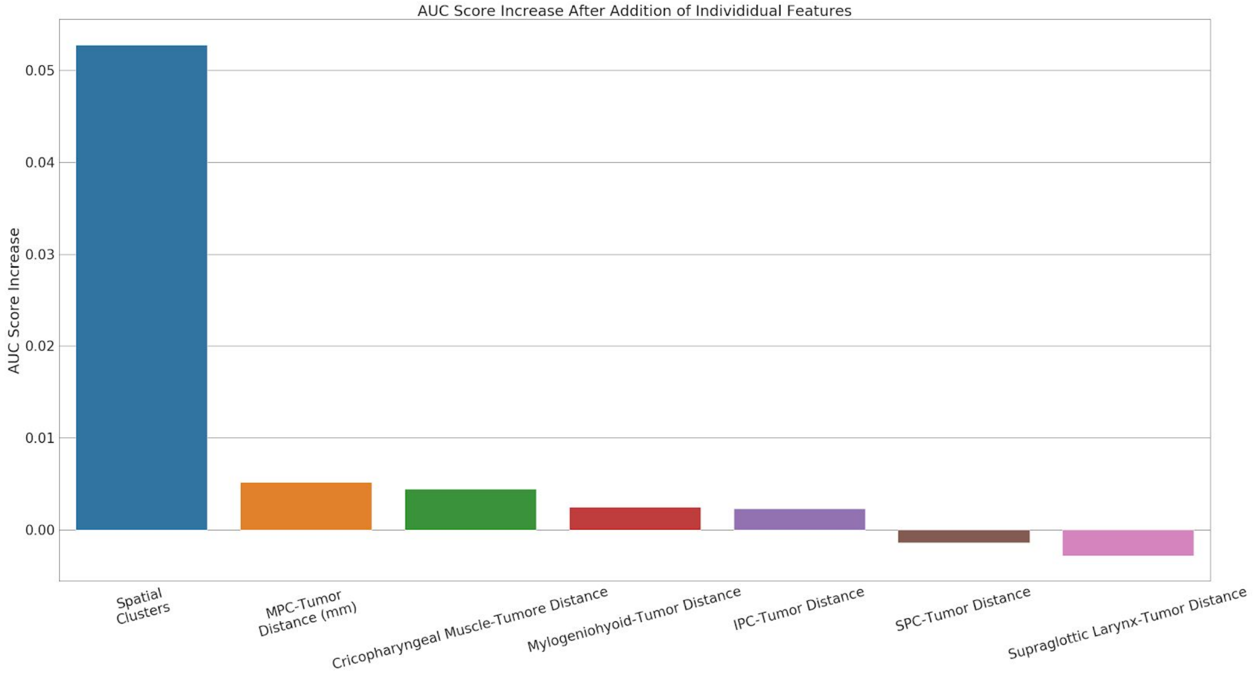

Figure D.2. Change in AUC score relative to baseline clinical features when introducing Swallowing Muscle.

We further created a logistic regression model to predict the probability of a patient being not in the high-risk group using the tumor-organ distances of the 6 swallowing organs: SPC, IPC, MPC, Cricopharyngeal Muscle, Supraglottic Larynx, and the Mylogeniohyoid muscle. By using a logistic regression model to estimate the likelihood of being in a low risk group, the resulting odds ratios can thus be interpreted as the relative importance that proximity to an organ represents for being in the high-risk group. The odds ratios from the fitted model are shown in Table D.1. Of the ROIs with a positive odds ratio, SPC was the most indicative (OR = 1.64, 97.5% CI = [1.41, 1.98]), suggesting that proximity to the SPC was the strongest single swallowing muscle for predicting RAD, followed by the Mylogeniohyoid (OR = 1.23, 97.5% CI = [1.05, 1.45]), and finally IPC (OR = 1.13, 97.5% CI = [.819, 1.59]). The logistic regression mimic model achieved an accuracy of 83.5%, with an AUC score for predicting RAD of 0.64.

Table D.1: Odds-ratios for tumor-organ distance and membership in the low-risk spatial cluster. Higher odds ratios indicate that tumors near the given ROI are more likely to be in the high risk group.

Organ	Odds-Ratio	2.5% CI	97.5% CI
Superior Pharyngeal Constrictor	1.64	1.41	1.98
Mylogeniohyoid Muscle	1.23	1.05	1.45
Inferior Pharyngeal Constrictor	1.13	.819	1.59
Cricopharyngeal Muscle	.996	.844	1.17
Medial Pharyngeal Constrictor	.963	.782	1.19
Supraglottic Larynx	.845	.701	1.01

Appendix E - Clinical Feature Clustering Analysis

Clustering was performed on only the non-spatial clinical features for each patient, to provide a baseline comparison of how well currently existing features perform when clustering. TMN staging information as well as common demographic and clinical features were collected from each patient in the cohort. All features except for age and dose-to-tumor were one-hot encoded [12]. Clustering was performed on these features using hierarchical clustering with the Manhattan distance function, which was chosen for its ability to work well with categorical data. We report results for *k* = 4 clusters, where *k* was chosen as it resulted in good correlation with RAD using Fisher’s exact test (p < .0001). For cases where the AJCC 8th edition classification was missing (51), the values were estimated using the patient’s AJCC 7th edition classification as all such patients were hpv negative. Categorical variables were encoding using dummy variables with one-hot encoding. Distance was computed using the manhattan distance with k-medoids clustering, which was chosen as it had the strongest correlation with RAD. Cluster breakdowns are reported in Table E.1. Cluster labels were significantly correlated with toxicity outcomes.

Table E.1: Clinical Cluster Characteristics

Clinical Cluster Characteristics				
Feeding Tube				
Cluster	**# Patients**	**# W/ Toxicity**	**% W/ Toxicity**	**P-Value**
Clinical Cluster 1	62	9	14.5	< 0.01
Clinical Cluster 2	59	2	3.4	
Clinical Cluster 3	42	10	23.8	
Clinical Cluster 4	37	1	2.7	
Aspiration				
Cluster	**# Patients**	**# W/ Toxicity**	**% W/ Toxicity**	**P-Value**
Clinical Cluster 1	62	10	16.1	< 0.01
Clinical Cluster 2	59	1	1.7	
Clinical Cluster 3	42	8	19	
Clinical Cluster 4	37	0	0	
RAD (Either)				
Cluster	**# Patients**	**# W/ Toxicity**	**% W/ Toxicity**	**P-Value**
Clinical Cluster 1	62	16	25.8	< 0.0001
Clinical Cluster 2	59	3	5.1	
Clinical Cluster 3	42	14	33.3	
Clinical Cluster 4	37	1	2.7	

AUC cross-validation score was calculated for these clinical cluster labels in combination with the original clinical features, the spatial clusters, and alone, which are reported in Table F.2. As these clinical cluster labels performed worse than T-staging alone, we instead compare our spatial clusterings to T-stage in the main results.

Table E.2: AUC cross-validation scores using logistic regression alone and in addition to other features. Clusters performed worse alone and in combination with spatial clusters than T-staging alone, and did not change performance when using all clinical features.

Leave-one-out Cross-Validation AUC Scores (Logistic Regression)			
	Feeding Tube	Aspiration	RAD (Either)

Clinical Clusters	0.64	0.66	0.68
Clinical Clusters + Spatial Clusters	0.71	0.74	0.73
All Clinical Features +Clinical Clusters	0.64	0.85	0.79

References

[1] A.S.R. Mohamed, M.-N. Ruangskul, M.J. Awan, C.A. Baron, J. Kalpathy-Cramer, R. Castillo, E. Castillo, T.M. Guerrero, E. Kocak-Uzel, J. Yang, L.E. Court, M.E. Kantor, G. Brandon Gunn, R.R. Colen, S.J. Frank, A.S. Garden, D.I. Rosenthal, C.D. Fuller, Quality Assurance Assessment of Diagnostic and Radiation Therapy–Simulation CT Image Registration for Head and Neck Radiation Therapy: Anatomic Region of Interest–based Comparison of Rigid and Deformable Algorithms, Radiology. 274 (2015). https://doi.org/10.1148/radiol.14132871.

[2] Vincent, Pascal, et al. "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion." *Journal of machine learning research* 11.Dec (2010): 3371-3408.

[3] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research, 2014.

[4] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.

[5] Mart´ın Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: A system for large-scale machine learning. In 12th Symposium on Operating Systems Design and Implementation, 2016.

[6] A Wentzel, P Hanula, T Luciani, B Elgohari, H Elhalawani, G Canahuate, D Vock, CD Fuller, and GE Marai. Cohort-based t-ssim visual computing for radiation therapy prediction and exploration. IEEE transactions on visualization and computer graphics, 2019.

[7] Upton, Graham JG. "Fisher's exact test." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 155.3 (1992): 395-402.

[8] Dale, Timothy, et al. "Beyond mean pharyngeal constrictor dose for beam path toxicity in non-target swallowing muscles: Dose–volume correlates of chronic radiation-associated dysphagia (RAD) after oropharyngeal intensity modulated radiotherapy." *Radiotherapy and Oncology*, 2016.

[9] Christopherson, Kaitlin M., et al. "Chronic Radiation-Associated Dysphagia in Oropharyngeal Cancer Survivors: Towards Age-Adjusted Dose Constraints for Deglutitive Muscles." *Clinical and Translational Radiation Oncology*, 2019.

[10] Kumar, Rachit, et al. "Radiation dose to the floor of mouth muscles predicts swallowing complications following chemoradiation in oropharyngeal squamous cell carcinoma." *Oral oncology*, 2014

[11] Timothy Dale, Katherine Hutcheson, Abdallah SR Mohamed, Jan S Lewin, G Brandon Gunn, Arvind UK Rao, Jayashree Kalpathy-Cramer, Steven J Frank, Adam S Garden, Jay A Messer, et al. Beyond mean pharyngeal constrictor dose for beam path toxicity in non-target swallowing muscles: Dose–volume correlates of chronic radiation-associated dysphagia (rad) after oropharyngeal intensity modulated radiotherapy. Radiotherapy and Oncology, 2016.

[12] Feinstein, Alvan R. "Clinical biostatistics." *Clinical Pharmacology & Therapeutics* 22.4 (1977): 485-498.