A. Workflow and Background

Fig A1. Physician Workflow. The main items of interest for this project are to help establish a stratification of patients' risk of certain side-effects from their radiation plan (Side-Effect red box), which can then be used to identify which patients require additional preventative treatment, as well as help identify dose thresholds (Dose thresholds red box) that can be input as soft-constraints during treatment planning.

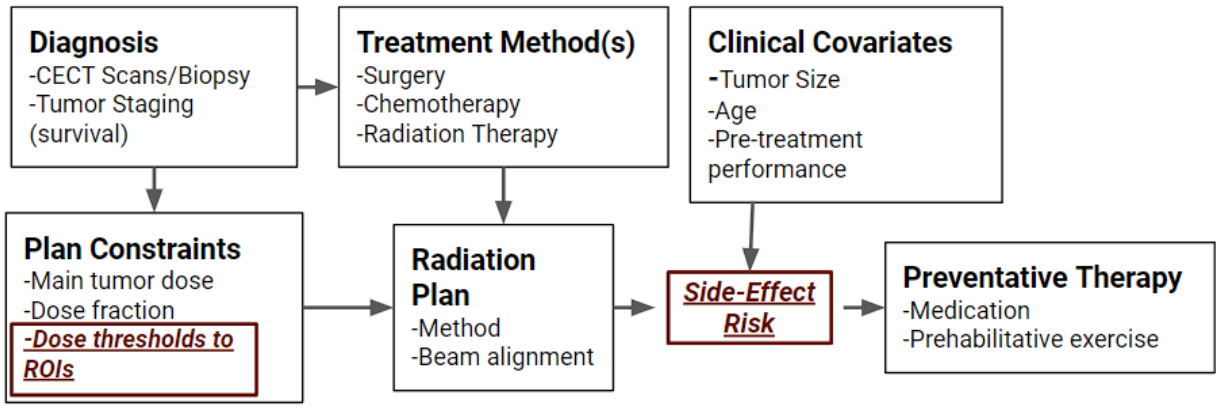


Fig A2. Full 2D Mapping of Regions of Interest. Regions represent approximate locations of important anatomical structures along the center, and sides of the head. The relative size of smaller regions are exaggerated to improve visibility.

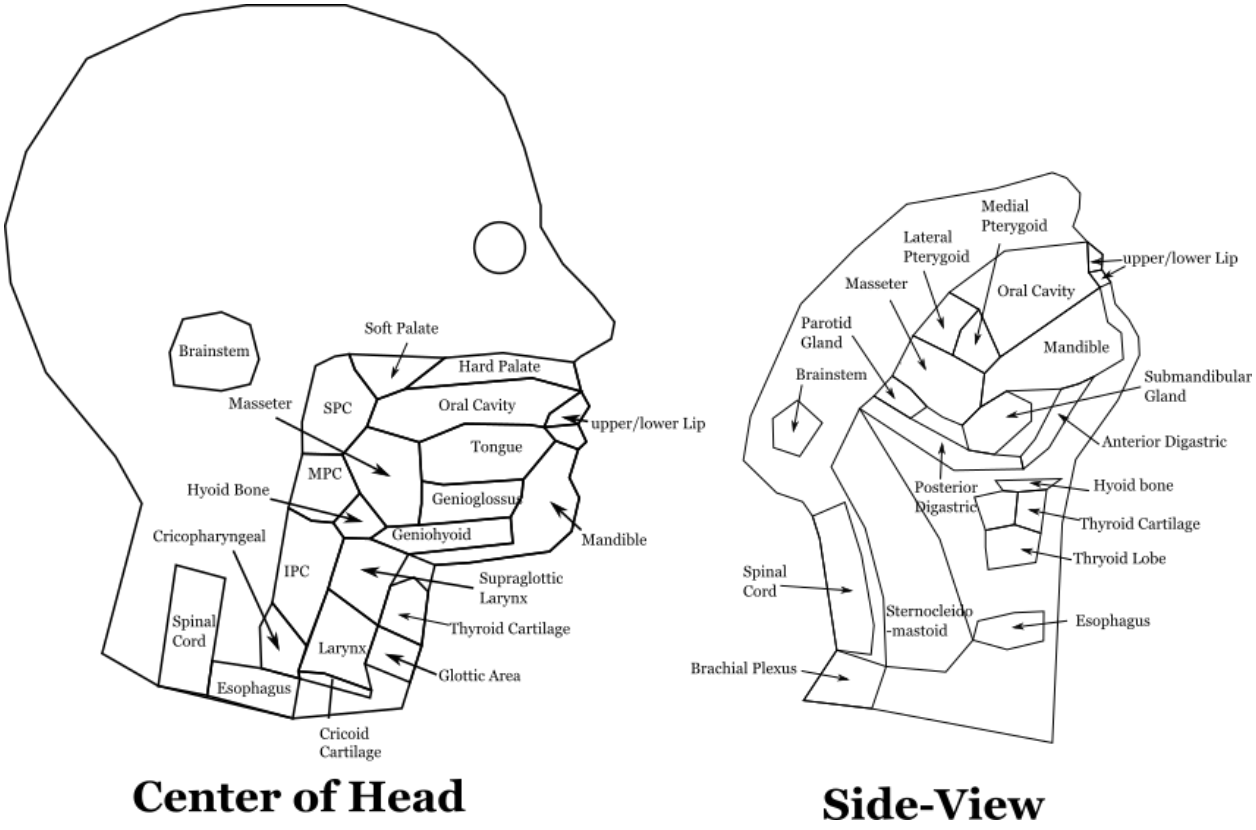
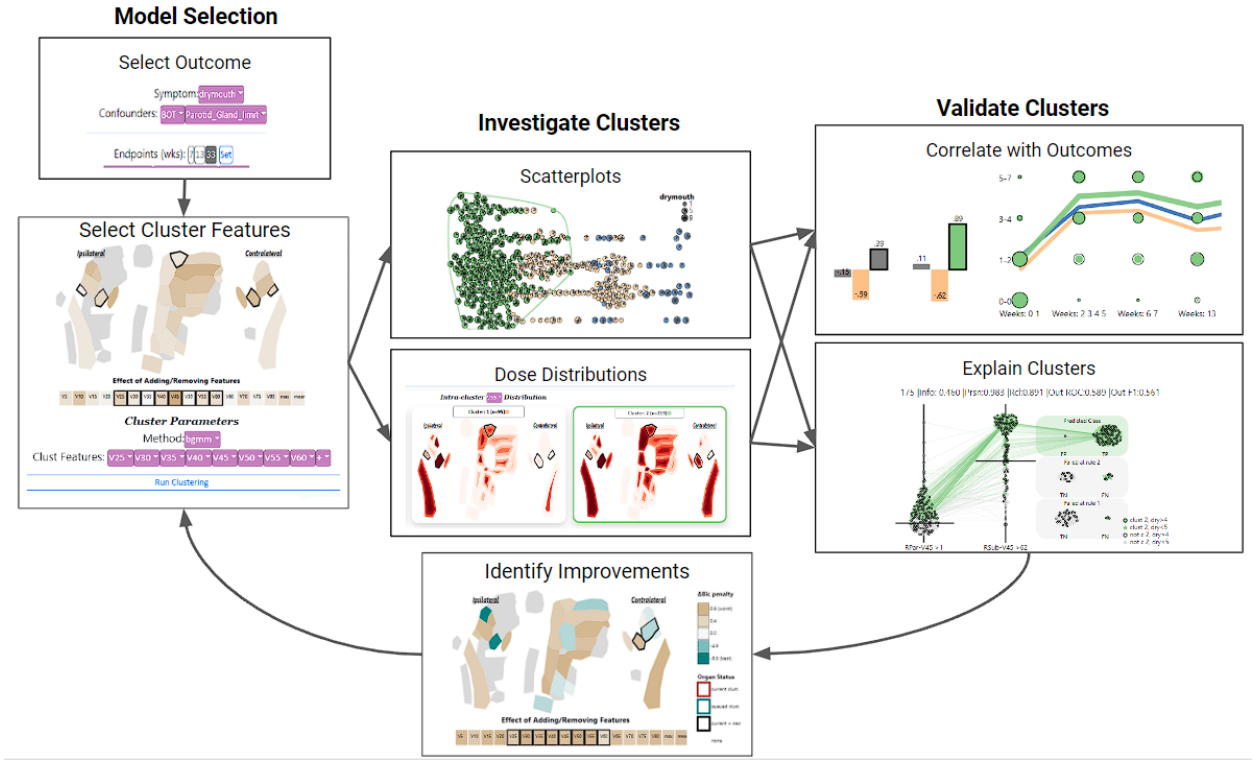


Fig A3. DASS Workflow. First, a desired outcome is selected, along with the initial clustering parameters, which can be drawn from prior literature or informed from information metrics in the explanation view. After clusters are generated, the inter-cluster distributions can be investigated using the dose distribution and configurable scatterplot views. Cluster performance can then be validated by investigating the inter-cluster symptom trajectories and correlations with outcomes. A rule-based classifier can also be used to produce explanations for the high or low-risk clusters based on dose thresholds. Once clusters are investigated, the Additive Effects panel can identify potential changes to the clustering parameters that could improve the model performance.



B. Prototypes and Designs

This section details some of the early prototypes and approaches that led up to our final design of DASS, which covers both algorithmic and visualization attempts.

Fig B1. Early prototype of the visual scaffolding interface used to inspect the results of clusters generated using grid-search or physician-selected features without visual steering. Each row shows the distribution of mean-dose to each ROI within each cluster. Bar charts on the left show odds-ratios of different symptoms. ROIs used in the cluster use a saturated red color scheme, while ROIs not used in the clusters use a desaturated grayscale color scheme. This was our original approach that inspired the design of MOTIV, as we felt that our clusters were not producing outcomes that achieved both good performance and explanations that aligned with clinical knowledge.

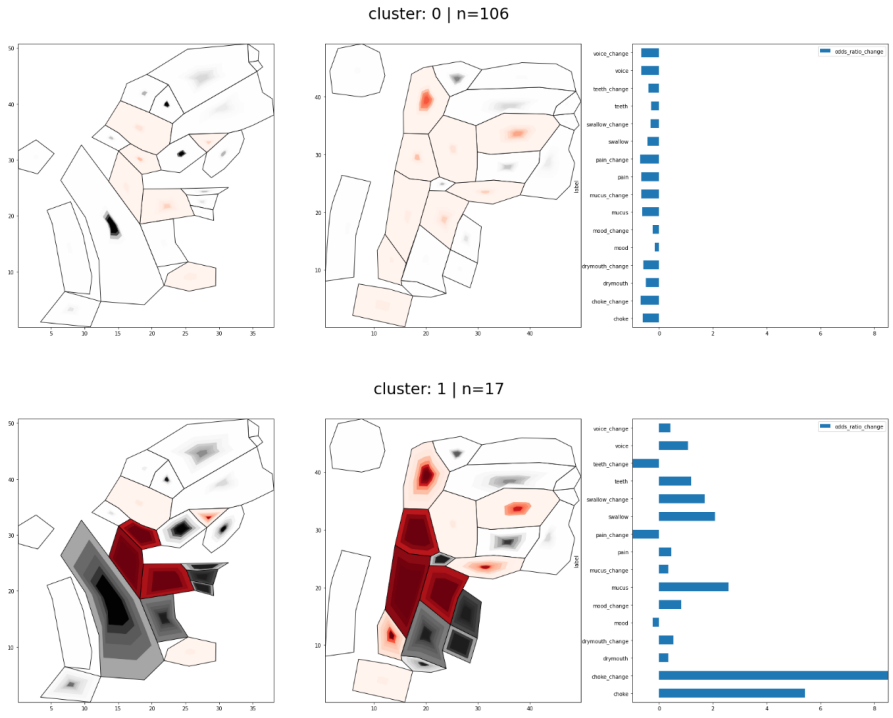


Fig B2. Alternative Approach to producing cluster explanations. For each cluster, we trained a spline-regression model to predict cluster membership using the dose-values used in the cluster. For each roi we selected the dose-value (e.g. V15, V20) to the roi that was the most strongly correlated with the cluster using a Chi-square test. Each row represents a cluster. Charts on the left are the shape-curves for each input to the spline model. Bar charts on the left encode the parameter coefficients for each dose-value.
This approach was abandoned as the model results do not correspond with the expected behavior, in which all shape curves should be monotonically decreasing or monotonically increasing.

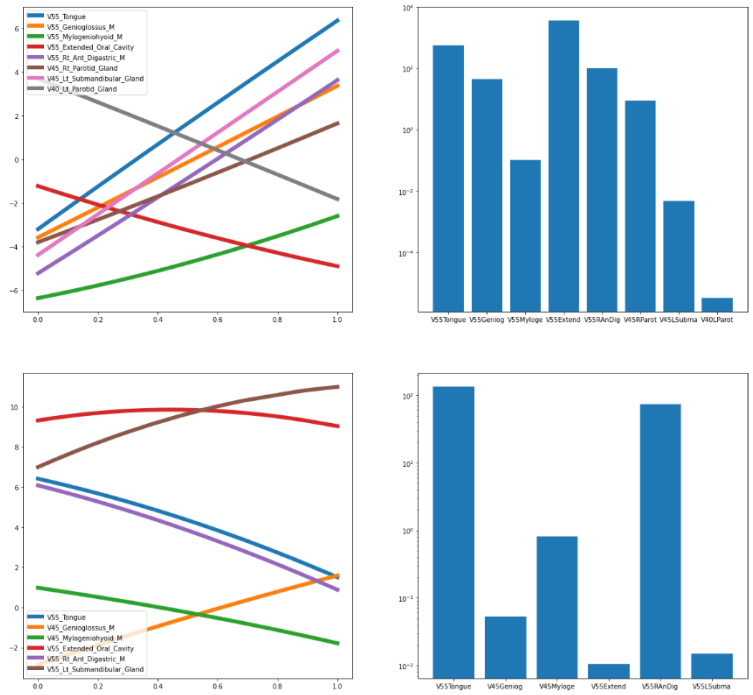


Fig B3. Graph of an earlier attempt at creating a clinical risk model. Dose clusters were generated using pre-defined groups of organs based on region or function. A constrained causal learning model (Fast Greedy Equivalent Search) was then run on the cluster memberships, clinical covariates, and severe symptom ratings at different time points, as well as membership in clusters of symptoms taken from prior work. This approach was implemented in hopes of establishing causal clusters, but was abandoned as it was difficult to consistently generate results that were useful for clinical workflows.

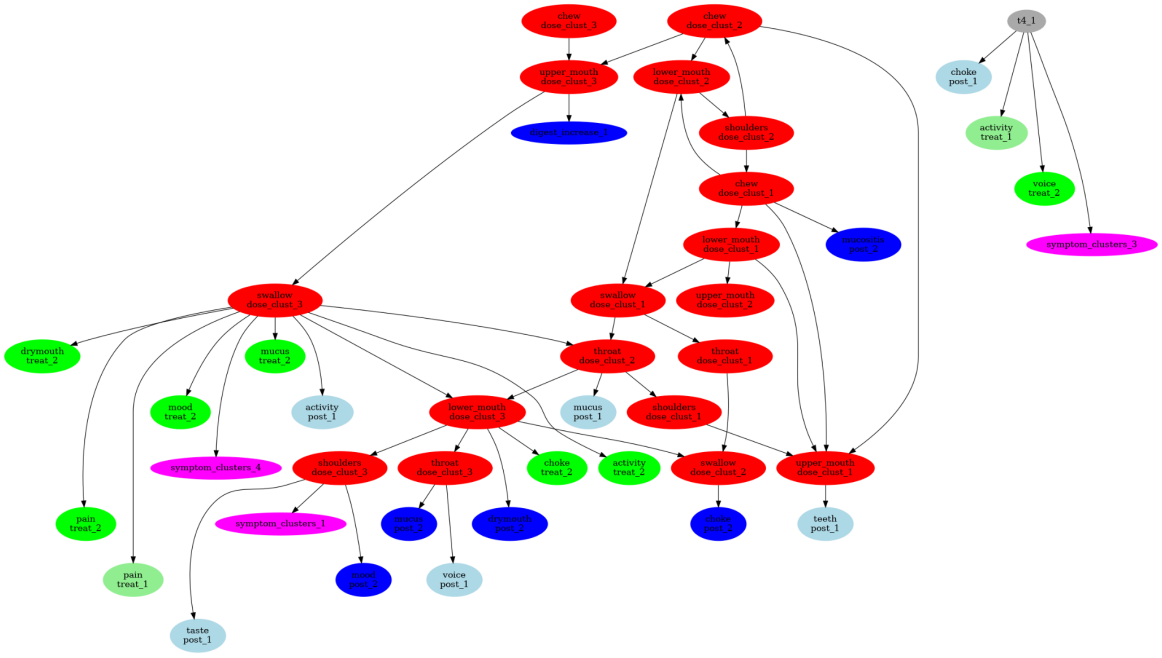


Fig B4. View that shows the effect of each cluster on a machine learning model for predicting different presence of a symptom at 6 months at different thresholds. A user selected model is trained to predict each outcome using clinical variables. Additional models are built that also include cluster labels (all or individual cluster labels), and cross-validation performance is measured. Outcomes are grouped by the metric used, and the dose threshold considered (3,5,7). This view was removed as our collaborators preferred to rely on traditional statistics with fewer thresholds when deciding on outcomes.

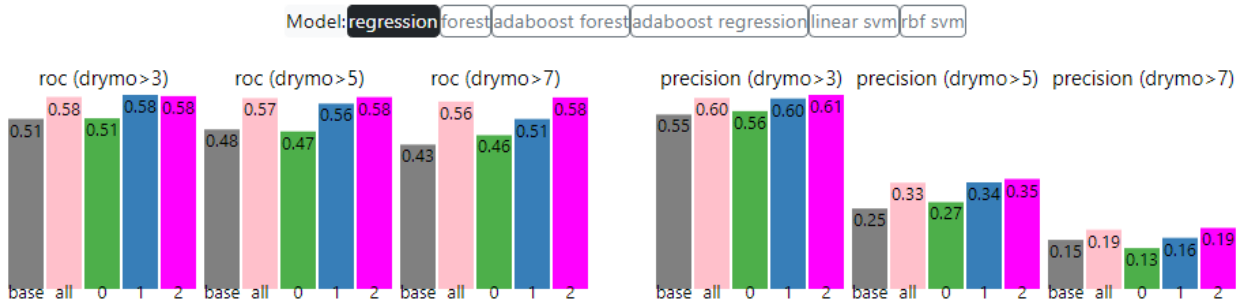


Fig B5. Early version of the interface using patient-specific view and star-charts to encode symptom ratings for patients and clusters. The patient view (top-right) used a single selected patient (133) within a selected cluster at the top. The additional patients on the left are the most similar within the same cluster, while the patients on the left show similar patients in different clusters, based on treatment categories and doses. Patients in the cluster are encoded with a user-selected dose value (e.g. V55) to each organ in red, while patients in other clusters are color-encoded with the difference in dose values from the selected patient using a blue-green color scale, where green indicates higher dose, and blue indicates lower doses. Star charts encode patient symptoms. This view was removed as we found that inspecting patients via a tooltip on the scatterplot was more efficient.

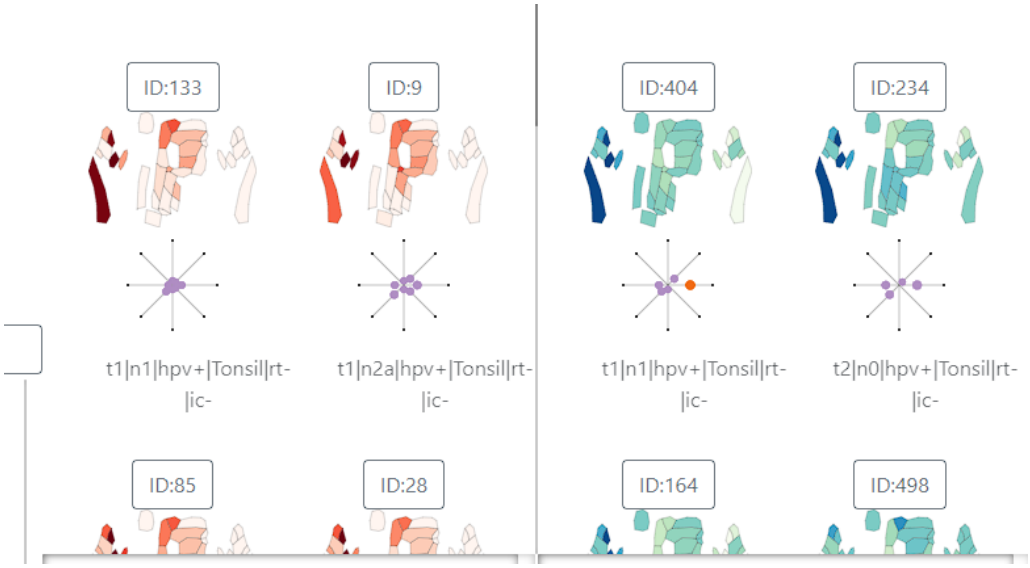


Fig B6. Alternative Variants of the patient encodings that use a color gradient to show the full dose-histogram within each organ. The outer layer represents the V5 while the innermost layer represents the V75 to the respective organ. Symptoms at different stages (baseline, during treatment, and post-treatment) are shown as bar charts using small multiples for a set of symptoms of interest. Patient color encodings for the counterfactual patients show absolute rather than relative doses.

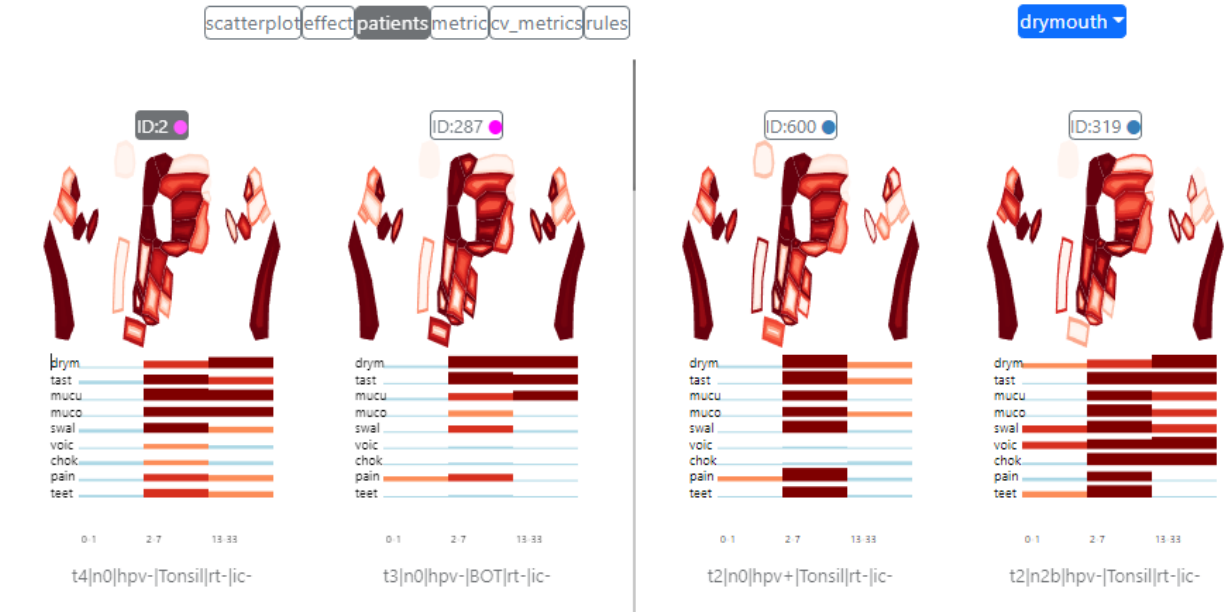


Fig B7. Alternative Version of the scatterplot that used shape instead of glyphs to encode symptom ratings. Circles, triangles, and diamonds encoded patients with symptom ratings of < 5, 5-7, or >7, respectively. Shape-encodings were changed to glyphs as our collaborators felt that the encodings were unintuitive.

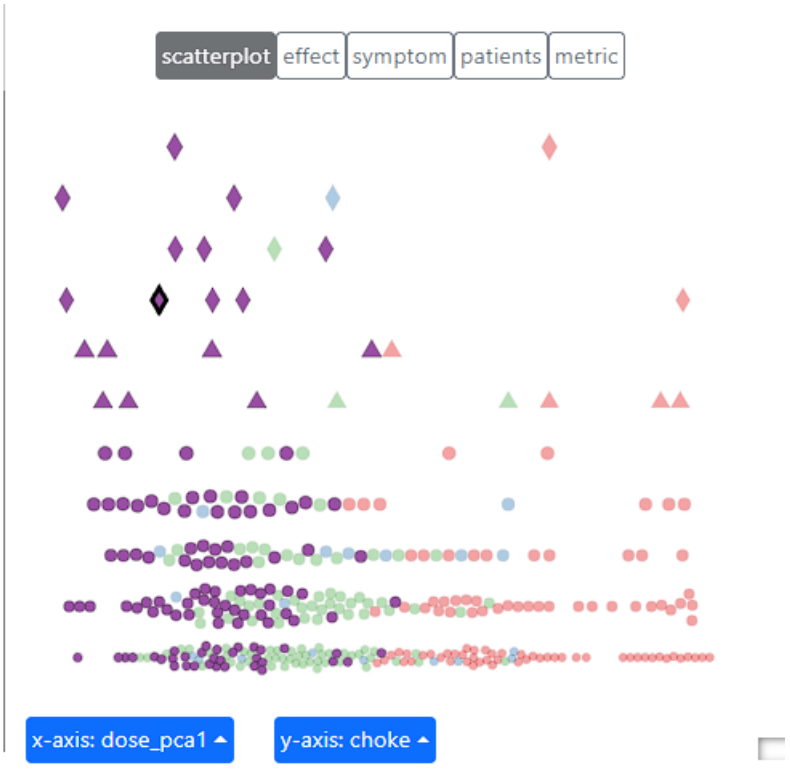


Fig B8. early version of the cluster view. The view on the right encodes cluster doses as in the final version. A star chart on the center-right encodes symptom ratings within the cluster, where the dot encodes median rating and the bold line indicates the 25-75% confidence interval. Circle size encodes odds-ratio for the cluster to experience severe symptoms. Finally, histograms on the right encode the distribution of clinical features within each cluster. The symptom charts were replaced with our current symptom plots as we wanted to focus on comparing trajectories between clusters. Clinical feature plots were removed to save space, as we did not use the information in the development of clusters outside of using them as covariates in our predictive models. Instead, extensive demographic distributions were calculated after clusters were finalized and thus did not need to be included in the interactive system.

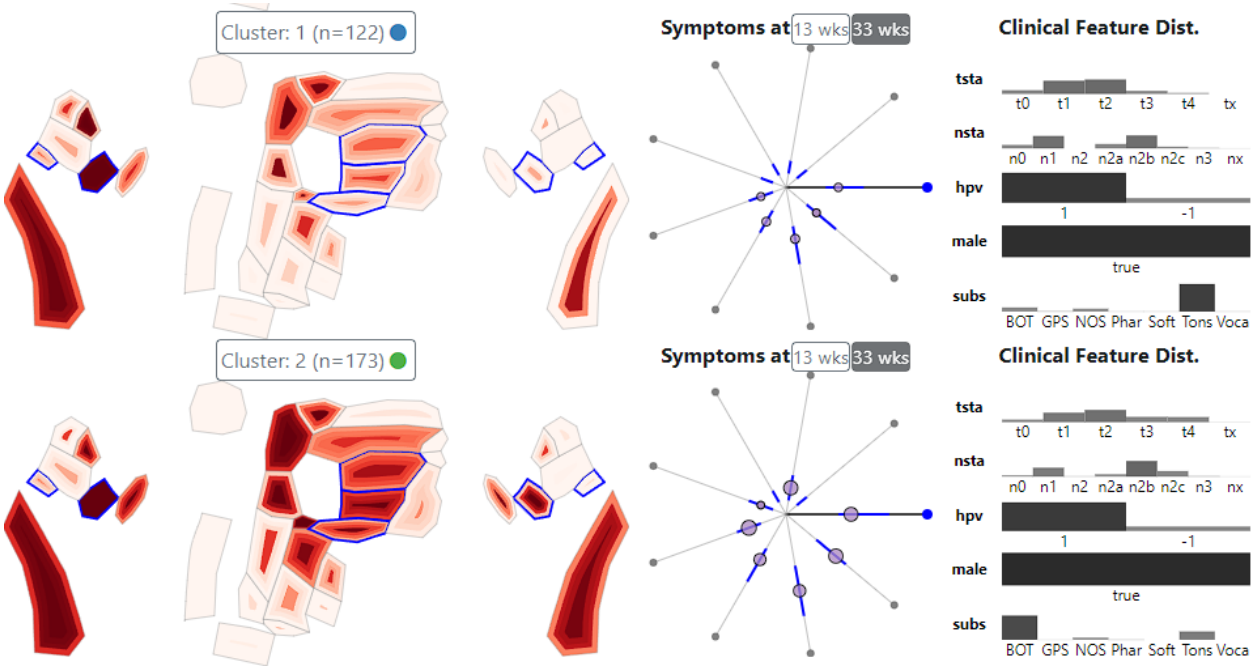
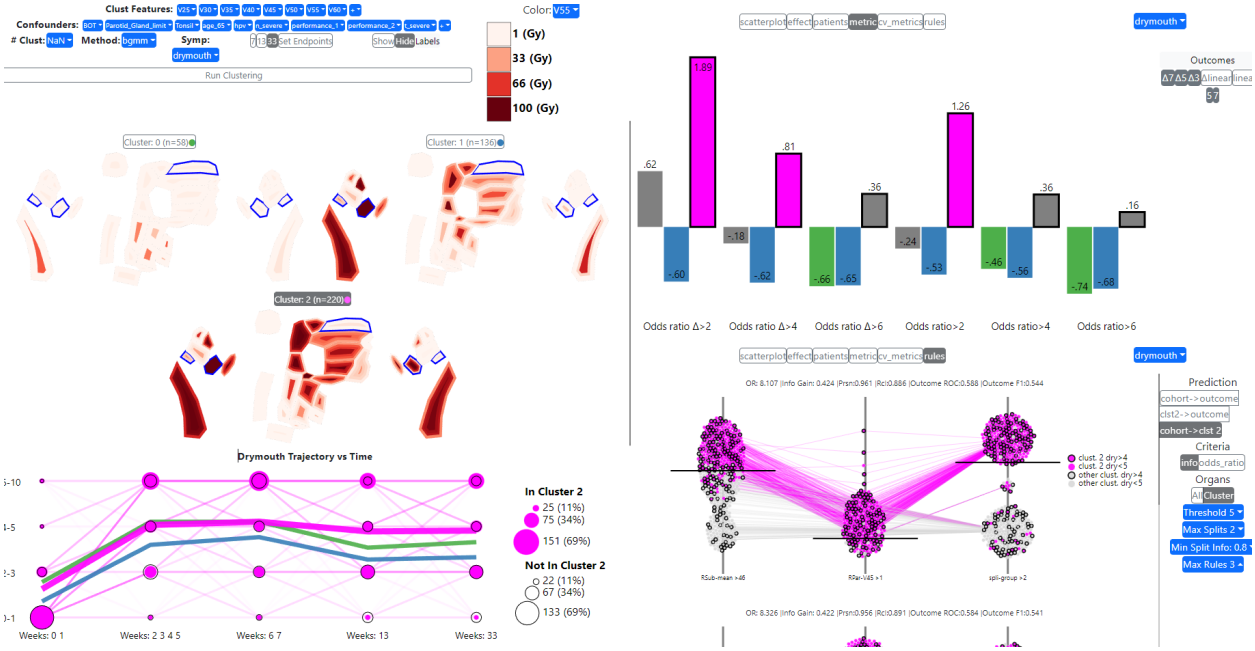
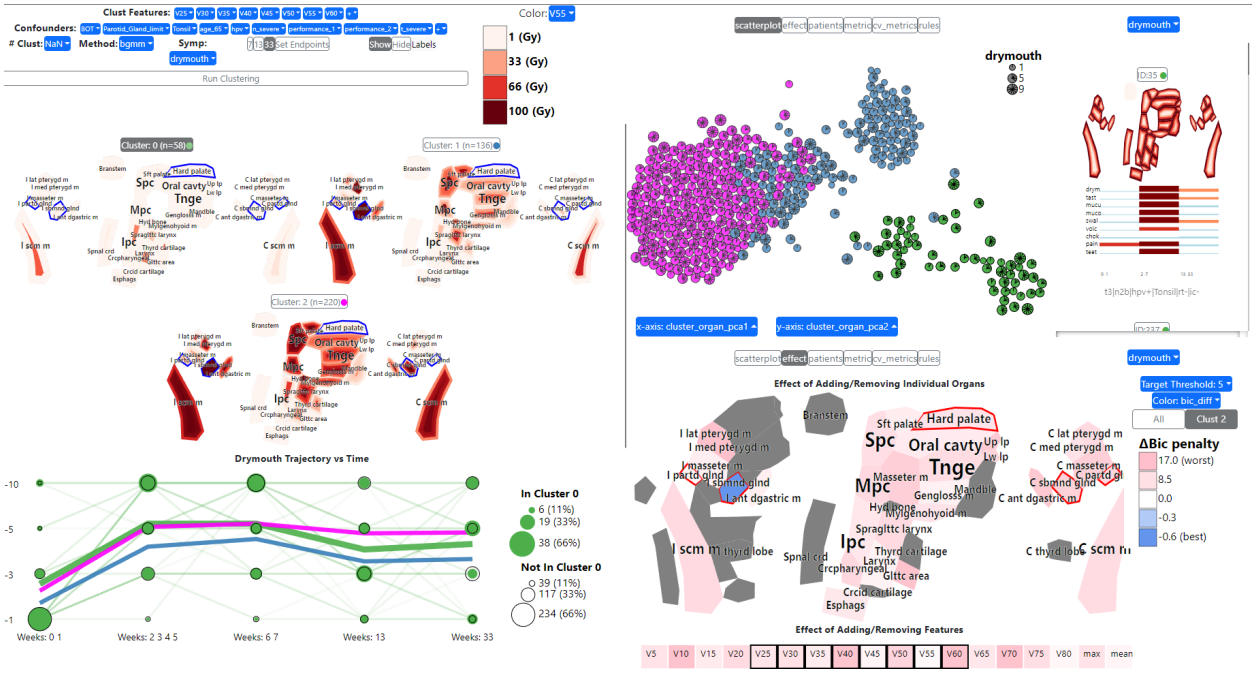


Fig B9. Early design of the layout of DASS in different configurations. The original layout was divided into 5 views: a control panel (top-left), the cluster view (center left), the symptom trajectory view (bottom-left), and two windows on the left that allow the user to toggle between views. Several adjustments to the layout and color schemes were made to de-clutter the interface based on feedback from collaborators. We also incorporated control-panels for each view into titles for each plot in order to make them more readable.





C. Cluster Quantitative Analysis

We include here a comparison of the performance of the clusters contained using our system to baseline clusters that are obtained via methods without visual steering. In this experiment, we use a baseline method of generating clusters to predict patient outcomes without use of the DASS interface. We then compare those results to clusters made from the Dass interface using two kinds of starting parameters: a default set determined by treatment guidelines, and the parameters used in the baseline clusters. Clusters are evaluated using a simple ablation study where a regression model is built to predict the desired outcome with and without the cluster label to test the effect of the model on cross validation model performance. We perform this experiment on three outcomes: late severe (rating >5) drymouth, swallow dysfunction, and choking.

To generate baseline clusters, we tested the results of clusters generated through a combination of input features and organs with 2-4 clusters, using a bayesian mixture model. Because we require that the organs in the model be linked to the outcome, we compiled lists of potential organ combinations for each outcome based on clinical literature. For drymouth, we consider the set of organs taken from NTCP models built for predicting xerostomia [1], and for swallow and choke, we use organs taken from NTCP models built for dysphagia [2]. In addition, we included sets of organs gathered from clinician feedback on the organs they believe would be related to each.

To compare our models, we used two variants of clusters that are obtained from DASS for each outcome. As a baseline, we found clusters following the regular approach, where we used a general starting point of the most important organ of interest. In addition, we consider models

that are obtained from Dass by using the model parameters in the baseline models found using grid-search. Simplified clusters for all cases are generated as described in the main text. The parameters of the final clusters are listed in Table C1.

For quantitative analysis, we convert our clusters into a clinical stratification during cross-validation as follows: first, we rank each cluster based on the number of patients that experience the given outcome in the training data split and assign risk to patients in the test data based on the rank of their clusters, where the highest-risk cluster is given a risk score of 1, the second highest-risk cluster is given a score of (number of clusters - 2)/(number of clusters - 1), etc, and the lowest risk is given a score of 0. For the simplified cluster, we always assign a risk of 1 to the high-dose cluster and 0 otherwise. We then calculate the cross-validation mean Area-under the curve score (AUC), and the Mathews correlation coefficient (MCC) for predicting each outcome for each cluster model. In addition, we calculate the results from using only the highest-dose (HD) cluster for each stratification, as this is a likely use case for clinical applications when deciding whether to prescribe preventative treatment. Thus, we generate AUC and MCC scores for 3 different cases for each clustering.

Results for AUC are shown in Figure C1. Drymouth clusters obtained using DASS, the main symptom of interest that we used to build the model, perform better for all measures compared to grid-search. Choke clusters perform better in general, although interestingly, grid-search simplified explanations tend to outperform other rule-based simplified explanations. For swallow, the second Dass clusters outperform the gridsearch cluster for the stratification but not the rule-based clusters.

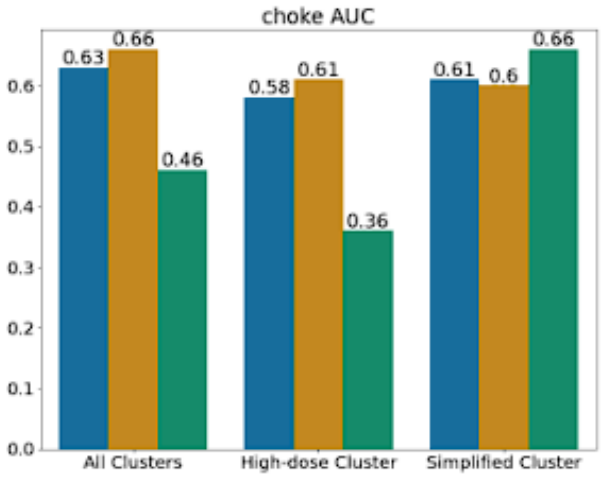
We note that while our swallow and choke clusters have mixed results, we can reasonably beat the baseline clusters for several measures without the need for extensive literature review to identify the optimal organs to consider when testing multiple clusters, in addition to the general benefits of interactive changing of the clusters to balance between other considerations such as the maximum number of clusters or comparison of performance across multiple thresholds. Drymouth results tended to be the most consistent, most likely due to the higher rates of severe drymouth, while severe swallow and choke are relatively rare, making the cluster predictions difficult. In addition, the organs linked to drymouth tend to be salivary glands rather than muscles, and are likely more sensitive to lower doses.

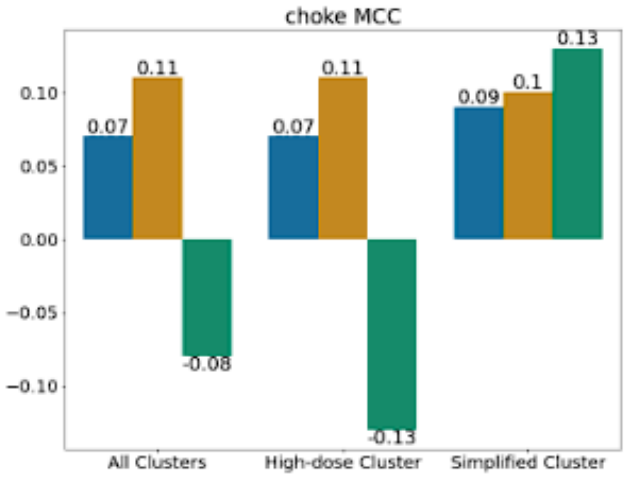
Table C1. Table of features used in each cluster for our analysis. The prefixes ‘Lt’ and “Rt” refer to the ipsilateral (same side as main tumor) and contralateral (opposite side as main tumor) organ, respectively. Legend: Dass: Models built using all clusters using DASS with a default starting point; Alt Dass: Models built using DASS with the baseline clusters at a starting point. Baseline: Models built using clusters obtained through standard grid search.

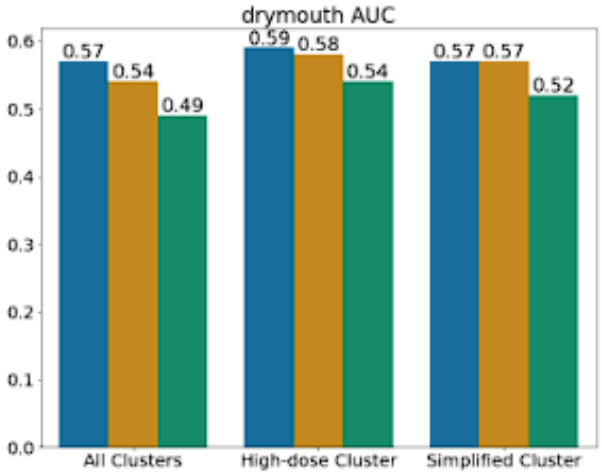
Symptom	organs	N Clusters	Organs	Dose Features	Simple Rule
Drymouth	Dass	3	Rt_Parotid_Gland, Lt_Parotid_Gland, Rt_Submandibular_Gland, Lt_Submandibular_Gland, Hard_Palate	V25-V60	V55_Rt_Submandibular_Gland>23.0, V65_Lt_Submandibular_Gland>51.0, V10_Rt_Parotid_Gland>18.0
	Dass Alt	3	Rt_Parotid_Gland, Lt_Parotid_Gland, Rt_Submandibular_Gland, Lt_Submandibular_Gland, Soft_Palate	V30-V55	V55_Rt_Submandibular_Gland>23.0, V65_Lt_Submandibular_Gland>51.0, V10_Rt_Parotid_Gland>18.0
	Baseline	3	Lt_Parotid_Gland, Rt_Parotid_Gland, Lt_Submandibular_Gland, Rt_Submandibular_Gland, Soft_Palate, Upper_Lip, Lower_Lip, Extended_Oral_Cavity, Mylogeniohyoid_M	mean dose, max dose	V60_Mylogeniohyoid_M>32.0, V55_Extended_Oral_Cavity>28.0, V30_Rt_Submandibular_Gland>62.0, V65_Lt_Submandibular_Gland>1.0
Choke	Dass	3	IPC, MPC, Supraglottic_Larynx, Esophagus, Mylogeniohyoid_M	V30-V65	V60_Supraglottic_Larynx>40.0, V55_MPC>17.0, V55_Mylogeniohyoid_M>24.0, V15_Esophagus>22.0
	Dass Alt	3	IPC, MPC, Supraglottic_Larynx, Cricopharyngeal_Muscle, Mylogeniohyoid_M	V30, V40, V50, V60	mean_dose_IPC>49.0, mean_dose_MPC>54.0, mean_dose_Cricopharyngeal_Muscle>16.0
	Baseline	3	SPC, IPC, Supraglottic_Larynx, Rt_Parotid_Gland, Cricopharyngeal_Muscle	mean dose, max dose	V5_Rt_Parotid_Gland>72.0, V65_SPC>3.0, V55_Supraglottic_Larynx>46.0, max_dose_IPC>46.0
Swallow	Dass	3	IPC, MPC, SPC' Supraglottic_Larynx, Esophagus	V35-V70	V60_Supraglottic_Larynx>40.0, V65_SPC>3.0, V15_Esophagus>22.0,

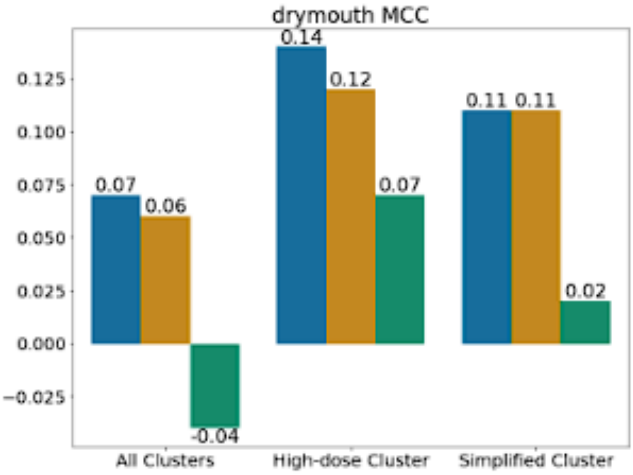
					V55_MPC>17.0
	Dass Alt	2	IPC, SPC, Cricopharyngeal_Muscle, Rt_Parotid_Gland	V65, max dose	max_dose_Cricopharyngeal_Muscle>63.0
	Baseline	2	SPC, IPC, Supraglottic_Larynx, Rt_Parotid_Gland, Cricopharyngeal_Muscle	mean_dose, V40-V65	V5_Rt_Parotid_Gland>72.0, V65_SPC>3.0, V55_Supraglottic_Larynx>46.0, max_dose_IPC>46.0

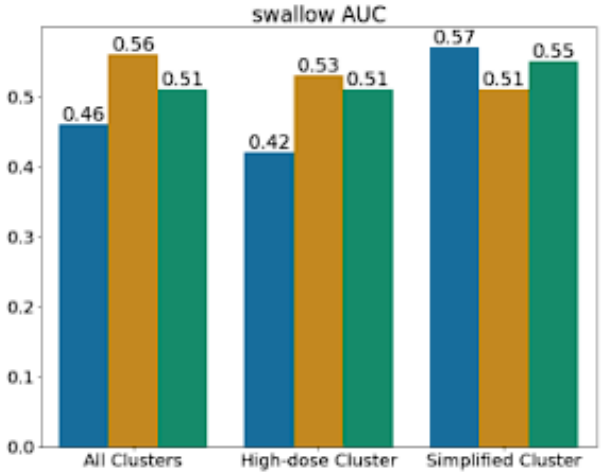
Figure C1. AUC and MCC scores for all model types. For cluster stratifications, Alt Dass clusters and high-dose clusters outperform the baseline models. Alt Dass performs better than Dass for Choke and Swallow, but not Drymouth.

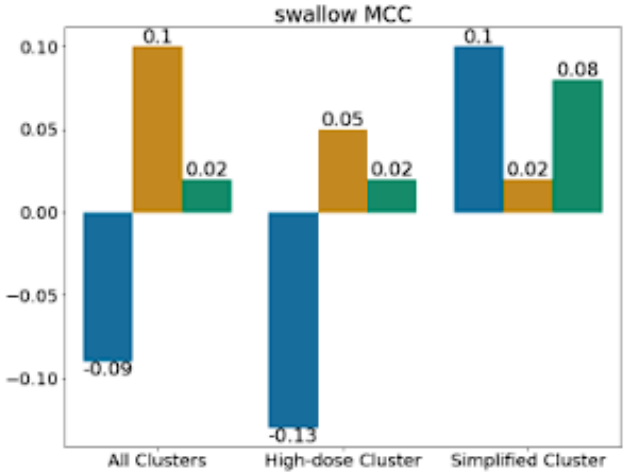

choke AUC
0.63
0.66
0.46
0.58
0.61
0.36
0.61
0.6
0.66
All Clusters
High-dose Cluster
Simplified Cluster


choke MCC
0.07
0.11
-0.08
0.07
0.11
-0.13
0.09
0.1
0.13
All Clusters
High-dose Cluster
Simplified Cluster


drymouth AUC
0.57
0.54
0.49
0.59
0.58
0.54
0.57
0.57
0.52
All Clusters
High-dose Cluster
Simplified Cluster


drymouth MCC
0.07
0.06
-0.04
0.14
0.12
0.07
0.11
0.11
0.02
All Clusters
High-dose Cluster
Simplified Cluster


swallow AUC
0.46
0.56
0.51
0.42
0.53
0.51
0.57
0.51
0.55
All Clusters
High-dose Cluster
Simplified Cluster


swallow MCC
-0.09
0.1
0.02
-0.13
0.05
0.02
0.1
0.02
0.08
All Clusters
High-dose Cluster
Simplified Cluster


Legend
DASS Clusters
Dass Clusters 2
Baseline

Table C2. Results from 5-fold cross-validation of different stratification models for predicting severe late symptoms. We consider both absolute ratings > 4 at 6 months, and increase in ratings > 4 from baseline at 6 months. Models include stratification with all clusters (stratification), stratification using only the highest dose clusters (High-dose), and the simplified high-risk rule explanations. Legend: Dass: Models built using all clusters using DASS with a default starting point; Alt Dass: Models built using DASS with the baseline clusters at a starting point. Baseline: Models built using clusters obtained through standard grid search.

Symptom	Change From Baseline	Model	ROC-AUC	MCC	Precision	Recall	F1
		Dass Stratification	0.628	0.070	0.059	0.500	0.106
		Alt Dass Stratification	0.659	0.106	0.080	0.429	0.135
		Baseline Stratification	0.462	-0.082	0.026	0.357	0.048
		Dass High-dose	0.584	0.070	0.059	0.500	0.106
		Alt Dass High-dose	0.611	0.106	0.080	0.429	0.135
		Baseline High-dose	0.357	-0.126	0.000	0.000	0.000
		Dass Simplified	0.611	0.088	0.060	0.643	0.110
		Al Dass Simplified	0.596	0.100	0.083	0.357	0.135
	FALSE	Baseline Simplified	0.662	0.127	0.066	0.786	0.122
		Dass Stratification	0.652	0.065	0.051	0.500	0.092
		Alt Dass Stratification	0.714	0.131	0.080	0.500	0.138
		Baseline Stratification	0.469	-0.085	0.021	0.333	0.039
Choke	TRUE						

		Dass High-dose	0.584	0.065	0.051	0.500	0.092
		Alt Dass High-dose	0.648	0.131	0.080	0.500	0.138
		Baseline High-dose	0.358	-0.116	0.000	0.000	0.000
		Dass Simplified	0.623	0.090	0.053	0.667	0.099
		Al Dass Simplified	0.627	0.122	0.083	0.417	0.139
		Baseline Simplified	0.685	0.135	0.060	0.833	0.112
		Dass Stratification	0.571	0.065	0.220	0.574	0.318
		Alt Dass Stratification	0.540	0.058	0.216	0.603	0.318
		Baseline Stratification	0.490	-0.037	0.179	0.426	0.252
		Dass High-dose	0.586	0.137	0.244	0.691	0.360
		Alt Dass High-dose	0.577	0.124	0.236	0.706	0.354
		Baseline High-dose	0.542	0.069	0.215	0.706	0.330
		Dass Simplified	0.568	0.108	0.241	0.574	0.339
		Al Dass Simplified	0.568	0.108	0.241	0.574	0.339
	FALSE	Baseline Simplified	0.515	0.025	0.208	0.382	0.269
		Dass Stratification	0.583	0.046	0.141	0.568	0.226
		Alt Dass Stratification	0.546	0.035	0.137	0.591	0.222
		Baseline Stratification	0.485	-0.059	0.105	0.386	0.165
		Dass High-dose	0.613	0.150	0.171	0.750	0.278
		Alt Dass High-dose	0.596	0.130	0.163	0.750	0.267
		Baseline High-dose	0.551	0.070	0.143	0.727	0.240
		Dass Simplified	0.573	0.097	0.160	0.591	0.252
Drymouth	TRUE						

		AI Dass Simplified	0.573	0.097	0.160	0.591	0.252
		Baseline Simplified	0.516	0.022	0.136	0.386	0.201
		Dass Stratification	0.465	-0.091	0.140	0.393	0.206
		Alt Dass Stratification	0.560	0.104	0.247	0.328	0.282
		Baseline Stratification	0.513	0.024	0.180	0.770	0.292
		Dass High-dose	0.418	-0.125	0.122	0.311	0.175
		Alt Dass High-dose	0.533	0.053	0.190	0.689	0.298
		Baseline High-dose	0.512	0.021	0.179	0.770	0.291
		Dass Simplified	0.567	0.102	0.218	0.557	0.313
		AI Dass Simplified	0.511	0.018	0.179	0.754	0.289
	FALSE	Baseline Simplified	0.550	0.075	0.205	0.557	0.300
		Dass Stratification	0.506	-0.021	0.116	0.465	0.186
		Alt Dass Stratification	0.538	0.083	0.173	0.326	0.226
		Baseline Stratification	0.523	0.040	0.130	0.837	0.225
		Dass High-dose	0.431	-0.092	0.090	0.326	0.141
		Alt Dass High-dose	0.510	0.014	0.127	0.651	0.212
		Baseline High-dose	0.494	-0.011	0.122	0.837	0.212
		Dass Simplified	0.577	0.101	0.160	0.581	0.251
		AI Dass Simplified	0.504	0.007	0.125	0.744	0.213
Swallow	TRUE	Baseline Simplified	0.560	0.079	0.151	0.581	0.239

Table C3. Distribution of outcomes in the cohort considered.

Symptom	Total >4 at 6M	Mean Rating	95% CI
Drymouth	153 (43.84%)	4.338109	0.4-9
Swallow	46 (12.18%)	2.137536	0-7
Choke	19 (5.44%)	1.111748	0-5

Table C4. Parameters used in the gridsearch for the baseline. All parameters were tested with a gaussian mixture models using 2-4 components

Dose Values (All Outcomes)
['mean_dose','V25','V30','V35','V40','V45','V50','V55','V60','V65','V70','V75','V80'], ['mean_dose','V25','V30','V35','V40','V45','V50'], ['mean_dose','V25','V30','V35','V40'], ['mean_dose','V30','V35','V40','V45','V50','V55','V60','V65','V70','V75','V80'], ['mean_dose','V40','V45','V50','V55','V60','V65','V70','V75','V80'], ['mean_dose','V55','V60','V65','V70','V75','V80'], ['mean_dose','V60','V65','V70','V75','V80'], ['mean_dose','V40','V45','V50','V55','V60','V65'], ['mean_dose','V40','V45','V50','V55','V60'], ['V25','V30','V35','V40','V45','V50','V55','V60','V65','V70','V75','V80'], ['V25','V30','V35','V40','V45','V50'], ['V25','V30','V35','V40'], ['V30','V35','V40','V45','V50','V55','V60','V65','V70','V75','V80'], ['V40','V45','V50','V55','V60','V65','V70','V75','V80'], ['V55','V60','V65','V70','V75','V80'], ['V60','V65','V70','V75','V80'], ['V40','V45','V50','V55','V60','V65'], ['V40','V45','V50','V55','V60'], ['mean_dose'], ['max_dose'], ['max_dose','mean_dose']
Organs (Drymouth)

['Tongue', 'Mylogeniohyoid_M', 'Genioglossus_M','Rt_Parotid_Gland','Lt_Parotid_Gland', 'Rt_Submandibular_Gland','Lt_Submandibular_Gland', 'Soft_Palate', 'Extended_Oral_Cavity','Supraglottic_Larynx','Larynx'] ['Lt_Parotid_Gland','Rt_Parotid_Gland','Lt_Submandibular_Gland','Rt_Submandibular_Gland','Soft_Palate', 'Upper_Lip','Lower_Lip', 'Extended_Oral_Cavity','Mylogeniohyoid_M']
Organs (Swallow)
['SPC','IPC','Supraglottic_Larynx','Rt_Parotid_Gland','Cricopharyngeal_Muscle'] ['IPC','MPC', 'SPC', 'Mylogeniohyoid_M','Tongue'] ['SPC','IPC','MPC','Supraglottic_Larynx', 'Rt_Parotid_Gland','Cricopharyngeal_Muscle']
Organs (Choke)
['SPC','IPC','Supraglottic_Larynx','Rt_Parotid_Gland','Cricopharyngeal_Muscle'] ['Rt_Masseter_M', 'Lt_Masseter_M', 'Rt_Medial_Pterygoid_M', 'Lt_Medial_Pterygoid_M', 'Rt_Lateral_Pterygoid_M', 'Lt_Lateral_Pterygoid_M'] ['Rt_Masseter_M','Lt_Masseter_M','Rt_Medial_Pterygoid_M', 'Lt_Medial_Pterygoid_M', 'Rt_Lateral_Pterygoid_M', 'Lt_Lateral_Pterygoid_M', 'Supraglottic_Larynx', 'Larynx','Glottic_Area','Thyroid_cartilage', 'Cricopharyngeal_Muscle', 'Cricoid_cartilage','Esophagus']

[1] 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.

[2] 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.

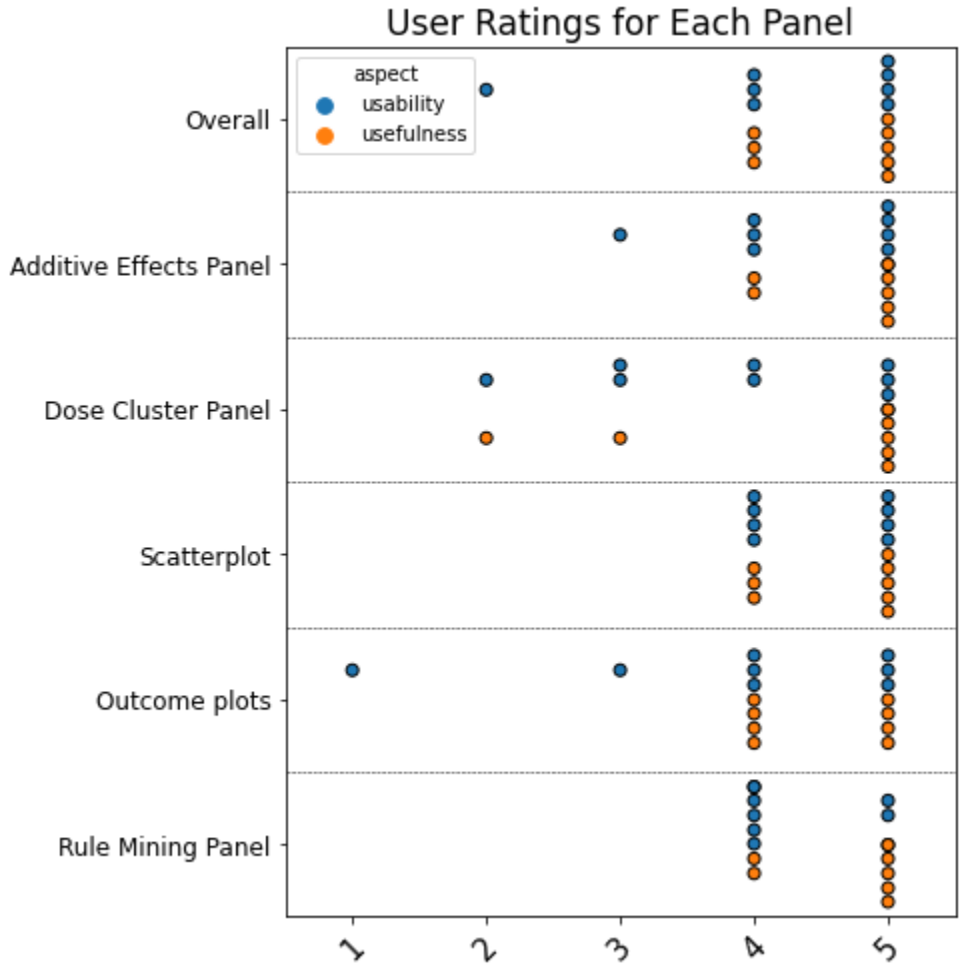
D. Qualitative Analysis Results

Table D1. Results From the usability evaluation questionnaire. Ratings are given on a range of 1 (not useful) to 5 (very useful).

	Additive Effects	Dose Clusters	Scatterplot	Outcome Plots	Rule Mining	Overall

									Explanations			
	Useful	Usable	Usefule	Usable	Usefule	Usable	Usefule	Usable	Usefule	Usable	Usefule	Usable
P1	4	4	5	4	5	4	5	4	5	4	5	4
P2	5	5	5	5	5	5	4	5	5	4	5	5
P3	5	5	5	5	5	5	4	3	4	4	5	5
P4	5	5	5	5	5	5	4	4	4	4	5	5
P5	5	4	5	3	5	5	5	5	5	4	4	4
P6	5	4	3	3	4	4	5	4	5	5	4	4
P7	4	3	2	2	4	4	4	1	5	4	4	2
P8	5	5	5	4	4	4	5	5	5	5	5	5
Average	4.75	4.38	4.38	3.88	4.63	4.50	4.50	3.88	4.75	4.25	4.63	4.25

Figure D1. Plot of user responses to the usability questionnaire.



Open ended comments

P6: The Dose Cluster Panel a bit complex at the beginning. But became clearer later on. Use cases are well explained in the end.
For the additive effects panel, it's easy to see the effect of adding other organs but I believe there is a factor of domain knowledge missing. For instance, adding an organ may look like a good idea but it may not be relevant to the symptom being investigated (based on AW's comments during presentation).

P7: The Dose Cluster panel encoding confused me because the distribution was mapped to spatial locations, and that matching was really trying to encode quantiles. Also it's quantized but the line chart is not.
I see a lot of potential in making the interface more actionable than descriptive. For instance, what's the effect of removing outliers or so, and comparing the cluster afterwards. Not sure if that's the purpose though
I'm not very convinced by the choice of the force layout graphs because they mutate the positions of the scatterplots or swarm for the sake of visibility. In doing so, they add some uncertainty of whether the location is accurate or not.
I liked that the outcome plots show what results are significant, but the p value could be parametrized given the different opinions of what threshold is actually significant. For the temporal trajectory, the choose of size and whether a circle is bigger than another is difficult for me to parse. I correlate the radius size with bigger being better or worse, but I believe it encodes population, and the relationship between which circle includes another circle is actually what matters. I cannot still read that well.
I liked the rule mining pcp the most, but didn't like the force layout for reasons described above. I think it could have more screen space to be able to add more rules if needed.