Multi-Organ Spatial Stratification of 3-D Dose Distributions Improves Risk Prediction of Long-Term Self-Reported Severe Symptoms in Oropharyngeal Cancer Patients Receiving Radiotherapy: Development of a Pre-Treatment Decision Support Tool.

†-Contributing authors:

Andrew Wentzel*[a], Abdallah S. R. Mohamed, MD, PhD[b], Mohamed A. Naser, PhD[b], Lisanne V. van Dijk, PhD[b], Katherine Hutcheson, PhD[b], Amy M. Moreno, MD[b], Clifton D. Fuller, MD, PhD[b], Guadalupe Canahuate[c], G. Elisabeta Marai[a]

Affiliations:

[a]Department of Computer Science, The University of Illinois Chicago, Chicago, IL, USA.

[b]Department of Radiation Oncology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA.

[c]Department of Electrical and Computer Engineering, University of Iowa, Iowa City, IA, USA.

Keywords

Radiation therapy; Oropharynx cancer; Head and Neck cancer; Stratification; Symptom Burden; Quality of Life

Prior presentation:

None

Statement of Impact

Recent advances in precision medicine for treating Oropharyngeal cancer (OPC) have led to significant improvements in survival for patients. However, OPC survivors often are left with severe side effects that result from damage to vital tissues in the head and neck caused by secondary radiation received during treatment. These side effects can persist for months after treatment cessation, and diminish the patients' quality of life. Existing guidelines for radiation therapy plans identify dose limits to individual organs with the goal to avoid or minimize permanent tissue damage and their related toxicity. However, existing tools fail to account for either the three-dimensional nature of dose distributions, or the fact that critical organs often work as a system that can have multiple failure points.

We present an unsupervised approach for stratifying patients based on three-dimensional dose distributions to multiple organs and evaluate their association with long-term patient-reported symptoms. We apply this method on predicting severe patient reported symptoms 6 months after ceasing radiation therapy for OPC patients. For actionable interpretability, we introduce a

rule-mining method that simplifies our strata into a set of dose-thresholds, and demonstrate that these thresholds outperform existing dose treatment guidelines.

Abstract

Purpose: Identify Oropharyngeal cancer (OPC) patients at high-risk of developing long-term severe radiation-associated symptoms using dose volume histograms for organs-at-risk, via unsupervised clustering.

Material and Methods: All patients were treated using radiation therapy for OPC. Dose-volume histograms of organs-at-risk were extracted from patients' treatment plans. Symptom ratings were collected via the MD Anderson Symptom Inventory (MDASI) given weekly during, and 6 months post-treatment. Drymouth, trouble swallowing, mucus, and vocal dysfunction were selected for analysis in this study. Patient stratifications were obtained by applying Bayesian Mixture Models with three components to patient's dose histograms for relevant organs. The clusters with the highest total mean doses were translated into dose thresholds using rule mining. Patient stratifications were compared against AJCC staging information using multivariate likelihood ratio tests. Model performance for prediction of moderate/severe symptoms at 6 months was compared against normal tissue complication probability (NTCP) models using cross-validation.

Results: A total of 349 patients were included for long-term symptom prediction. High-risk clusters were significantly correlated with outcomes for severe late drymouth ($p < .0001$, OR = 2.94), swallow ($p = .002$, OR = 5.13), mucus ($p = .001$, OR = 3.18), and voice ($p = .009$, OR = 8.99). Simplified clusters were also correlated with late severe symptoms for drymouth ($p < .001$, OR = 2.77), swallow ($p = .01$, OR = 3.63), mucus ($p = .01$, OR = 2.37), and voice ($p < .001$, OR = 19.75). Proposed cluster stratifications show better performance than NTCP models for severe drymouth (AUC .598 vs .559, MCC .143 vs .062), swallow (AUC .631 vs .561, MCC .20 vs -.030), mucus (AUC .596 vs .492, MCC .164 vs -.041), and voice (AUC .681 vs .555, MCC .181 vs -.019). Simplified dose thresholds also show better performance than baseline models for predicting late severe ratings for all symptoms.

Conclusion: Our results show that leveraging the 3-D dose histograms from radiation therapy plan improves stratification of patients according to their risk of experiencing long-term severe radiation associated symptoms, beyond existing NTPC models. Our rule-based method can approximate our stratifications with minimal loss of accuracy and can proactively identify risk factors for radiation-associated toxicity.

Introduction

With advancements in precision radiation therapy and the emerging dominance of HPV-driven tumors [elhalawani2020tobacco] over smoking-related tumors, patient survival has improved significantly for Oropharyngeal Cancer (OPC) patients [ang2010human][fakhry2008improved]. Despite this, survivors that receive radiation therapy frequently suffer severe lasting side effects that can significantly reduce quality of life following treatment as a side effect of radiation-induced damage to organs, such as xerostomia (drymouth) or difficulty swallowing [eisbruch2004dysphagia]. Damage to vital organs such as salivary glands and swallowing

muscles from radiation is a major factor in reduced quality of life, and precisely determining the risk associated with patient treatment plans can help physicians improve patient endpoints in two ways [langendijk2008impact]. First, it allows oncologists to identify which organs to prioritize when designing individualized treatment plans. Second, when risk of organ damage is unavoidable, oncologists can prescribe preventative treatments, such as occupational therapy, to minimize side effects.

Existing approaches to radiation treatment planning often consider single-value dose thresholds for key organs. For xerostomia, existing guidelines recommends limiting the mean dose to the parotid glands to under 20Gy to the contralateral side, or 25Gy for the ipsilateral side [emami2013tolerance], although other research suggests higher dose thresholds of 35.7Gy [sanguineti2015parotid]. Single-dose thresholds are useful in their practicality for clinical researchers, but suffer from poor generalizability and fail to take into account interactions between multiple organs, or effects from different dose distributions that yield similar mean doses.

Other approaches such as Normal Tissue Complication Probability (NTCP) models attempt to account for 3-dimensional dose distributions to organs by considering the contribution for different parts of the dose-volume histogram to output a final risk probability [marks2010use]. Existing xerostomia NTCP models mainly consider mean doses to organs at risk [beetz2012NTCP]. NTCP models can outperform dose thresholds, but suffer from higher complexity that may lead to overfitting on the data, and are difficult to use for dose planning. More complex deep learning models have shown good performance in predicting patient endpoints [men2019a]. However, research has suggested that despite improvements in performance from deep learning models, they don’t outperform standard statistical approaches in practice due to their poor transparency and generalizability [marawaha2022crossing].

To address this problem, we present an unsupervised learning method for stratifying patients based on 3D dose distributions to relevant organs-at-risk, in order to identify clusters of patients that are at risk of radiation-associated long-term severe symptoms after treatment. By using clusters as proxies for risk, these clusters can serve as risk stratifications for patient symptoms that account for complex dose distributions to multiple organs at risk, while maintaining simplicity and actionability not seen in NTCP or more complicated models. In order to translate these stratifications into more actionable doses, we also propose a method of producing a set of dose thresholds to approximate the high-risk group. Focusing on predicting patient-reported drymouth, we compare our risk stratification to existing dose-based models and models using clinical factors to show that our cluster-based and simplified threshold-based stratifications can be used to improve risk predictions of self-reported symptoms.

Methods

Overview

We detail our methods in the following sections as follows: (1) diagnostic and treatment data is collected and preprocessed from a cohort of Oropharyngeal cancer patients. We then filter out relevant patients and preprocess relevant features. (2) Patient treatment plans are fed into a clustering algorithm in order to extract patient risk clusters. (3) Ruling mining is used to produce a set of dose thresholds that approximate the high risk cluster. (4) We perform multivariate correlation testing to show that the clusters are correlated with severe long-term toxicities. (5) We perform cross-validation using logistic regression to compare the performance of our clusters to normal-tissue complication probability models. An overview of our process is shown in (Figure 1). The remainder of this subsection details an overview of our methodology.

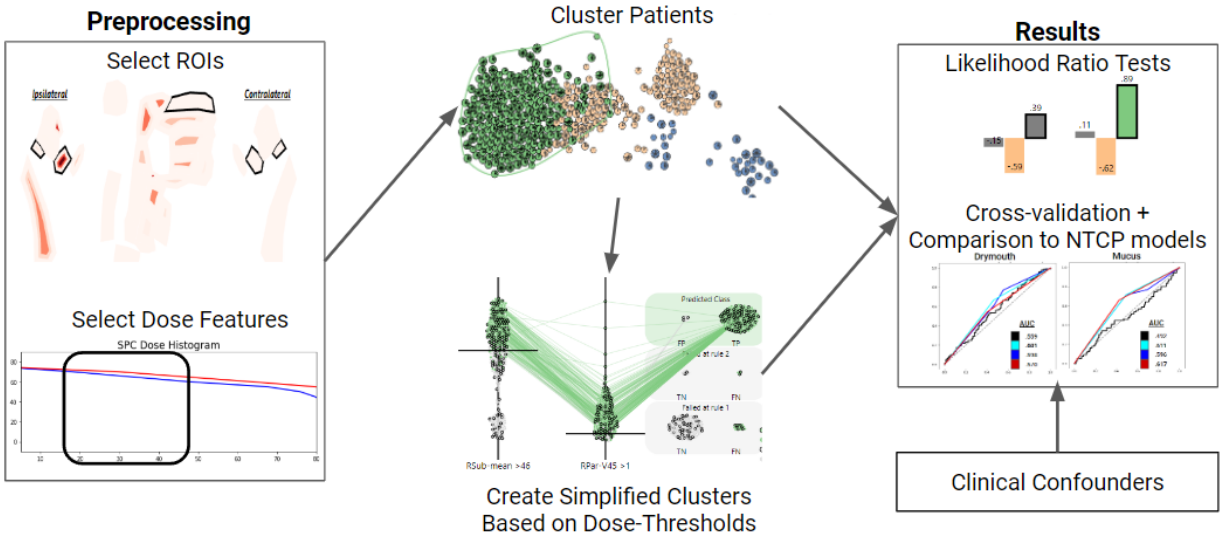


Figure 1) Overview of the methods used for each symptom of interest. First, relevant ROIS and DVH features are selected. These features are used to vectorize each patient and cluster them using a Gaussian Mixture Model. Clusters are then converted into a set of dose thresholds to approximate the high-risk group. Both clusters and simple clusters are evaluated using multivariate likelihood-ratio tests and cross-validation against NTCP models with clinical covariates to assess how predictive they are of the symptom of interest at 6 months.

First, we select a set of dosimetric features for organs relevant to each toxicity, and cluster patients based on these features into three clusters that correspond to low, medium, and high dose groups. We then identify the high-dose group, which is assumed to be the group at higher risk of long term drymouth due to damage to the relevant organs. Thus, inclusion in this high dose cluster can be used as a stratification metric for risk of tissue damage. In order to produce a more actionable and explainable stratification, we also identify a minimal set of dose thresholds to organs at risk which closely models membership in this high risk group.

For the purpose of this paper, we consider the following four self-reported symptoms: drymouth, difficulty swallowing (swallow), excessive mucus (mucus), and voice dysfunction (voice).

Drymouth has been shown to be an accurate indication of salivary function [Nittayananta2013relationship], and other symptoms are included as we theorize that they are also causally linked to damage to key tissues. Separate feature sets (choice of organ and dose thresholds) and clusters are generated for each symptom.

Our symptom data is self-reported ratings of symptoms at their worst between 0 (none) and 10 (the worst I can imagine) taken from the MD Anderson Symptom Inventory [rosenthal2007measuring]. To identify long-term outcomes, we consider the reported symptom rating during the patient's 6 month (late) followup. We consider whether the reported symptom is > 4 (severe), as well as the change in reported symptom from the patient's reported drymouth at the start of treatment is > 4 (severe change). These result in 2 binary outcomes for each symptom. Values measured during treatment were only used for imputing baseline values.

We demonstrate that our stratifications are highly correlated with self-reported late symptoms using multivariate likelihood ratio tests, and well as cross-validation in order to demonstrate that the clusters provide better predictive performance for late symptoms relative to existing clinical and normal tissue complication probability models [emami2013tolerance][beetz2012NTCP][kierkels2016multivariable], while being more explainable and accessible in real settings.

Data Collection and Preprocessing

Data were collected retrospectively from a continuously enrolled cohort of Oropharyngeal patients treated using curative-intent Radiation Therapy at the MD Anderson Cancer Center between 2010 and 2021. DVH histograms were collected from pre-treatment CECT scans taken prior to the start of treatment. Organs of interest were automatically segmented, and dose-volume histograms were extracted using commercially available software [velocity], as described in [dale2016beyond]. Additional information such as T-stage, N-stage [ICONS], HPV/p16 status, tumor location, demographic information, and initial ECOG performance score [oken1982toxicity] was collected from electronic health record data. T and N stage are existing risk stratifications based on the size and spread of primary and secondary tumors, respectively, while ECOG performance score is an indicator of the patient's level of functioning at the start of treatment.

To collect symptom information, patients were asked to fill out an MD Anderson symptom inventory (MDASI) questionnaire [rosenthal2007measuring] at weekly intervals during treatment, as well as during follow up sessions at 6 weeks, and 6 months after treatment, for a maximum of 9 time points. These questionnaires asked patients to rate the severity of 28 side effects, including drymouth, on a scale of 0-10.

Inclusion criteria for the patients were: 1) presence of OPC confirmed via biopsy; 2) patient was treated using curative-intent IMRT; 3) dose-volume histogram data available for organs at-risk in the head and neck; 4) at least 70% of the items on the MDASI questionnaire are available in the

time period from the start of treatment until 6 months after treatment; 5) symptom ratings available at 6 months; 6) patients survived long enough for a 6 month follow up appointment. The final cohort consisted of 349 patients.

Because baseline ratings were not available for 59 (16.9%) patients, we used a denoising neural network [abiri2019establishing] to impute missing values from related symptoms and the ratings at other time points for patients with a sufficient number of symptom ratings. To train the symptom imputation model, all symptom ratings from all 10 time points were used as input data. To ensure that enough symptom data was available to impute missing values, we only considered patients with a baseline drymouth rating and with at least 70% of all symptom ratings across all timesteps available. In order to train the network to learn to impute missing data, we used gaussian dropout during training, where values were randomly get to 0 with a 50% change during training, and the network was trained to reconstruct the original values using the other symptom ratings. The denoiser used two fully connected layers with a ReLU activation function followed by batch normalization. The model was trained using the Adam optimizer and mean-squared-error loss with a learning rate of .001 for 2000 with early stopping. The final model had a mean reconstruction error of 6.18%.

Clustering

In order to demonstrate that our approach can be generalized to any outcome that is associated with radiation-induced tissue damage, we apply our methodology for identifying high-risk clusters for predicting late severe ratings for four different symptoms: drymouth, swallow, mucus, and voice. Optimal cluster parameters were identified using a previously published visual analytics system developed for this project [wentzel2023dass]. For all outcomes, we use 3 clusters, and consider the cluster with the highest total mean dose to organs at risk to be the “high-dose cluster”. Organs and DVH values used for each symptom cluster are given in (Table 1). To account for bilaterality of the head, we consider the side with the higher total mean dose as the primary side, and encode the parotid and submandibular glands on that side as the “ipsilateral side”, and the organs on the other side as the “contralateral” side.

For example, when creating clusters for drymouth, we used the doses to both parotid glands, both submandibular glands, and the hard palate. We then considered the following DVH features from each organ of interest: The dose delivered to 25% of the volume (V25) through the dose delivered to 60% of the volume (V60), collected in increments of 5%, which were selected by identifying the dose features with the maximum mutual information with all late patient symptoms. Each patient was thus encoded as a vector of 40 (5 organs x 8 features) values.

The patient dose distribution was modeled using a Bayesian Gaussian Mixture Model (BGMM), an unsupervised machine learning model that learns from the distribution of the data [attias1999a]. We chose to use mixture models as we found that they proved to be effective at modeling patterns in the dose distribution due to difference in the position of the underlying tumors [tssim]. We consider the bayesian variant of the model as it is traditionally less sensitive

to the choice of parameters [blei2006variational]. After training a three cluster BGMM, the patients were clustered by assigning them to the component with the maximum likelihood.

Simplified Cluster Generation

In this paper we are mainly interested in the high-dose, high-risk patients. To define the high dose (HD) group as follows. First, we calculate the mean dose for the organs of interest used to define the clusters. We then calculate the sum of the mean doses for each cluster, and consider the cluster with the highest total mean dose to be the HD group. We verify that this HD group is also the group with the highest incidence of severe late symptom ratings.

To make the model more accessible for users without access to the original model, we also generate a "simplified" high risk group (SHD) as follows. First, we look at all dose features for all organs used in the cluster (e.g. V55 to the parotid gland). For each feature, we test different value thresholds to split the cohort into 2 groups (e.g. V55 to the parotid > 1). We then calculate the mutual information between this split, and the HD cluster, and select the 25 feature splits with the highest mutual information gain. For each ruel, we then repeat this process only on the sub-cohort that meets the criteria of the first rule, and select the 25 sets of 1-2 feature splits with the highest mutual information gain. We repeat this process iteratively until we identify a set of dose thresholds that maximize the mutual information with cluster membership. The group that exceeds all thresholds in the data is considered the “simplified” high dose (SHD) group. This results in a set of rules that can quickly approximate the original HD group, while providing thresholds that may be used for soft constraints when planning treatment plans.

Once the high-risk and simplified high-risk clusters were identified, we performed a chi2 test between clinical covariate and membership in either the original clusters and the simplified cluster. T-test statistic and significance levels were collected for the following covariates: Sex (male/female), T-stage, N-stage, HPV p16 status, primary tumor subsite, radiation treatment type, if the patient had surgery prior to treatment, age, total dose to the primary tumor, and the dose-fraction.

LRT Tests

For each endpoint we assess the predictive power of the original and simplified clusters using a likelihood ratio test (LRT). For this, we build maximum likelihood estimation models that consider clinical covariates as well as models that include both clinical covariates and either all clusters or each cluster individually. We then perform an LRT to identify if the goodness of fit of the model with clusters added has a statistically significant better fit than the baseline cluster with only clinical covariates. Additionally, we consider the linear case where we model the outcome on a 10-point scale using linear regression. We report the p-values from the likelihood ratio test, the odds ratios are taken from the model coefficients for each cluster, and the change in Akaike (AIC) and Bayesian (BIC) information criteria between each model and the clinical baseline mode. AIC and BIC are estimates of the goodness of fit of a model that includes a penalty for the number of variables considered, in order to prevent overfitting, where lower

scores indicate better fits [konishi2008information]. For BIC, reductions in score relative to the baseline model of at least 2 indicate reasonable evidence, while reductions of at least 6 indicate “strong” evidence of improvement [raftery1995bayesian].

For the purpose of testing our models, we consider the following covariates that serve as our clinical confounders: T-stage > 2 (T-stage); N-stage > 1 (Nn-stage); HPV/p16 status (hpv); primary tumor at the base of the tongue (BOT); primary tumor at the Tonsil (Tonsil); age >= 65 years at the time of diagnosis (age); ECOG performance score = 1; ECOG performance score = 2 (ECOG score); and if the patient had a mean dose of > 20 Gy to both parotid glands, or > 25 Gy to one parotid gland (Parotid Limit). These encodings were chosen as they are clinically relevant confounders that have been found to be most relevant when considering treatment type and outcomes. Sex was not included as it was found to not have any correlation with any outcome ($p > .8$) via chi-squared test, and 90% of the cohort was male. We chose to include T-stage, N-stage, and HPV status separately as our earlier work suggested that T-stage was more predictive of dysphagia than AJCC status [wentzel2020precision], which was designed to be predictive of survival, and our cohort had a combination of AJCC 8th edition and 7th edition ratings.

In order to understand how our baseline confounders compare to our clusters, we performed multivariate maximum likelihood estimation to determine the odds ratio and p-value from the likelihood ratio test between each confounder and outcome individually. Additionally, we tested the correlation between published dose thresholds to organs in the head and neck and severe late drymouth. We also looked at correlations with published dose limits [emami2013tolerance] to organs of interest. Rules for dose limits are described in (Table 2).

Cross-Validation

In order to compare our model to existing models, we compare cross-validation performance of our clusters (3-level stratification) to a baseline NTCP model based on previous literature. For the NTCP model, we use logistic regression with clinical covariates as well as the dosimetric values to organs at risk that best approximated existing clinical models based on available segmentation data [beetz2012NTCP][beetz2012NTCP][kierkels2016multivariable]. For each outcome, we re-calibrate the NTCP model on the training data during cross-validation in order to ensure the optimal performance of the NTCP model for comparison. All dosimetric values for NTCP models consider the mean dose to the organs considered. For example, the final dose values considered in the NTCP model are the mean doses to the following organs: parotid glands, submandibular glands, soft palate, upper lip, lower lip, oral cavity, and mylogeniohyoid muscle. We included the mylogeniohyoid muscle as we did not have seperate contour data for sublingual salivary glands.

When evaluating the performance of our clusters during cross-validation, we rank each cluster based on the number of patients that experience the given outcome in the training data and assign risk to patients in the test data based on the rank of their clusters. In this way, the

highest-risk cluster is given a risk score of 1, while the second highest-risk cluster is given a risk score of .5. For the simplified cluster, we always assign a risk of 1 to the high-dose cluster and 0 otherwise. For the whole dataset, this is the equivalent of using the clusters as a xerostomia risk stratification.

We report the area under the receiver-operator curve (AUC-ROC score), which is a measure of the specificity of a test as the sensitivity threshold changes [fawcett2006an]; and the Mathew's correlation coefficient (MCC) [matthews1975comparison], which is a special case of a correlation coefficient that has been shown to be useful for evaluating binary outcomes for imbalance data [chicco2020the], of our risk stratification compared to the baseline and NTCP models for all binary outcomes.

Results

Demographics

The distribution of patient symptom ratings are shown in (Table 3). We see drymouth is the most prevalent symptom, with late severe drymouth occurring in 43.8% of patients and an average rating of 4.34 at 6 months, followed by severe mucus, which only occurs in 16% of patients (mean rating 2.26). Voice had the lowest number of patients with an average rating of 1.07 and only 4% of patients reporting severe voice dysfunction and only 1.7% reporting an increase of at least 5 point from baseline at 6 months.

Demographics, and demographics within the high-dose and simplified high-dose clusters for each outcome are shown in (Table 4). The cohort was predominantly male (90%) and HPV/p16 positive (81%), with a mean age of 59 (95% CI 58-60). A majority of patients were treated with volume-modulated arc therapy or intensity modulated proton therapy (63%), while only 2 patients received 3d conformational therapy. 10% of patients underwent surgery prior to radiation therapy.

Results of chi-squared tests between demographic features and cluster membership is shown in (Figure 2). A significant correlation was found between all cluster memberships and T-stage ($p < .0001$), tumor subsite ($p < .0001$), and treatment modality ($p < .05$), while N-stage was correlated with all but simplified swallowing risk ($p < .05$). Patients in high risk clusters had higher rates of stage T4 (10% vs 17-31%) and N2C/N3 tumors (14% vs 18-23%), which correspond to patients likely to receive the most aggressive treatment. Additionally, all high risk groups had higher incidences of tumors at the base of the tongue (BOT), and lower incidence of tumors in the Tonsil. There was also a higher rate of patients that received VMAT/IMPT in the high-risk clusters (63 vs 69-87%). All standard clusters as well as simplified voice clusters were correlated with lower rates of pre-treatment surgery ($p < .05$, 10% vs 0-7%). No significant difference was found between ECOG performance score and clusters. Drymouth and Mucus clusters were not correlated with HPV status ($p > .05$), but there were fewer HPV+ patients in the swallow high-dose (81% vs 78%, $p < .01$) and simplified high dose clusters 81% vs 80%, ($p < .05$), as well as simplified voice (81% vs 76%, $p < .001$).

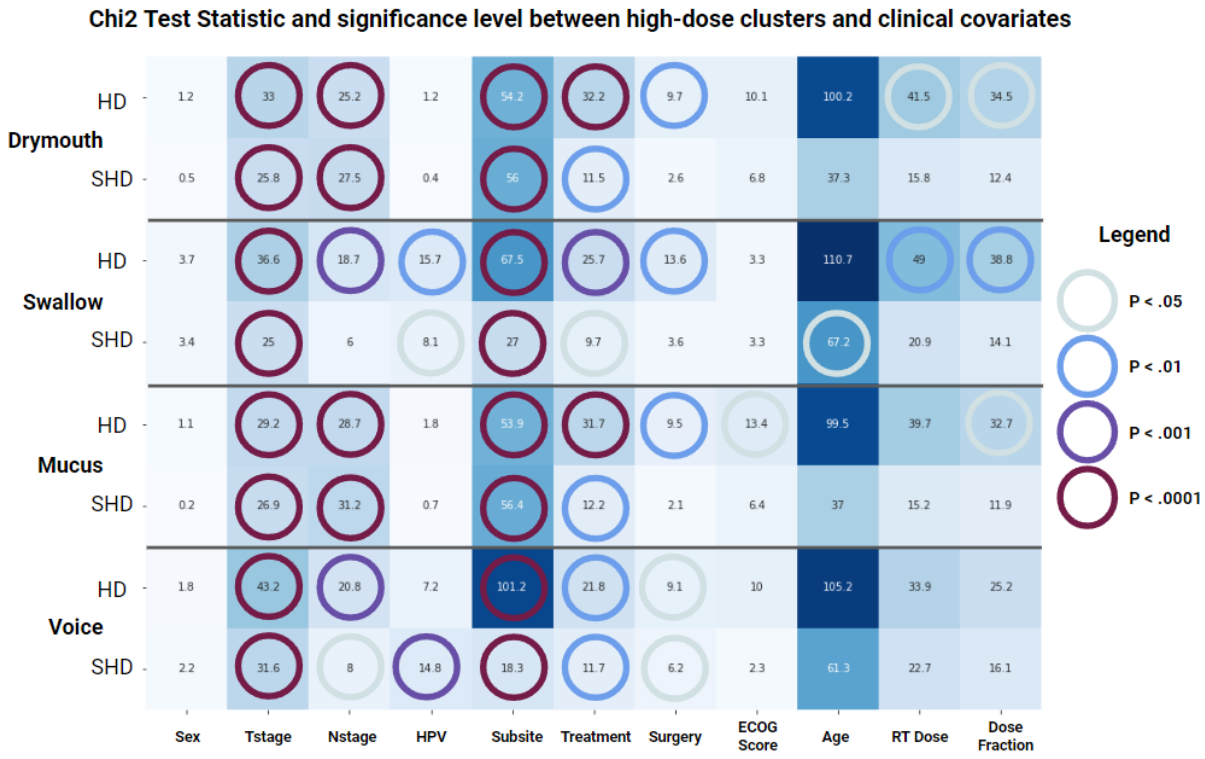


Figure 2) Results of a chi-squared test between covariates and membership in each set of clusters for each outcome. (HD) Standard clusters, (SHD) Simplified clusters. Color and annotations encode t-statistic values while colored circles represent the significance level based on the p-value.

Results for the correlation tests between baseline confounders, existing dose guidelines, and late severe symptoms are shown in (Figure 3). The factors most correlated with severe drymouth were ECOG performance score >= 2, and primary tumor at the base of the tongue (BOT). Oddly, T-stage 4 was negatively correlated with drymouth, while the less-severe T-stage 3 was positively correlated. The strongest predictors of negative outcomes are high doses to the larynx and superior pharyngeal constrictor, which are traditionally associated with swallowing complications and not drymouth. The dose limits to the parotid glands intended to predict xerostomia were negatively correlated with high drymouth, which is likely due to the fact that the majority of patients whose doses were within acceptable limits were within the low-dose cluster, which had anomalously high rates of drymouth relative to the moderate dose group (38.3% vs 92.92%, respectively).

Figure 3. Heatmap of odds-ratios from fisher's-exact test between late severe (> 4) ratings for each symptom, and confounders used in the data, (top) as well as published dose limits. Statistically significant values (p < .05) are marked with green circles. Values < 1 indicate lower than average risk while values > 1 indicate above average risk. Legend: (BOT) Subsite at Base of Tongue; (HPV+) HPV/p16 positive; (IMPT) Intensity Modulated Proton Therapy); (IMRT) Intensity Modulated Radiation Therapy; (Tonsil) Subsite at Tonsil; (VMAT) Volumetric Modulated Arc Therapy; (Concurrent) Chemotherapy concurrent with radiation therapy; (IC) Induction Chemotherapy; (ECOG) Eastern Cooperative Oncology Group Performance Score.

Cluster Analysis

The final parameters for each outcome are shown in (Table 1). Interestingly, we found similar simplified rules for predicting late severe voice dysfunction (IPC V55 > 34) and late severe swallowing issues (IPC V50 > 40). Similarly, rules for the high-risk mucus and drymouth clusters show similar rules for thresholds to the contralateral parotid glands (V45 > 61 and V50 > 48), and for the contralateral parotid gland (V45 > 0). Notably, the optimal DVH values were lowest for predicting drymouth than other symptoms with values ranging from V25-V65, compared to V20-V60 for drymouth. Clusters for swallow and voice also had higher optimal DVH values, and generally included more muscles instead of salivary glands.

Comparison of high-dose and low/moderate-dose-volume histograms of the organs used for the high-dose clusters are in (Figure 4). We can see that rules generally correspond to the ROIs that show the highest difference in mean dose between high- and low/moderate-dose groups. We see larger separations for the contralateral submandibular glands, inferior pharyngeal constrictors, and supraglottic larynx. We can also see that in the high risk group, mean dose to

the submandibular glands tends to be relatively high even at 80% penetration, while the dose to the dose to the parotid gland will drop off to low or zero values at around 45% penetration for the low/moderate dose groups. We also see relatively high levels of dose for the MPC and SPC (Figure 4-swallow column) even at 80% penetration with limited dropoff.

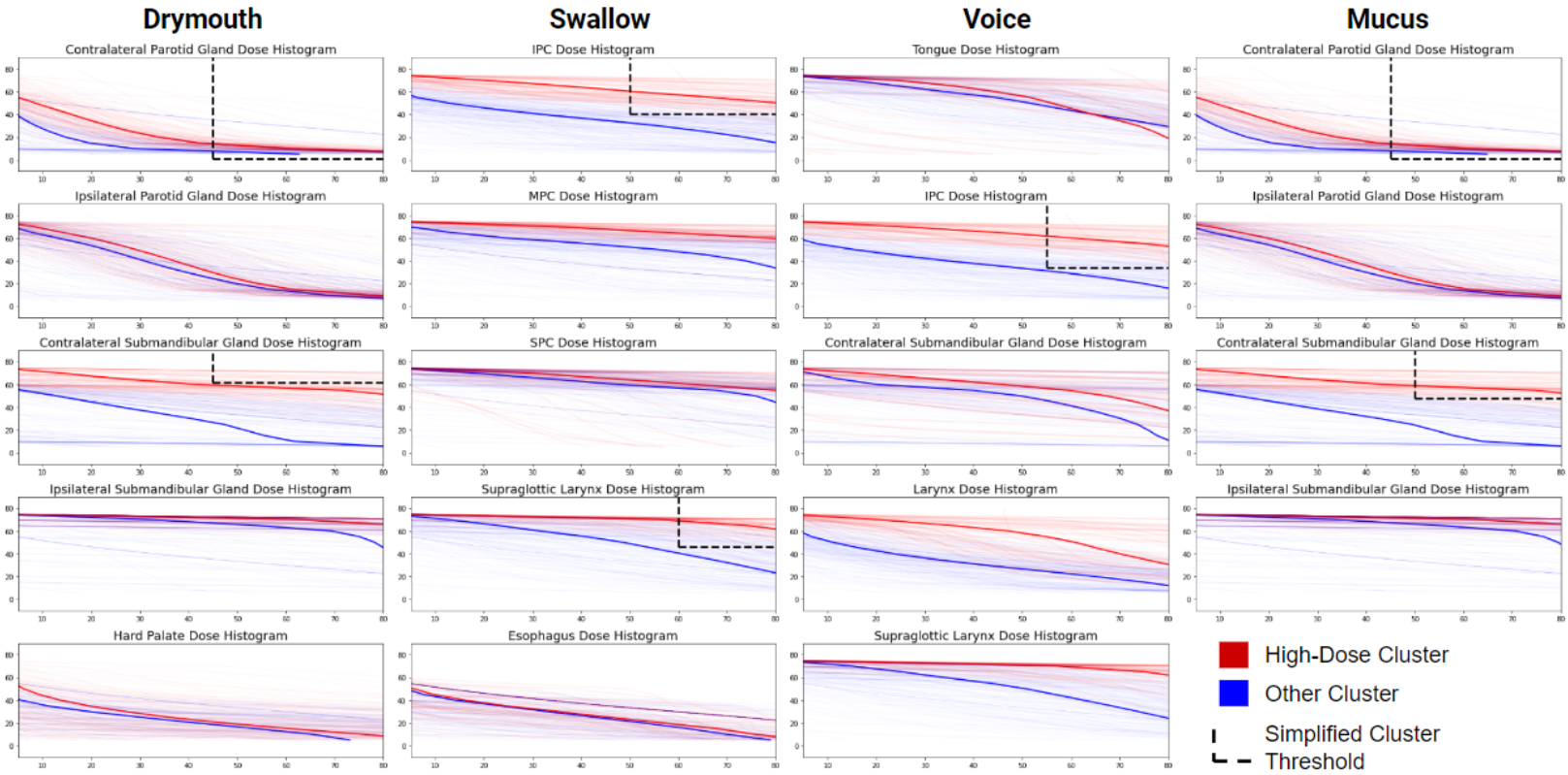


Figure 4) Comparison of Dose-volume features between patients in each high-dose cluster (red) and those in low- or moderate-dose clusters (blue). Each plot shows the dose-volume histogram for each patient. Darker lines show the median values within each group. Dashed lines show the thresholds used for producing the simplified cluster, excluding rules that use max-dose to the ROI. Patient histograms that pass through the upper-right window of all plots in their row are in the simplified high-dose cluster.

The distribution of symptoms at the start of RT treatment and at 6 months for each high-risk and low/moderate risk groups are shown in (Figure 5). Mean ratings for all groups increase between baseline and 6 months, although the difference in change is higher for the high-dose groups. All high-dose clusters show a slightly higher mean symptom rating at baseline than the low/moderate dose groups, with differences of .14, .83, .01, and .78 for drymouth, swallow, mucus, and voice, respectively. This difference increases at 6 months for all cases to 1.27, .126, .91, and 1.02 for drymouth, swallow, mucus, and voice, respectively. The larger baseline difference for swallow and voice likely correspond to the higher rates of stages T4 and N3 in these groups at the start of treatment, which we don't see in drymouth or mucus. The most

significant change is in the high-dose drymouth group, which has a mean symptom rating increase of 3.87 between baseline and 6-months.

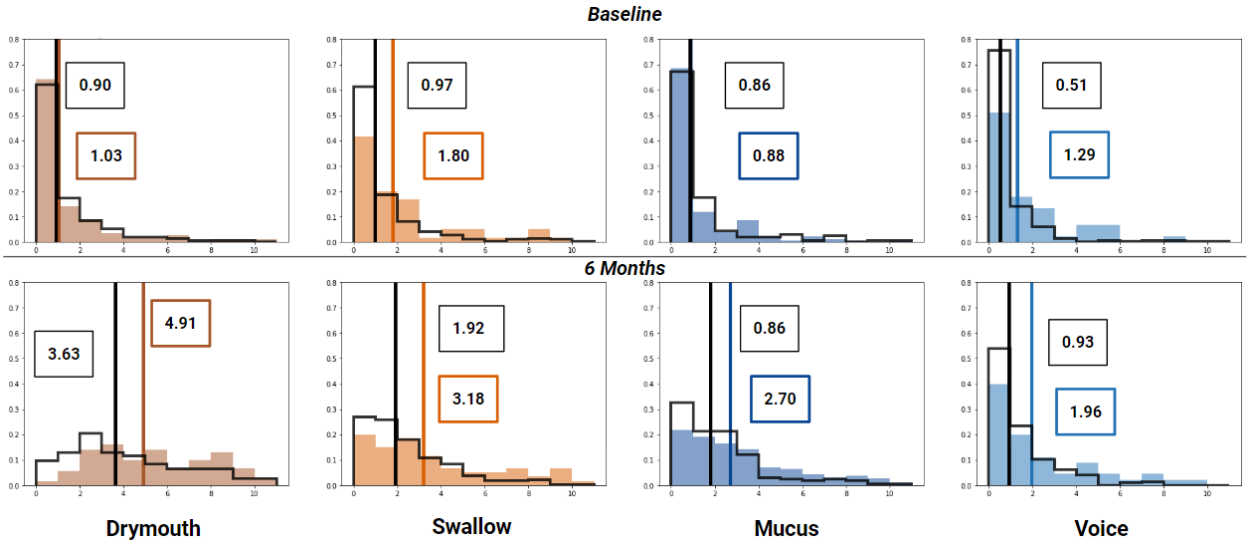


Figure 5) Histogram of symptom ratings before treatment (top), 6 week (middle) and 6 months (bottom) after treatment for each cluster (colored bars) compared to the rest of the cohort (black outline). Lines show median rating for patients within (colored) and patients not in the cluster (black). Mean values for high-dose clusters are labeled in colored boxes while the moderate/low dose clusters are labeled with black boxes.

LRT test results

Results for LRT tests on all outcomes with clinical confounders are reported in (Table 5). All outcomes show significant (< .01) correlation between 3-level cluster stratifications and severe late symptoms. When considering the change from baseline rating, we have significant correlations for the high-dose clusters with all outcomes except for “voice”, which may be because we only have 6 patients with a change in voice ratings above 4 in the dataset (1.7%) (Table 3).

For absolute outcomes (rating > 4), Drymouth high-dose (HD) and simplified high-dose (SHD) clusters had the highest significance level ($p < .0001$) with odds-ratios of 2.942 and 2.767 for severe late drymouth, respectively. Voice had the highest odds-ratios of all symptoms for severe voice dysfunction with values of 8.99 and 19.75 for the HD and SHD, respectively ($p < .01$). Swallow HD and SHD clusters had odds ratios of 5.129 ($p = .002$) and 3.625 ($p = .01$), respectively. Finally, mucus HD and SHD clusters had odds ratios of 3.18 ($p = .001$) and 2.37 ($p = .01$), respectively.

For relative outcomes (rating change from baseline > 4), we see similar or slightly lower odds ratios but lower p-values, due to the smaller number of measured outcomes, for Drymouth HD (OR = 2.38, $p = .002$), Drymouth SHD (OR = 2.447, $p < .002$), Swallow HD (OR = 4.73, $p = .014$),

Swallow SHD (OR = 3.76, p = .028), Mucus HD (OR = 3.382, p < .001), and Mucus SHD (p = 2.17, p = .032). However, there is no correlation between Voice HD (OR = .96, p = .96) or Voice SHD (OR = 2.55, p = .42) and change in voice ratings > 4.

Comparing 3-level cluster stratifications, HD cluster, and SHD clusters, HD clusters tend to perform slightly better, except in the case of predicting severe late drymouth and severe late voice, in which the SHD clusters do marginally better. Inclusion of the 3-level stratifications over the High-dose only clusters didn't have a notable difference in significance level. With the exception of change in swallow >4 from baseline, 3-level stratification tended to perform worse in terms of change in Bayesian Information Criteria, suggesting that majority of the information gain comes from the high-dose clusters.

Cross-Validation Results

We report results from performing cross-validation for several alternative patient outcomes in (Table 6). ROC curves for each outcome on severe ratings are shown in (Figure 6). In terms of ROC and MCC, cluster stratification (3 clusters) outperformed baseline NTCP models for all outcomes. Performance differences between only the high-dose clusters (HD), simplified clusters (SHD), and all clusters (3-level stratification) were mixed, with the high-dose cluster outperforming all clusters for late mucus and drymouth, but not voice or swallow.

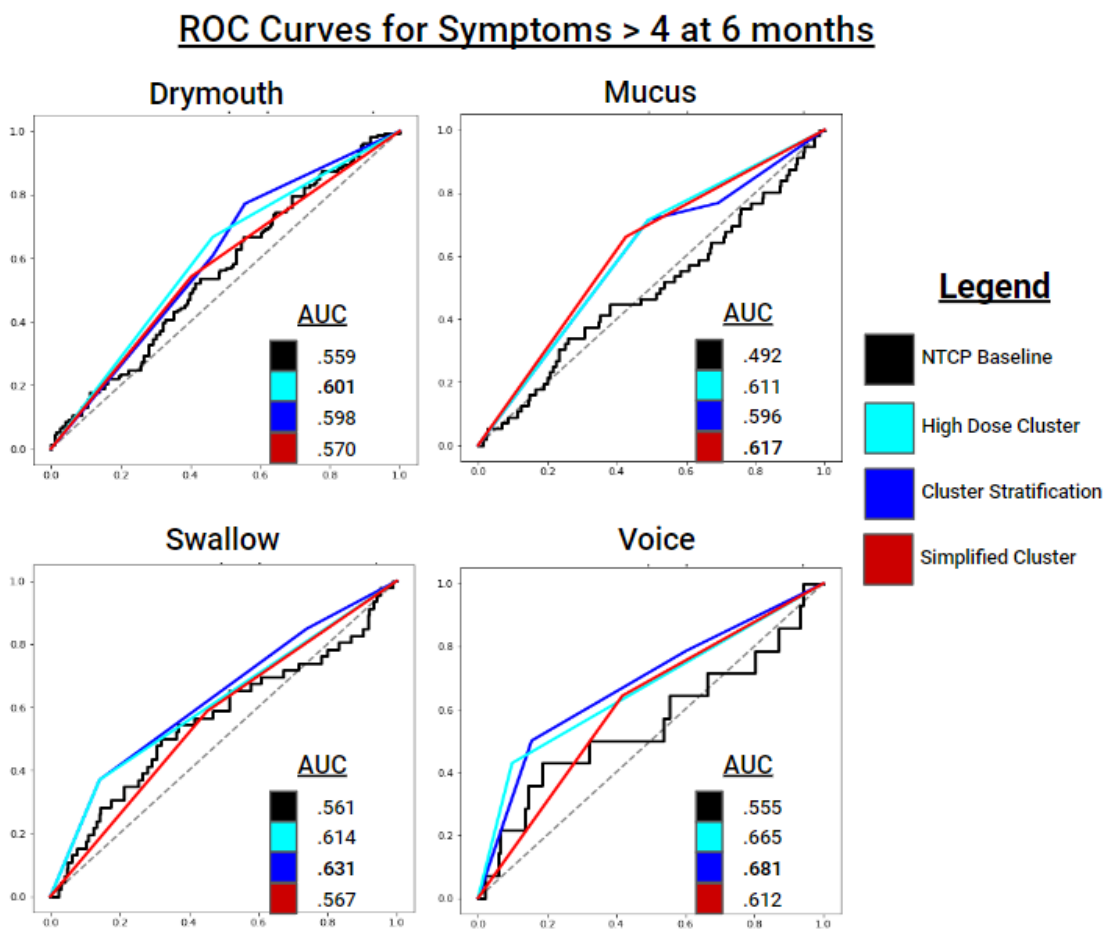


Figure 6) ROC Curves for predicting symptom ratings > 4 at 6 months for each symptom ratings. Cluster stratifications include: all clusters (blue), the high dose cluster (cyan), and the simplified high dose cluster (red). Baseline models for comparison are NTCP logistic regression models which include dosimetric variables and clinical variables.

For Drymouth outcomes, the HD cluster alone performed the best for all measures, with an AUC of .6 for severe drymouth vs .56 for the NTCP + clinical covariates model. Using the 3-level stratification achieved the same AUC score as the HD cluster, but lower MCC, due to the higher number of clusters. SHD slightly outperformed the NTCP model for absolute rating > 4 (AUC .57 vs .56), but not for change in rating > 4 (AUC .55 vs .57), although the SHD had a higher MCC for both outcomes.

For Swallow, 3-level stratifications performed the best in terms of AUC for rating > 4 (AUC = .63) and change in rating > 4 (AUC = .61). In all cases for swallow, the 3-clusters performed the best, followed by the HD, SHD, and NTCP models performed the worst.

For Mucus, the SHD performed the best in terms of AUC for both Mucus > 4 (AUC = .62) and change from baseline > 4 (AUC = .64). Voice had mixed results in terms of performance. For Voice > 4, the 3-level model performed the best (AUC = .68), followed by HD (AUC = .67), and SHD (AUC = .61), and finally the NTCP model (AUC = .56). For change from baseline, all models

performed close to chance due to the lower number of positives, with the highest performance from HD (AUC = .53).

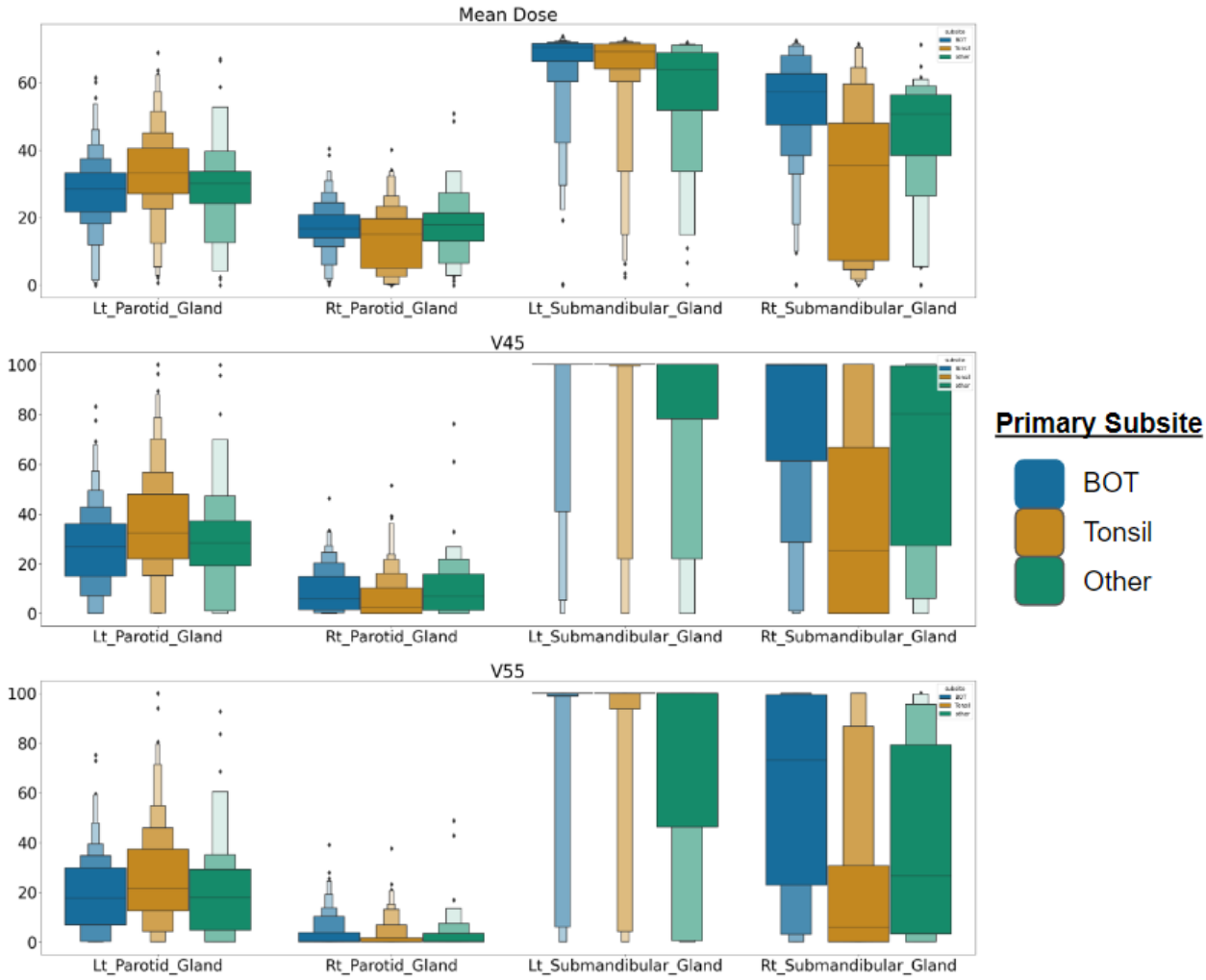


Figure 7. Dose distribution for patients with tumors at the BOT, Tonsil, and any other subsite for the parotid and submandibular salivary glands. Each row represents Mean Dose, V45, and V55, respectively. BOT subsite is associated with higher average doses to the contralateral submandibular glands, suggesting more frequent bilateral irradiation. Each rectangle in the plot represents the value range for a quantile.

Discussion

Our results demonstrate the benefits of grouping OPC RT patients based on multi-organ key 3D dose spatial distribution metrics related to patient outcomes. By identifying organs that may serve as failure points for essential functions, we were able to identify a high-dose/high-risk group of patients. Both the original high-dose cluster and the simplified version of this cluster are strongly correlated with the severity of self-reported symptoms that persist up to 6 months after treatment and improve predictive models after accounting for clinical confounders and overfitting. This methodology can serve as a valuable tool for identifying potential causes of

lasting toxicities as a result of radiation-induced damage that outperforms existing models, and can be used alongside NTCP risk prediction models. Additionally, we provide a rule mining algorithm that can simplify our rule set into a set of actionable dose thresholds that can be used without access to the original model.

Existing approaches for normal tissue toxicity probability (NTCP) calculation for risk prediction rely on summary dosimetric parameters [kuperman2009general], such as generalized equivalent uniform dose [widescott2008role], maximum, or mean dose to a region of interest. Normal Tissue Complication Probability (NTCP) models can address three-dimensional dose distributions to individual organs in order to predict outcomes. Existing models suffer from limitations imposed by challenges of dealing with correlated dose features, assumptions of linear relationships between dose and effect, and reliance on simplifying 3-dimensional dose distributions to a single unit [van2020key]. We attempt to address these issues with the use of clustering on 2-dimensional dose-volume histograms, which allows us to capture patterns in the dose distribution that encompass relationships between many correlated features in a way that does not assume linearity or uncorrelated dose features. Additionally, our simplified stratifications are transparent, which makes them more convenient to use when incorporating them into existing treatment guidelines and accounting for patient-specific information. Finally, we note that while we directly compare our model to NTCP models, these metrics can be used alongside each other, as NTCP models are designed for use in calculating specific risks when using dose planning software, while our methods are designed to provide convenient risk stratification for identifying high-risk patients and giving simple dosing guidelines.

Outside of NTCP models, the most common risk stratification for OPC patients is AJCC TNM staging. T, N, and M-staging criteria consider the size and spread of the primary tumors, secondary tumors, and distant metastasis, respectively, to predict survival [ICONS]. While TNM staging is not directly related to late toxicity risk, it can serve as a proxy for the aggressiveness of treatment and is correlated with radiation-associated dysphagia in patient outcomes [wentzel2020precision].

In our cohort, the predominant lasting toxicity was severe drymouth, which occurred in 43.8% of patients, while only 5.2% of patients reported no drymouth at 6 months, which makes it of particular interest for clinical applications. Our cluster parameters for drymouth include the submandibular glands, and the hard palate, which are all possibly causally linked to patients experiencing drymouth. When considering the simplified cluster, we found using the V45 to the contralateral submandibular gland and the V45 to the contralateral parotid gland achieved a sensitivity and specificity of .89 and .98, respectively. This suggests that treatment planning should prioritize reducing the dose delivered bilaterally to the submandibular salivary glands, as well as sparing at least 55% of the contralateral salivary gland from irradiation. These findings suggest that damage to both sets of salivary glands, rather than one, is a major factor in determining severe drymouth, as sparing a single set of glands may be able to mitigate the severity of experienced drymouth. At the same time, high-dose to the contralateral side of the

head is also correlated with larger and more extensive tumor spread, which may be a confounding factor that we would like to investigate in future work [wentzel2020precision].

When comparing our clusters for different symptoms see that the optimal parameters for predicting both drymouth and mucus include the parotid glands and submandibular glands, which indicate that mucosal dysfunction may be related to drymouth. Our parameters for swallow and voice issues consider larger sets of muscles closer to the area around the neck and base of the tongue, while mucus and drymouth focus on salivary glands in the mouth. Additionally, we see that the optimal parameters for swallow and voice consider radiation at larger levels of penetration into the volume (V30-V65), and contain smaller high-risk clusters (Table 6). This may reflect a greater tolerance in muscle tissue over salivary glands to radiation. Overall, the alternative symptoms considered were reported as severe (> 5) less frequently than drymouth, which may explain the larger p-values on LRT tests relative to drymouth, even when performed on predictive models was good for high-dose and simplified high-dose clusters.

While our models represent an improvement over existing tools, overall performance remains relatively low, with clinical baseline models performing only slightly above chance, which may reflect the difficulty in precisely identifying patients at high risk of symptoms using only EHR and dosimetric data. Notably, the previously suggested dose limits for the parotid gland to limit xerostomia are not correlated with drymouth, with most outcomes yielding a negative odds ratio, likely due to other confounders in the data. Of our confounders, we found that the strongest predictors were ECOG performance score, having a tumor at the base-of-tongue, and receiving proton therapy. The relationship between tumors at the BOT supports the theory that higher doses to the submandibular glands are related to drymouth. Preliminary analysis suggests that patients with a primary subsite at the BOT are associated with higher doses to the contralateral submandibular gland (Figure 7), with an average mean dose of 66Gy and 54Gy to the ipsilateral and contralateral submandibular glands, respectively, vs 62Gy and 34Gy for other subsites. On the other hand, BOT tumors are not associated with higher doses to the parotid glands.

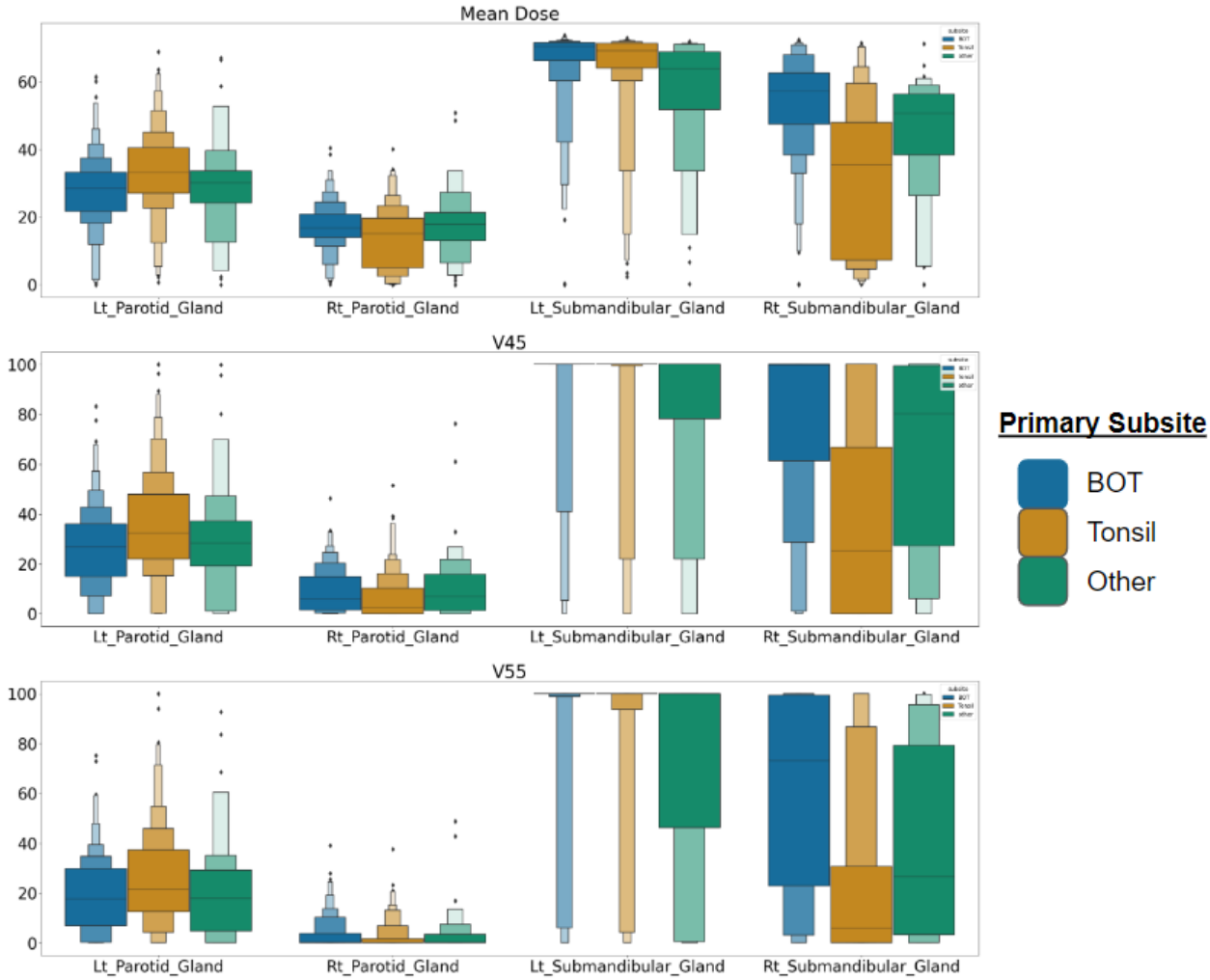


Figure 7. Dose distribution for patients with tumors at the BOT, Tonsil, and any other subsite for the parotid and submandibular salivary glands. Each row represents Mean Dose, V45, and V55, respectively. BOT subsite is associated with higher average doses to the contralateral submandibular glands, suggesting more frequent bilateral irradiation. Each rectangle in the plot represents the value range for a quantile.

Interestingly, we also found that late T-staging (T4) and N-staging (N3) was strongly predictive of severe swallow and voice dysfunction, but not mucus or drymouth. Both swallow and voice also had a higher difference in baseline symptom ratings between the high-dose and moderate-dose groups, as well as higher rates of tumors at the base-of-tongue. This suggests that there may be additional effects caused by the tumor itself in addition to radiative damage. Regarding treatment modality, we didn't find a correlation between method and outcomes (Table 3), but we did find a correlation between treatment method and cluster, with the HD clusters being a higher portion of patients that received VMAT or IMPT, especially for the swallow and voice clusters.

Our results consider both overall severity at 6 months (rating > 4), as well as severe change in rating relative to baseline ratings (change > 4). The inclusion of the severe change outcome is

designed to filter out patients with high baseline symptoms, whose toxicity may not be related to radiation-induced damage. Results show that our model still improves over the baseline in these cases, with a slight decrease in measured effect size, which is likely due to the smaller number of outcomes. We do find that all models approximate change when considering a severe change in voice outcomes, which may be because only 1.7% (6) of patients in the data report this outcome (Table 3). Additionally, we see that the high-risk clusters have a lower incidence of patients with prior surgery than the main cohort, or the low-risk group. These findings support the idea that the differences in patient outcomes are likely related to radiation-driven effects, and not confounders due to the impact of prior treatment.

With respect to our study's limitations, while our methodology attempts to identify the organs most likely to have a causal effect on outcomes, the nature of radiation dosing makes identifying causal relationships difficult due to the highly correlated nature of the doses. Spatially adjacent organs have highly correlated doses which makes disentangling their effects difficult without very large datasets. Additionally, our results are sensitive to the choice of dose parameters, and require parameter tuning in order to translate our results to other cohorts. Although we focus on HNC cancer here, our method could be generalized to other types of cancer that are linked to radiation-associated side effects, although other localized considerations may need to be taken, such as greater shape variability in the case of bladder cancer. Since the thresholds may be affected both by the specific organ and treatment methods, generalizing these results to other cohorts requires calibration of dose-volume parameters used in the clustering. Additionally, our reliance on imputation for 17% of the baseline symptoms may introduce some bias. Finally, while we attempt to use baseline features to correct for high initial symptoms, this approach may under-count patients whose initial symptoms were caused by the tumor itself as the initial symptoms not due to radiation damage would decrease after completion of treatment.

Future work could also consider modifying the dose distributions on a per-organ basis, as the submandibular glands may have lower threshold tolerances than larger muscles such as the tongue. The model may be further improved by using segmentation of specific sublingual and salivary glands in the mouth, beyond the two sets that we consider. Additionally, while we only consider dose plans prepared before treatment, future research could consider the impact of anatomical data as well as the impact of changes in dose due to temporal anatomical changes in response to treatment [marzi2012anatomical]. We also plan on incorporating additional information that may provide additional insight into patient risks, such as tumor location and bilaterality. Finally, other work may look into correlating doses to more complicated patterns of symptom progression rather than simply considering late severe symptoms, such as those being investigated in other works such as [floricel2022thalis].

In conclusion, our paper presents an unsupervised methodology for identifying patients with high doses to a set of organs, which we have shown are associated with a higher risk of lasting severe symptoms. Our model uses unsupervised Gaussian Mixture Models and approaches

based in rule mining to find stratification rules that consider failure points at multiple organs in order to identify high-risk patients.

Conflict of Interest

The authors declare no conflicts of interest.

Acknowledgments

This work is supported by NIH award NCI-R01 CA258827.

Data Availability

Data is embargoed until publication; thereafter anonymized data can be accessed at 10.6084/m9.figshare.21545436

Tables

Table 1. Table of rules used to approximate the high-dose clusters for alternative outcomes, along with the precision, recall, and info gain associated with each set of simplified clusters, in order to show how well the simplified clusters approximate the high-dose group.

Outcome	Cluster Organs	Cluster DVH Features	Thresholds	N (HD)	N (Simplified HD)	Cluster Precision	Cluster Recall	NTCP Organs
Drymouth	Both Parotid Glands, Both Submandibular Glands, Hard Palate	V25-V60	Contralateral submandibular gland V45 > 61 Contralateral parotid gland V45 > 0	219	205	0.98	0.89	Parotid glands, Submandibular glands, soft palate, upper lip, lower lip, oral cavity, mylogeniohyoid
Swallow	IPC, MPC, Supraglottic Larynx, Esophagus, Mylogeniohyoid Muscle	V30-V65	IPC V50 > 40 Supraglottic Larynx V60 > 46	60	65	0.892	0.967	IPC, SPC, Supraglottic Larynx, Parotid Gland, Cricopharyngeal Muscle
Mucus	Both Parotid Glands, Both Submandibular Glands	V25-V65	Contralateral Submandibular Gland V50 > 48 Contralateral Parotid Glannd V45 > 0	184	171	0.988	0.918	Soft Palate, Hard Palate, Oral Cavity, Mandible, Tonge, Parotid Glands

Voice	Tongue, IPC, Larynx, Supraglottic Larynx, Contralateral Submandibular Gland	V45-V65	IPC V55 > 34 Larynx Max Dose > 66	45	5.50E+01	8.00E-01	0.978	Larynx, Supraglottic Larynx, Tongue, Genioglossus Muscle, Mylogeniohyoid Muscle

Table 2. Description of the dose limits considered to different organs [emami2013tolerance], and the toxicity they are designed to avoid.

Organ	**Dose Limit (Gy)**	**Outcome**
Spinal Cord	Max dose > 50	Myelopathy
Parotid Gland	Mean dose > 25 for one OR Mean dose > 20 for both	Xerostomia
Inferior Pharyngeal Constrictor (IPC)	Mean dose > 50	Feeding Tube
Inferior Pharyngeal Constrictor (IPC 2)	Mean dose > 60	Aspiration
Medial Pharyngeal Constrictor (MPC)	Mean dose > 50	Feeding Tube
Medial Pharyngeal Constrictor (MPC 2)	Mean dose > 60	Aspiration
Superior Pharyngeal Constrictor (SPC)	Mean dose > 50	Feeding Tube
Superior Pharyngeal Constrictor (SPC 2)	Mean dose > 60	Aspiration
Mandible	Max dose > 70	Osteoradionecrosis
Larynx	V50 > 27	Edema
Brachial Plexus	Max dose > 60	Nerve Damage
Esophagus	V35 > 50 OR V50 > 40 OR V70 > 20 OR V60 > 30	Esophagitus

Table 3. Distribution of each symptom rating at 6 months, as well as the number of patients who have ratings or change in ratings above different thresholds, corresponding to “any”, “moderate”, and “severe”.

Symptom	**Avg Rating**	**Rating 5% CI**	**Rating Median**	**Rating 95% Ci**	**Threshold**	**Above Threshol**	**Above Threshol**	**Change Above**	**Change Above**

						d	d (%)	Threshold	Threshold (%)
					0	331	94.8%	295	84.5%
					2	241	69.1%	203	58.2%
Drymouth	4.34	0.4	4	9	4	153	43.8%	114	32.7%
					0	259	74.2%	199	57.0%
					2	112	32.1%	73	20.9%
Swallow	2.14	0	2	7	4	46	13.2%	29	8.3%
					0	255	73.1%	202	57.9%
					2	120	34.4%	86	24.6%
Mucus	2.26	0	2	8	4	56	16.0%	42	12.0%
					0	167	47.9%	133	38.1%
					2	51	14.6%	34	9.7%
Voice	1.07	0	0	4	4	14	4.0%	6	1.7%

Table 4. Patient demographics, treatment information of the cohort, as well as the distribution of features within the high-dose (HD) and simplified high-dose (SHD) clusters for each outcome. Continuous values show mean values and 95% confidence intervals within each group. Legend) T-stage: AJCC 8th edition T-staging; N-stage: AJCC 8th edition N-staging; HPV) Whether the patient was HPV/p16+; Subsite: site of primary tumor (BOT, Tonsil, other); BOT: Base of Tongue; VMAT: volumetric modulated arc therapy; IMPT: intensity modulated proton therapy; IMRT: intensity modulated proton therapy; ECOG Perf. Score: Eastern Cooperative Oncology Group pre-treatment performance score; RT Dose: total prescribed RT dose the the main tumor; Dose-fraction: weekly dose delivered to the main tumor.

			Drymouth		Swallow		Mucus		Voice	
	Value	Cohort	HD	SHD	HD	SHD	HD	SHD	HD	SHD
Feature	Total	349	193	175	60	65	184	171	45	55
Sex	Male	314 (90%)	171 (89%)	155 (89%)	58 (97%)	63 (97%)	163 (89%)	152 (89%)	43 (96%)	53 (96%)
T-stage	T3	48 (14%)	31 (16%)	24 (14%)	13 (22%)	15 (23%)	28 (15%)	22 (13%)	11 (24%)	13 (24%)
	T4	34 (10%)	32 (17%)	31 (18%)	16 (27%)	15 (23%)	31 (17%)	31 (18%)	14 (31%)	15 (27%)
N-stage	N2a/N2b	164 (47%)	88 (46%)	81 (46%)	31 (52%)	36 (55%)	85 (46%)	81 (47%)	21 (47%)	28 (51%)
	N2C/N3	48 (14%)	40 (21%)	40 (23%)	13 (22%)	12 (18%)	40 (22%)	40 (23%)	10 (22%)	13 (24%)
HPV	Unknown	43 (12%)	23 (12%)	20 (11%)	3 (5%)	4 (6%)	22 (12%)	19 (11%)	3 (7%)	3 (5%)
	HPV +	282 (81%)	155 (80%)	142 (81%)	47 (78%)	52 (80%)	147 (80%)	139 (81%)	35 (78%)	42 (76%)
Subsite	BOT	162 (46%)	117 (61%)	114 (65%)	45 (75%)	49 (75%)	113 (61%)	112 (65%)	32 (71%)	40 (73%)
	Tonsil	145 (42%)	52 (27%)	40 (23%)	11 (18%)	13 (20%)	48 (26%)	38 (22%)	10 (22%)	11 (20%)
Treatment	3D Conf.	2 (1%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)

	VMAT/IMPT	221 (63%)	134 (69%)	120 (69%)	50 (83%)	52 (80%)	128 (70%)	118 (69%)	39 (87%)	46 (84%)
	IMRT	74 (21%)	41 (21%)	39 (22%)	6 (10%)	8 (12%)	39 (21%)	38 (22%)	2 (4%)	5 (9%)
Prior Surgery	Yes	36 (10%)	13 (7%)	13 (7%)	1 (2%)	2 (3%)	12 (7%)	13 (8%)	0 (0%)	0 (0%)
ECOG Perf. Score	1	64 (18%)	41 (21%)	35 (20%)	13 (22%)	17 (26%)	37 (20%)	34 (20%)	12 (27%)	13 (24%)
	2	6 (2%)	3 (2%)	0 (0%)	1 (2%)	1 (2%)	1 (1%)	0 (0%)	0 (0%)	0 (0%)
	Unknown	18 (5%)	8 (4%)	8 (5%)	3 (5%)	3 (5%)	9 (5%)	8 (5%)	3 (7%)	3 (5%)
Age	Mean (95% CI)	59 (58 - 60)	59 (58 - 61)	60 (58 - 61)	63 (61 - 66)	64 (61 - 66)	59 (58 - 61)	60 (58 - 61)	64 (61 - 67)	63 (61 - 65)
RT Dose	Mean (95% CI)	53 (51 - 55)	51 (47 - 55)	51 (48 - 55)	51 (43 - 57)	52 (46 - 58)	52 (49 - 55)	51 (47 - 55)	53 (46 - 60)	54 (47 - 61)
Dose-Fraction	Mean (95% CI)	26 (25 - 28)	26 (23 - 28)	25 (23 - 26)	24 (21 - 27)	25 (22 - 28)	26 (24 - 28)	24 (23 - 26)	25 (21 - 29)	26 (22 - 28)

Table 5. Results from LRT tests for severe late drymouth and severe late change in drymouth for swallowing, mucus, and voice outcomes using their clusters. Legend: All) all clusters, results do not include odds ratio; HD) Highest dose cluster; SHD) Simplified high-dose cluster using the threshold rules; ΔAIC) Change in Aikake Information Criteria from inclusion of the cluster in a regression model; ΔBIC) Change in Bayesian Information Criteria from inclusion of the cluster in a regression model.

		Swallow			Mucus			Voice			Drymouth		
Outcome		All	HD	SHD	All	HD	SHD	All	HD	SHD	All	HD	SHD
Rating > 4	P-value	**0.001**	**0.002**	**0.010**	**0.003**	**0.001**	**0.010**	**0.004**	**0.009**	**0.000**	**0.000**	**0.000**	**0.000**
	Odds Ratio	NA	5.129	3.625	NA	3.182	2.373	NaN	8.987	19.749	NA	2.942	2.767
	ΔAIC	-10.6	-7.6	-4.7	-8.0	-9.3	-4.7	-7.2	-4.8	-11.1	-12.3	-14.2	-12.0
	ΔBIC	-2.9	-3.7	-0.9	-0.2	-5.4	-0.8	0.5	-0.9	-7.2	-4.6	-10.4	-8.2
ΔRating > 4	P-value	**0.002**	**0.014**	0.028	**0.002**	**0.001**	**0.032**	0.046	0.976	0.409	**0.009**	**0.002**	**0.002**
	Odds Ratio	NA	4.726	3.762	NA	3.382	2.171	NA	0.960	2.559	NA	2.382	2.447
	ΔAIC	-8.8	-4.1	-2.8	-8.1	-8.4	-2.6	-2.1	2.0	1.3	-5.4	-7.3	-7.4
	ΔBIC	-1.1	-0.2	1.0	-0.4	-4.5	1.3	5.6	5.9	5.2	2.3	-3.5	-3.6

Table 6) Area-under the curve score (AUC) and Mathew's correlation coefficient (MCC) scores from 5-fold cross-validation testing using cluster stratification and NTCP models for severe (> 4) self-reported symptoms at 6 months. Legend: All) Stratification with all clusters; NTCP) Fitted NTCP logistic regression model; HD) Stratification with only the high-dose cluster; SHD) Stratification with only the simplified high-dose cluster rules.

Outcome		Swallow				Mucus				Voice				Drymouth			
Metric		All	NTCP	HD	SHD	All	NTCP	HD	SHD	All	NTCP	HD	SHD	All	NTCP	HD	SHD
Rating > 4	AUC	**0.63**	0.56	0.61	0.57	0.60	0.49	0.61	**0.62**	**0.68**	0.56	0.67	0.61	0.60	0.56	**0.60**	0.57
	MCC	**0.20**	-0.03	**0.20**	0.09	0.16	-0.04	0.16	**0.17**	0.18	-0.02	**0.21**	0.09	0.14	0.06	**0.20**	0.14
ΔRating > 4	AUC	**0.61**	0.50	0.60	0.57	0.63	0.50	**0.64**	**0.64**	0.52	0.52	**0.53**	0.45	0.58	0.57	**0.58**	0.55
	MCC	**0.14**	-0.02	**0.14**	0.08	0.19	-0.03	**0.19**	0.18	0.00	-0.01	**0.02**	-0.03	0.11	0.08	**0.16**	0.09

References

[ang2010human] Ang, K. Kian, et al. "Human papillomavirus and survival of patients with oropharyngeal cancer." New England Journal of Medicine 363.1 (2010): 24-35.

[fakhry2008improved] Fakhry, Carole, et al. "Improved survival of patients with human papillomavirus–positive head and neck squamous cell carcinoma in a prospective clinical trial." Journal of the National Cancer Institute 100.4 (2008): 261-269.

[elhalawani2020tobacco] Elhalawani H, Mohamed ASR, Elgohari B, Lin TA, Sikora AG, Lai SY, Abusaif A, Phan J, Morrison WH, Gunn GB, Rosenthal DI, Garden AS, Fuller CD, Sandulache VC. Tobacco exposure as a major modifier of oncologic outcomes in human papillomavirus (HPV) associated oropharyngeal squamous cell carcinoma. BMC Cancer. 2020 Sep 23;20(1):912. doi: 10.1186/s12885-020-07427-7. PMID: 32967643; PMCID: PMC7513300.

[eisbruch2004dysphagia] Eisbruch, Avraham, et al. "Dysphagia and aspiration after chemoradiotherapy for head-and-neck cancer: which anatomic structures are affected and can they be spared by IMRT?." International Journal of Radiation Oncology* Biology* Physics 60.5 (2004): 1425-1439.

[langendijk2008impact] Langendijk, Johannes A., et al. "Impact of late treatment-related toxicity on quality of life among patients with head and neck cancer treated with radiotherapy." Journal of clinical oncology 26.22 (2008): 3770-3776.

[emami2013tolerance] Emami, Bahman, et al. "Tolerance of normal tissue to therapeutic irradiation." International Journal of Radiation Oncology* Biology* Physics 21.1 (1991): 109-122.

[sanguineti2015parotid] Sanguineti, Giuseppe, et al. "Parotid gland shrinkage during IMRT predicts the time to Xerostomia resolution." Radiation Oncology 10.1 (2015): 1-6.

[marks2010use] Marks, Lawrence B., et al. "Use of normal tissue complication probability models in the clinic." International Journal of Radiation Oncology* Biology* Physics 76.3 (2010): S10-S19.

[beetz2012NTCP] Beetz, Ivo, et al. "NTCP models for patient-rated xerostomia and sticky saliva after treatment with intensity modulated radiotherapy for head and neck cancer: the role of dosimetric and clinical factors." Radiotherapy and Oncology 105.1 (2012): 101-106.

[men2019a] Men, Kuo, et al. "A deep learning model for predicting xerostomia due to radiation therapy for head and neck squamous cell carcinoma in the RTOG 0522 clinical trial." International Journal of Radiation Oncology* Biology* Physics 105.2 (2019): 440-447.

[marawaha2022crossing] Marwaha, Jayson S., and Joseph C. Kvedar. "Crossing the chasm from model performance to clinical impact: the need to improve implementation and evaluation of AI." NPJ digital medicine 5.1 (2022): 1-2.

[wentzel2020precision] Wentzel, Andrew, et al. "Precision toxicity correlates of tumor spatial proximity to organs at risk in cancer patients receiving intensity-modulated radiotherapy." Radiotherapy and Oncology 148 (2020): 245-251.

[nittayanata2013relationship] Nittayananta, Wipawee, et al. "Relationship between xerostomia and salivary flow rates in HIV-infected individuals." Journal of investigative and clinical dentistry 4.3 (2013): 164-171.

[kuperman2009general] Kuperman, V. Y. (2008). General properties of different models used to predict normal tissue complications due to radiation. Medical physics, 35(11), 4831-4836.

[widescott2008role] Widesott, L., Strigari, L., Pressello, M. C., Benassi, M., & Landoni, V. (2008). Role of the parameters involved in the plan optimization based on the generalized equivalent uniform dose and radiobiological implications. Physics in Medicine & Biology, 53(6), 1665.

[van2020key] Van den Bosch, L., Schuit, E., van der Laan, H. P., Reitsma, J. B., Moons, K., Steenbakkers, R., Hoebers, F., Langendijk, J. A., & van der Schaaf, A. (2020). Key challenges in normal tissue complication probability model development and validation: towards a comprehensive strategy. Radiotherapy and oncology : journal of the European Society for Therapeutic Radiology and Oncology, 148, 151–156. https://doi.org/10.1016/j.radonc.2020.04.012

[ICONS] O'Sullivan, Brian, et al. "Development and validation of a staging system for HPV-related oropharyngeal cancer by the International Collaboration on Oropharyngeal cancer Network for Staging (ICON-S): a multicentre cohort study." The Lancet Oncology 17.4 (2016): 440-451.

[dale2016beyond] 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 118.2 (2016): 304-314.

[velocity] "Velocity." Varian Oncology (2018). https://www.varian.com/products/software/information-systems/velocity

[oken1982toxicity] Oken, Martin M., et al. "Toxicity and response criteria of the Eastern Cooperative Oncology Group." American journal of clinical oncology 5.6 (1982): 649-656.

[rosenthal2007measuring] Rosenthal, David I., et al. "Measuring head and neck cancer symptom burden: the development and validation of the MD Anderson symptom inventory, head and neck module." Head & Neck: Journal for the Sciences and Specialties of the Head and Neck 29.10 (2007): 923-931.

[abiri2019establishing] Abiri, Najmeh, et al. "Establishing strong imputation performance of a denoising autoencoder in a wide range of missing data problems." Neurocomputing 365 (2019): 137-146.

[wentzel2023dass] Wentzel, A., Floricel, C., Canahuate, G., Naser, M.A., Mohamed, A.S., Fuller, C., van Dijk, L. and Marai, G.E. (2023), DASS Good: Explainable Data Mining of Spatial Cohort Data. Computer Graphics Forum, 42: 283-295. https://doi.org/10.1111/cgf.14830

[attias1999a] Attias, Hagai. "A variational baysian framework for graphical models." Advances in neural information processing systems 12 (1999).

[tssim] Wentzel, Andrew, et al. "Cohort-based T-SSIM visual computing for radiation therapy prediction and exploration." IEEE transactions on visualization and computer graphics 26.1 (2019): 949-959.

[blei2006variational] Blei, David M., and Michael I. Jordan. "Variational inference for Dirichlet process mixtures." Bayesian analysis 1.1 (2006): 121-143.

[konishi2008information] Konishi, Sadanori, and Genshiro Kitagawa. "Information criteria and statistical modeling." (2008): 978-0.

[raftery1995bayesian] Raftery, Adrian E. "Bayesian model selection in social research." Sociological methodology (1995): 111-163.

[fawcett2006an] Fawcett, Tom. "An introduction to ROC analysis." Pattern recognition letters 27.8 (2006): 861-874.

[matthews1975comparison] Matthews, Brian W. "Comparison of the predicted and observed secondary structure of T4 phage lysozyme." Biochimica et Biophysica Acta (BBA)-Protein Structure 405.2 (1975): 442-451.

[chicco2020the] Chicco, Davide, and Giuseppe Jurman. "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation." BMC genomics 21.1 (2020): 1-13.

[kierkels2016multivariable] Kierkels, Roel GJ, et al. "Multivariable normal tissue complication probability model-based treatment plan optimization for grade 2–4 dysphagia and tube feeding dependence in head and neck radiotherapy." Radiotherapy and Oncology 121.3 (2016): 374-380.

[marzi2012anatomical] Marzi, S., et al. "Anatomical and dose changes of gross tumour volume and parotid glands for head and neck cancer patients during intensity-modulated radiotherapy: effect on the probability of xerostomia incidence." Clinical Oncology 24.3 (2012): e54-e62.

Appendix

A) Full Tables

Table A1. Table of patient demographics, treatment information, and outcomes within each cluster and simplified cluster, as well as the total cohort. We also include a list of patients that exceed dose limits to the parotid glands and pharyngeal muscles listed in Table 1. All p-values are reported using a Chi-squared test.

	Cluster 1		Cluster 2		Cluster 3		Simplified Cluster 3		Total	
	Total	(%)	Total	(%)	Total	(%)	Total	(%)	Total	(%)
Count	47	12.40%	113	29.82%	219	57.78%	205	54.09%	379	100.00 %
Sex										
Male	42	89.36%	102	90.27%	197	89.95%	184	89.76%	0.884	0.23%
Chi2-p		0.912		0.949		0.874		0.985		
T-stage										
t0/tx	5	10.64%	7	6.19%	23	10.50%	24	11.71%	35	9.23%
t1	13	27.66%	52	46.02%	52	23.74%	52	25.37%	117	30.87%

t2	14	29.79%	43	38.05%	72	32.88%	62	30.24%	129	34.04%
t3	8	17.02%	8	7.08%	35	15.98%	29	14.15%	51	13.46%
t4	4	8.51%	3	2.65%	34	15.53%	35	17.07%	41	10.82%
Chi2-p		0.91		**0.00000611**		**0.000169**		**0.0000458**		**0.000192**
N-stage										
n0/nx	5	10.64%	11	9.73%	16	7.31%	15	7.32%	32	8.44%
n1	18	38.30%	41	36.28%	52	23.74%	48	23.41%	111	29.29%
n2a/n2b	17	36.17%	54	47.79%	106	48.40%	96	46.83%	177	46.70%
n2c/n3	4	8.51%	7	6.19%	42	19.18%	43	20.98%	53	13.98%
Chi2-p		**0.14**		**0.0268**		**0.005**		**0.00065**		**0.0097**
HPV/p16										
Positive	36	76.60%	92	81.42%	176	80.37%	164	80.00%	304	80.21%
Unknown	7	14.89%	16	14.16%	26	11.87%	25	12.20%	49	12.93%
Chi2-p		0.794		0.447		0.587		0.681		0.739
Induction Chemo										
IC	13	27.66%	12	10.62%	39	17.81%	40	19.51%	64	16.89%
Unknown	0	0.00%	**0**	0.00%	**2**	0.91%	**2**	0.98%	2	0.53%
Chi2-p		0.092		**0.045**		0.359		0.124		**0.052**
OS										
Survival	36	76.60%	108	95.58%	203	92.69%	189	92.20%	347	91.56%
Chi2-p		0.065		0.134		0.945		0.827		0.0305
Subsite										
BOT	14	29.79%	28	24.78%	128	58.45%	129	62.93%	170	44.85%
GPS	2	4.26%	0	0.00%	2	0.91%	2	0.98%	4	1.06%
NOS	6	12.77%	5	4.42%	19	8.68%	20	9.76%	30	7.92%
Pharyngeal Wall	2	4.26%	2	1.77%	2	0.91%	1	0.49%	6	1.58%
Soft Palate	1	2.13%	0	0.00%	1	0.46%	0	0.00%	2	0.53%
Tonsil	20	42.55%	77	68.14%	63	28.77%	49	23.90%	160	42.22%
Vocal Cord	0	0.00%	1	0.88%	0	0.00%	0	0.00%	1	0.26%
Nasopharynx	**0**	0.00%	**0**	0.00%	**1**	0.46%	**1**	0.49%		0.00%

Chi2-p		**0.0339**		**0.0000000436**		**0.0000000411**		**0**		**0.0000000077**
Age	59	15	59	9	59	11	59	11	59	11
p-value		0.836		0.356		0.614		0.829		0.699
Followup Days	631	429	835	431	890	426	894	420	845	433
p-value		0.51		0.55		0.55		0.53		0.551
Sequence of Treatment										
Induction Chemo	6	12.77%	25	22.12%	25	11.42%	24	11.71%	56	14.78%
p-value		0.92		0.79		0.73		0.9		0.88
Radiation	6	12.77%	14	12.39%	24	10.96%	23	11.22%	44	11.61%
p-value		0.98		0.89		0.76		0.92		0.9
Concurrent Chemo	6	12.77%	14	12.39%	24	10.96%	23	11.22%	44	11.61%
p-value		0.98		0.89		0.76		0.92		0.9
Surgery	7	14.89%	15	13.27%	28	12.79%	27	13.17%	50	13.19%
p-value		0.89		0.89		90		0.89		0.93
Previously Treated	2	4.26%	3	2.65%	15	6.85%	15	7.32%	20	5.28%
p-value		0.99		0.22		0.17		0.09		0.25
Treatment Type										
IMPT	11	23.40%	18	15.93%	20	9.13%	19	9.27%	49	12.93%
IMRT	6	12.77%	23	20.35%	48	21.92%	43	20.98%	77	20.32%
VMAT	8	17.02%	51	45.13%	128	58.45%	119	58.05%	187	49.34%
3D Conformal	0	0.00%	2	1.77%	0	0.00%	0	0.00%	2	0.53%
p-value		**0.001**		**0.145**		**0.002**		**0.009**		**0.0006**
Max Post-treatment Symptom										
Drymouth >= 3										
Count	41	87.23%	72	63.72%	190	86.76%	178	86.83%	303	79.95%
p-value		0.25		**0.00000056**		**0.0018**		**0.00046**		**0.0000018**

Drymouth >= 5										
Count	28	59.57%	43	38.05%	142	64.84%	134	65.37%	213	56.20%
p-value	0.733		**0.000006**		**0.0001**		**0.0001**		**0.000017**	
Drymouth >= 7										
Count	16	34.04%	22	19.47%	94	42.92%	88	42.93%	132	34.83%
p-value	0.97		**0.000071**		**0.00017**		**0.00049**		**0.0001**	
Dose Limit Violations										
Parotid										
Count	18	38.30%	105	92.92%	210	95.89%	195	95.12%	333	87.86%
p-value	**0**		**0.073**		**0.000000053**		**0.0000056**		**0**	
IPC										
Count	8	17.02%	12	10.62%	51	23.29%	50	24.39%	71	18.73%
p-value	0.9		**0.01**		**0.01**		**0.003**		**0.019**	
MPC										
Count	8	17.02%	42	37.17%	175	79.91%	167	81.46%	225	59.37%
p-value	**0.00000000074**		**0.000000019**		**0**		**0**		**0**	
SPC										
Count	0	0.00%	66	58.41%	216	98.63%	204	99.51%	282	74.41%
p-value	**7.90E-35**		**0.0000061**		**5.50E-36**		**2.20E-33**		**2.40E-48**	

Table A2. Results for LRT tests from alternative symptoms. Full table includes correlations for low and mid-dose clusters in addition to the high-risk clusters

Correlations With Severe Outcomes At 6 Months						
Symptom	Change From Baseline	Clusters	P-value	Odds Ratio	AIC Change	BIC Change
Swallow	FALSE	All	**0.001**	NA	-10.586	-2.876
		Low Dose	**0.004**	0.332	-6.106	-2.251
		Moderate Dose	**0.001**	0.369	-9.991	-6.136
		High Dose	**0.002**	5.129	-7.573	-3.718
		Simplified HD	**0.010**	3.625	-4.723	-0.868
	TRUE	All	**0.002**	NA	-8.798	-1.087
		Low Dose	**0.014**	0.328	-4.071	-0.216

		Moderate Dose	**0.001**	0.307	-9.941	-6.086
		High Dose	**0.014**	4.726	-4.097	-0.242
		Simplified HD	0.028	3.762	-2.825	1.030
Mucus	FALSE	All	**0.003**	NA	-7.958	-0.248
		Low Dose	**0.019**	0.291	-3.468	0.387
		Moderate Dose	**0.002**	0.351	-7.679	-3.824
		High Dose	**0.001**	3.182	-9.265	-5.410
		Simplified HD	**0.010**	2.373	-4.702	-0.847
	TRUE	All	**0.002**	NA	-8.099	-0.389
		Low Dose	**0.002**	0.093	-8.078	-4.223
		Moderate Dose	**0.001**	0.282	-9.020	-5.165
		High Dose	**0.001**	3.382	-8.361	-4.506
		Simplified HD	**0.032**	2.171	-2.594	1.261
Voice	FALSE	All	**0.004**	NaN	-7.221	0.489
		Low Dose	**0.034**	0.298	-2.484	1.371
		Moderate Dose	**0.002**	0.161	-8.078	-4.223
		High Dose	**0.009**	8.987	-4.798	-0.943
		Simplified HD	**0.000**	19.749	-11.059	-7.204
	TRUE	All	0.046	NA	-2.144	5.566
		Low Dose	0.062	0.266	-1.489	2.366
		Moderate Dose	**0.032**	0.210	-2.602	1.253
		High Dose	0.976	0.960	1.999	5.854
		Simplified HD	0.409	2.559	1.319	5.174
Drymouth	Fasle	All	**0.000**	NA	-12.269	-4.558
		Low Dose	0.966	1.017	1.998	5.853
		Moderate Dose	**0.001**	0.406	-9.508	-5.653
		High Dose	**0.000**	2.942	-14.233	-10.378
		Simplified HD	**0.000**	2.767	-12.030	-8.175
	TRUE	All	**0.009**	NA	-5.435	2.275
		Low Dose	0.237	0.610	0.599	4.454
		Moderate Dose	**0.025**	0.531	-3.045	0.810
		High Dose	**0.002**	2.382	-7.308	-3.453

		Simplified HD	**0.002**	2.447	-7.446	-3.591

Table A3. Distribution of each symptom rating at 6 months, as well as the number of patients who have ratings or change in ratings at or above different thresholds, corresponding to “any”, “moderate”, and “severe”.

Symptom	Avg Rating	Rating 5% CI	Rating Median	Rating 95% Ci	Threshold	Above Threshold	Above Threshold (%)	Change Above Threshold	Change Above Threshold (%)
					1	331	94.8%	295	84.5%
					3	241	69.1%	203	58.2%
Drymouth	4.34	0.4	4	9	5	153	43.8%	114	32.7%
					1	177	50.7%	137	39.3%
					3	52	14.9%	34	9.7%
Choke	1.11	0	1	5	5	19	5.4%	14	4.0%
					1	255	73.1%	202	57.9%
					3	120	34.4%	86	24.6%
Mucus	2.26	0	2	8	5	56	16.0%	42	12.0%
					1	167	47.9%	133	38.1%
					3	51	14.6%	34	9.7%
Voice	1.07	0	0	4	5	14	4.0%	6	1.7%

Table A4. Correlation between various boolean confounders and severe late or severe change in late symptoms for drymouth using Fisher's Exact test.

Confounder	outcome	t-statistic	p-value
Subsite BOT	change >=5	1.69697	0.022525
Subsite BOT	absolute >=5	1.453435	0.104584
HPV+	change >=5	1.077853	0.8851
HPV+	absolute >=5	0.885484	0.682651
IMPT	change >=5	1.295655	0.496065
IMPT	absolute >=5	1.140152	0.748254
IMRT	change >=5	1.241379	0.485331
IMRT	absolute >=5	1.114286	0.694039
N-stage 2	change >=5	1.04138	0.909165
N-stage 2	absolute >=5	0.878514	0.589695

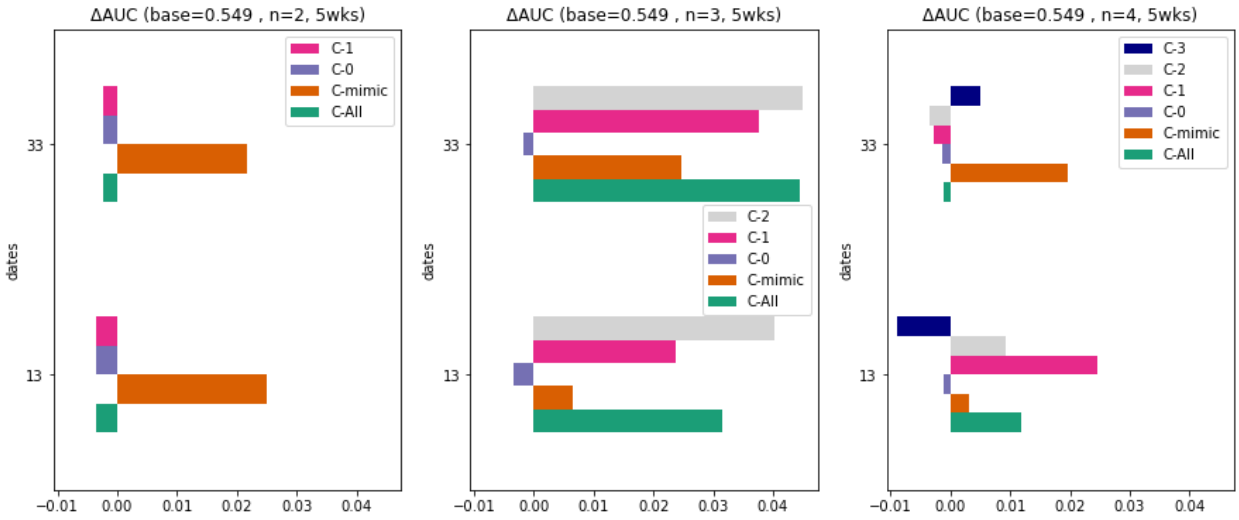
N-stage 3	change >=5	0.650327	0.24953
N-stage 3	absolute >=5	0.737778	0.353052
T-stage 3	change >=5	1.572948	0.184565
T-stage 3	absolute >=5	1.210154	0.638805
T-stage 4	change >=5	0.72	0.563916
T-stage 4	absolute >=5	1.316176	0.471117
Subsite Tonsil	change >=5	0.503452	0.005331
Subsite Tonsil	absolute >=5	0.570212	0.012053
VMAT	change >=5	0.925287	0.819491
VMAT	absolute >=5	0.947532	0.829667
age\n65	change >=5	1.236219	0.444943
age\n65	absolute >=5	1.027946	0.904885
concurrent	change >=5	0.760435	0.279862
concurrent	absolute >=5	0.683824	0.111774
ic	change >=5	0.880782	0.757154
ic	absolute >=5	1.132808	0.769032
ECOG perf. score = 1	change >=5	0.923598	0.883048
ECOG perf. score = 1	absolute >=5	0.995556	1
ECOG perf. score = 2	change >=5	2.09009	0.396696
ECOG perf. score = 2	absolute >=5	6.587838	0.090371

B) Effect of Number of Clusters on Performance for Drymouth

To determine if the results are sensitive to the number of clusters, we tested the effect of varying the number of clusters. For each test, we re-ran the clustering using a different number of clusters. Rule-thresholds were automatically identified each time. We then built logistic regression models using clinical covariates along with labels for each individual cluster, and tested the effect that the cluster label had on 10-fold cross-validation performance for predicting intermediate and late drymouth using ROC-AUC scores. AUC performance change for different number of clusters are shown in (Figure B1). We see that 2 and 4 clusters show

drastically reduced cluster performance, although the simplified high risk cluster performed well for 2 clusters.

Figure B1. Change in AUC Performance for predicting severe drymouth (> 4) for different cluster sizes at 6 weeks and 6 months post-treatment.



C) Correlations between DVH levels and Drymouth for Organs Of Interest

To better understand how different dose levels affect late drymouth, we performed experiments looking at the correlation between different DVH values and drymouth for each organ of interest. We consider 3 different metric: mutual information, multivariate regression coefficients, and f-statistics. For mutual information and multivariate regression, out outputs consider all other variable simultaneously, as well as HPV status, and tumor subsite (bottom of tongue, tonsil, or other) as confounders.

We consider severe drymouth, for both raw values and change from baseline. We report the odds ratios, calculated as the natural exponent of the model coefficients. Results for severe late (6 months) drymouth are shown in (Figure C1), while results for severe intermediate (6 weeks) are shown in figure (Figure C2).

Figure C1. Heat Maps showing different correlation measures between the dose-histogram values for each organ in the cluster, and different drymouth thresholds at 6 months after treatment cessation. (Left) Mutual information gain, including HPV and subsite as confounders. (Center) Odds ratios derived from multivariate logistic regression models which include HPV and subsite as confounders. (Right) F-statistic taken from a univariate anova test between each

value and the output, which does not consider confounders.

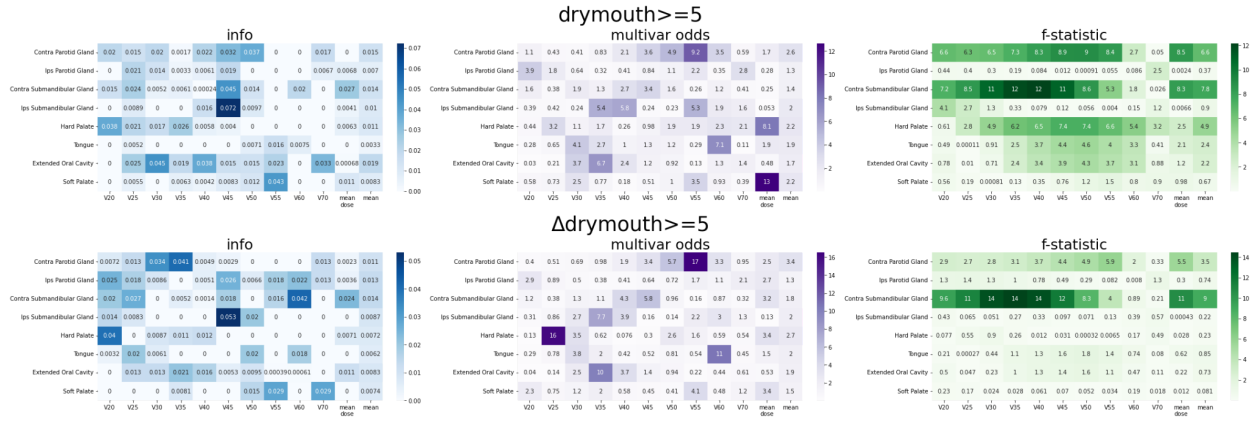
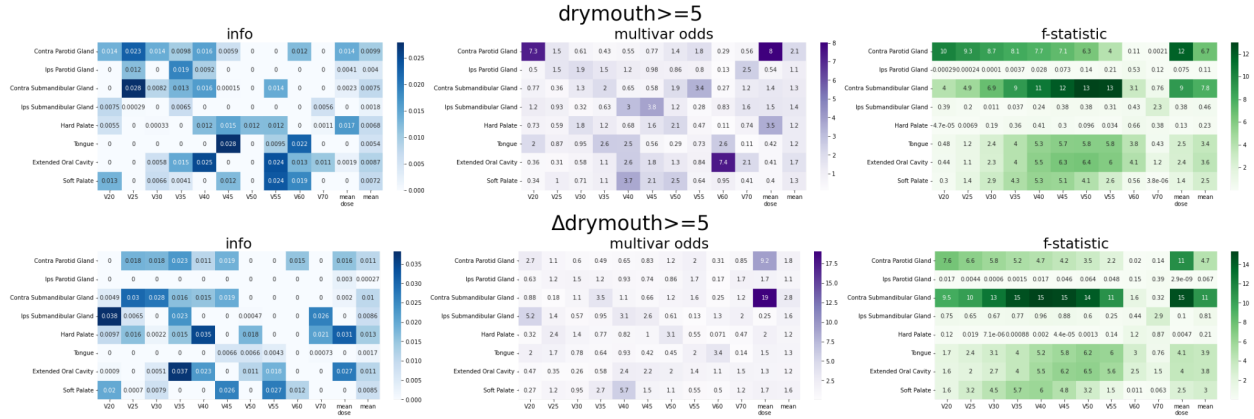


Figure C2. Heat Maps showing different correlation measures between the dose-histogram values for each organ in the cluster, and different drymouth thresholds at 6 Weeks after treatment cessation.



D) Cluster Effect Estimation with Propensity Matching

In order to attempt to estimate the causal relationship between each cluster assignment and patient outcomes, we applied doubly-robust average-treatment-effect (ATE) estimation [funk2011doubly], which is a measure of the change in likelihood of a patient experiencing and outcome if they recieve a treatment, while correcting for confounders that affect both the likelihood of experincing the treatment, and the likelihood of experiencing an outcome. For our model, we consider membership in a dose-cluster as a treatment. ATE is written generally as:

$$ATE = E(Y|X=1) - E(Y|X=0)$$

Where E(Y|X=x) is the likelihood of an outcome for a patient if the patient recives a treatment x.

ATE estimation in retrospective data is usually estimated using propensity matching, which assigns a “propensity score”, or the likelihood of the patient being in the cluster given the clinical

confounders. Doubly-robust estimation is a form of matching that uses a model for predicting both the likelihood of treatment, and likelihood of experiencing an outcome given the confounders, which allows the estimation to correct for poorly-defined propensity models, which we found gave more consistent and conservative results than other methods such as inverse-probability of treatment weighing.

To determine confounders, we tested clinical confounders and dose limits in <Table Appendix Dose Limits>, and tested which confounders were potentially correlated with both dose-clusters, and moderate drymouth. To allow all possible confounders, we tested correlations between each potential confounder and both treatment and drymouth using a chi2 test, and included all confounders with a p-value of at most .25 for both values. The final confounders are:

- T-stage 3 OR T-stage 4
- N-stage 3 OR N-stage 4
- HPV p16 status
- Primary tumor at base-of-tongue
- Primary tumor at Tonsil
- SPC mean dose > 50Gy
- MPC mean dose > 50Gy
- Esophagus V35 > 50 OR V50 > 40 OR V70 > 20 OR V60 > 30
- Larynx V50 > 27

We ran 60 iterations for each threshold. For each iteration, the data was resampled with replacement. To ensure that the clustering method is robust, clustering and simplified cluster estimation was performed in an unsupervised manner on the resampled dataset for each run. For both propensity estimation and outcome prediction, we used logistic regression with balanced class weights and elasticnet regularization. Models were calibrated using Platt's method [platt1999probabilistic] to ensure that model outputs were true probabilities. Random forests and calibration was implemented using scikit-learn [sklearn].

We report the median and 95% confidence intervals for each outcome and each cluster across all 60 runs in table (Table D1). High-dose (cluster 3) and simplified high-dose clusters have a positive treatment effect for severe drymouth (> 4) in the 95% confidence. Simplified clusters had a lower median affect size than high-dose clusters, with a non-signficant effect in predicting severe change in drymouth (> 4). We found that the low-dose cluster also had a significant positive treatment effect, which indicates that the relatively high rate of severe symptoms are not fully captured by the confounders in the data, and thus there are likely unmeasured confounders that affect patient outcomes that aren't captured in the data.

To assess quality of the effect estimated, we report the difference in propensity scores for treated and untreated groups for each cluster in (Table D2). Higher differences in average

propensity score indicate strong differences in results that can't be fully corrected with propensity methods.

Table D1. Average Treatment effect estimation using doubly-robust estimation and inverse-probability of treatment weighing. 95% Confidence intervals are estimated from bootstrap using 60 samples. All clusters are statically significant within a 95% confidence for all outcome.

Time Period	Threshold	Cluster	5% CI	Median	95% CI
6 Months	Drymouth > 4	1	0.060	0.094	0.128
		2	-0.150	-0.139	-0.059
		3	0.078	0.167	0.177
		3 (Simplified)	0.079	0.084	0.090
	Drymouth Change > 4	1	0.055	0.074	0.109
		2	-0.121	-0.068	-0.045
		3	0.042	0.101	0.123
		3 (Simplified)	-0.004	0.002	0.012
6 Weeks	Drymouth > 4	1	-0.026	-0.009	0.017
		2	-0.157	-0.091	-0.009
		3	0.015	0.133	0.155
		3 (Simplified)	0.101	0.107	0.113
	Drymouth Change > 4	1	-2.208	-2.163	-0.446
		2	-1.706	-0.974	0.098
		3	-0.550	0.819	1.429
		3 (Simplified)	0.444	0.492	0.544

Table D2. Propensity score distribution between treated group and untreated group for each cluster. Large differences in score indicate stronger separation in the dataset.

	Mean		25% CI		Median		75% CI	
Treatment (Cluster)	**In Cluster**	**Not In Cluster**	**In Cluster**	**Not In Cluster**	**In Cluster**	**Not In Cluster**	**In Cluster**	**Not In Cluster**
1	0.469924	0.068491	0.365035	0.010387	0.483098	0.014755	0.560135	0.025089
2	0.49016	0.276622	0.293881	0.169618	0.576026	0.277927	0.669384	0.31556
3 (High Dose)	**0.738513**	0.317638	0.686174	0.049667	0.743916	0.256948	0.868469	0.501791
3 (Simplified)	0.49016	0.276622	0.293881	0.169618	0.576026	0.277927	0.669384	0.31556

Table D3. Results of a chi2 test between covariates and membership in each high-dose and simplified high-dose cluster. Values report the t-statistic and p-value.

	Drymouth		Swallow		Mucus		Voice	
	HD	SHD	HD	SHD	HD	SHD	HD	SHD
Gender	1 (p > .05)	0 (p > .05)	4 (p > .05)	3 (p > .05)	1 (p > .05)	0 (p > .05)	2 (p > .05)	2 (p > .05)
T-stage	33 (p < .0001)	26 (p < .0001)	37 (p < .0001)	25 (p < .0001)	29 (p < .0001)	27 (p < .0001)	43 (p < .0001)	32 (p < .0001)
N-stage	25 (p < .0001)	27 (p < .0001)	19 (p = 0.001)	6 (p = 0.05)	29 (p < .0001)	31 (p < .0001)	21 (p < .0001)	8 (p = 0.018)
HPV status	1 (p > .05)	0 (p > .05)	16 (p = 0.003)	8 (p = 0.017)	2 (p > .05)	1 (p > .05)	7 (p > .05)	15 (p = 0.001)
Subsite	54 (p < .0001)	56 (p < .0001)	68 (p < .0001)	27 (p < .0001)	54 (p < .0001)	56 (p < .0001)	101 (p < .0001)	18 (p < .0001)
Treatment	32 (p < .0001)	12 (p = 0.009)	26 (p < .0001)	10 (p = 0.021)	32 (p < .0001)	12 (p = 0.007)	22 (p = 0.001)	12 (p = 0.009)
Prior Surgery	10 (p = 0.008)	3 (p > .05)	14 (p = 0.001)	4 (p > .05)	10 (p = 0.009)	2 (p > .05)	9 (p = 0.011)	6 (p = 0.012)
ECOG Score	10 (p > .05)	7 (p > .05)	3 (p > .05)	3 (p > .05)	13 (p = 0.037)	6 (p > .05)	10 (p > .05)	2 (p > .05)
Age	100 (p > .05)	37 (p > .05)	111 (p > .05)	67 (p = 0.018)	99 (p > .05)	37 (p > .05)	105 (p > .05)	61 (p > .05)
RT-Dose	41 (p = 0.049)	16 (p > .05)	49 (p = 0.008)	21 (p > .05)	40 (p > .05)	15 (p > .05)	34 (p > .05)	23 (p > .05)

Figure 5. ROC curves for predicting drymouth ratings > 4 at 6 months (left) and change from baseline drymouth ratings > 4 at 6 months for the cluster stratifications and baseline models. Cluster stratifications include: all clusters (blue), the high dose cluster (cyan), and the simplified high dose cluster (red). Baseline models for comparison are a logistic regression with only clinical covariates (baseline-gray), and an NTCP logistic regression model which includes dosimetric variables to xerostomia-related organs. All-values are taken from probabilities

predicted using 10-fold cross-validation.

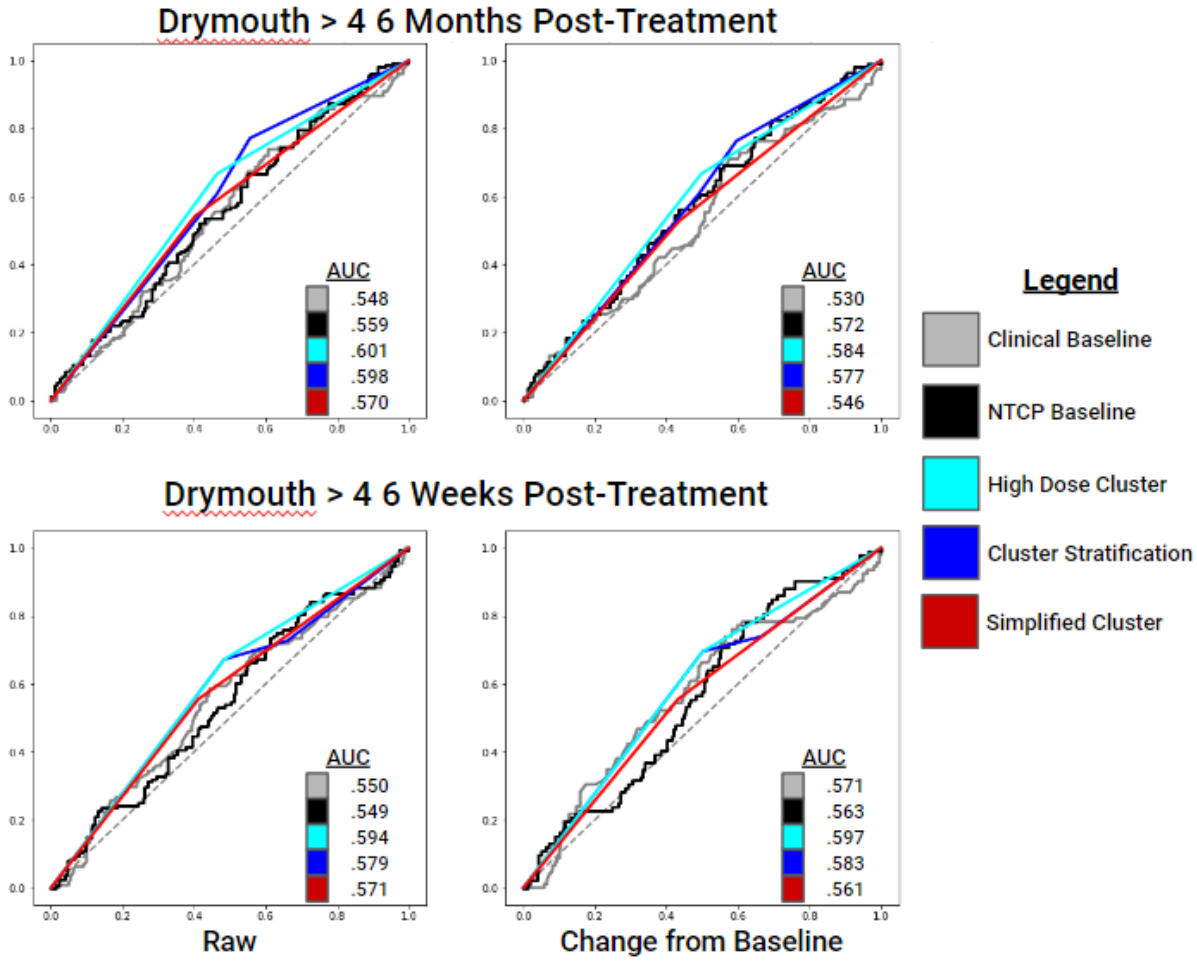


Table 5.

Outcome	Time	Group	Mean	95% CI	Difference
Drymouth	Baseline	**High Dose**	1.04	0 - 6	0.14
		Low/Moderate	0.90	0 - 4.25	
	6M	**High Dose**	4.91	1 - 9	1.27
		Low/Moderate	3.63	0 - 8.25	
Swallow	Baseline	**High Dose**	1.80	0 - 8	0.83
		Low/Moderate	0.97	0 - 4.6	
	6M	**High Dose**	3.18	0 - 9	1.26
		Low/Moderate	1.92	0 - 6	
Mucus	Baseline	**High Dose**	0.88	0 - 4	0.01

		Low/Moderate	0.86	0 - 5	
		High Dose	2.70	0 - 8	
	6M	**Low/Moderate**	1.78	0 - 6.8	0.91
		High Dose	1.29	0 - 5	
	Baseline	**Low/Moderate**	0.51	0 - 2	0.78
		High Dose	1.96	0 - 7	
Voice	6M	**Low/Moderate**	0.93	0 - 4	1.02

Appendix References

[funk2011doubly] Funk, Michele Jonsson, et al. "Doubly robust estimation of causal effects." American journal of epidemiology 173.7 (2011): 761-767.

[platt1999probabilistic] Platt, John. "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods." Advances in large margin classifiers 10.3 (1999): 61-74.

[sklearn] Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

[floricel2022thalis] Floricel C, Nipu N, Biggs M, Wentzel A, Canahuate G, Van Dijk L, Mohamed A, Fuller CD, Marai GE. THALIS: Human-Machine Analysis of Longitudinal Symptoms in Cancer Therapy. IEEE Trans Vis Comput Graph. 2022 Jan;28(1):151-161. doi: 10.1109/TVCG.2021.3114810. Epub 2021 Dec 24. PMID: 34591766; PMCID: PMC8785360.