Fall 2018

Leveraging clustering for dimensionality reduction and improved prognosis in head and neck cancer patients

Joel E Tosado
University of Iowa

Follow this and additional works at: https://ir.uiowa.edu/etd
Part of the Electrical and Computer Engineering Commons

LEVERAGING CLUSTERING FOR DIMENSIONALITY REDUCTION AND IMPROVED PROGNOSIS IN HEAD AND NECK CANCER PATIENTS

by

Joel E. Tosado

A thesis submitted in partial fulfillment
of the requirements for the Doctor of Philosophy
degree in Electrical and Computer Engineering in the
Graduate College of
The University of Iowa

December 2018

Thesis Supervisor: Associate Professor Guadalupe M. Canahuate

ABSTRACT

Survival outcomes, such as overall survival or recurrence-free survival, are called right-censored because for many patients the event has not yet occurred at the last follow-up time. With an increased number of available features and relatively small number of patients and even smaller number of events, dimensionality reduction is needed to reduce the sparsity of the data and make standard approaches such as Cox Proportional Hazards (Cox) model effective. Clustering is used to identify similar groups within the data and can be thought as a dimensionality reduction technique when the cluster label is used in the analysis. Our goal is to identify similar groups of patients that exhibit the same response to treatment or expected outcomes in order to improve the prediction accuracy for new patients.

In this thesis, we explore different ways of leveraging clustering for improved prognosis for head and neck cancer patients. To circumvent the right-censoring of survival outcomes, we use the residuals from a Cox as the dependent variable for guiding clustering of the data. We propose two approaches. The first one, Supervised Scaled Clustering (SSC), uses the residuals to scale the features using a regression model before clustering the patients using K-medians and consensus clustering. The second one, Supervised Domain Clustering (SDC), considers groups of features and uses the residuals to learn the most suitable dissimilarity for clustering. Cluster labels are then used as covariates within a Cox model and/or other survival models. A rigorous experimental evaluation summarizes, compares and contrasts different metrics for model comparison and performance evaluation. Results show that our approaches find significantly discriminative groupings w.r.t. to the outcomes, and can serve as a feature extraction method that can improve performance while considerably reducing the dimensionality of the original feature space.

PUBLIC ABSTRACT

Survival outcomes, such as overall survival or recurrence-free survival, are called right-censored because for many patients the event has not yet occurred at the last follow-up time. With an increasing number of potential risk factors available that can aid in improving prognosis, standard statistical modeling approaches such as Cox Proportional Hazards may not be as effective in incorporating them. Clustering is a machine learning task with the ultimate goal of identifying similar groups within the data and effectively condensing multiple risk factors represented by the cluster label. In this manner we are able to summarize the increasing number of risk factors and find labels that identify not obvious, yet salient, similarities that result from simultaneously considering these multiple risk factors. Once one or multiple groupings have been identified we evaluate how these groupings discriminate against the survival outcomes of interest. Finally we incorporate clustering into standard approaches for risk modeling and evaluate and quantify the improvement in prognosis.

TABLE OF CONTENTS

LIST OF TABLES

Table

LIST OF FIGURES

Figure

CHAPTER 1
INTRODUCTION

Every year over 50,000 new cases of head and neck cancers are diagnosed in the United States. This number is projected to rise in the future, especially for oropharyngeal cancers, recently been associated with the incidence of HPV16 genotype infections [1].The American Joint Committee on Cancer (AJCC) and the Union for International Cancer Control, maintains an internationally used standardized TNM Staging System. This system serves as a way to systematically assess the severity of the cancer on individual subjects [2]. The vast majority of risk stratification of head neck cancer patients uses staging systems that sub classify patients into four or less groups, based primarily on committee derived treatment standards and approaches using existing data sets. These consider physical examinations, imaging and laboratory tests, pathology and surgical reports, etc. Establishing the AJCC stage for a patient considers various important anatomic classifications and other risk factors that contribute to the overall assessment such as T, N and M Categories. T Category relates to the extent of the primary tumor, N Category relates to the spread to lymph nodes, and M Category indicates the spread outside the T and N related areas. These classifications play a critical role in the ultimate diagnosis and prognosis of an event (or outcome). The ability to more accurately assess the underlying condition such that it improves the prediction on various outcomes is a long standing clinical goal.

In the era of personalized cancer medicine, innovative sources of meaningful data are critically needed and increasingly becoming available. Radiomics is one such “big data” approach that applies advanced image refining/data characterization algorithms to generate imaging features that may be used to quantitatively classify tumor phenotypes in a noninvasive manner [3]. As the number of radiomic features is very large, methods to extract or identify meaningful radiomic signatures that have statistically significant correlations to patient outcomes are needed [4]–[6].

There are multiple outcomes that can be considered in the context of head and neck cancer, such as local control (primary site recurrence of tumor), regional control (recurrence of tumor in non-primary site such as lymph nodes), distant control (distant metastases, spread of the cancer outside of primary), loco-regional control (combination of local and regional), overall survival, and recurrence free survival or RFS (local, regional and distant control).

These patient outcomes are said to be right-censored because for some patients the time-to-event may be unknown. This is the case for patients where the outcome has not been observed up to the last known follow-up time. Right-censored data poses challenges to training methods, specially those that require a known target. Nevertheless, the patients that have yet to incur an event can still provide some useful information in order to predict the probability of an event occurring at a certain time. Survival analysis attempts to use these right-censored outcomes in a meaningful way rather than discarding them or ignoring the censored status.

Right-censored outcomes is not the only challenge for risk modelling. A potentially limiting issue is the sparsity of the data. The relatively small sample size compared to an increasingly high dimensionality, requires us to address models that may result in overfitting, and may be sensitive to noise as a result of it. Moreover, missing data and their imputation, especially for prognosis of new patients, may exacerbate these issues by introducing biases and uncertainty in the analysis.

For prognosis, an important requirement is the interpretability of the results with respect to any meaningful features and perhaps further, with respect to feature values. To this extent, machine learning approaches that facilitate this interpretability are preferred such as Decision Trees (DTs) - and by extension Random Forests which have variable importance measures across iterations. Other widely used highly interpretable methods are Logistic Regression and the ubiquitous Cox Proportional Hazards (cox) regression model from where Hazard Ratios are commonly reported.

In this thesis, we explore different ways of leveraging clustering for improved prognosis for head and neck cancer patients. In this work we combine clinical data from various sources such as Electronic Health Record (EHR), diagnosis, demographic, and radiomic signatures in order to build risk prediction models for OS and RFS. To simplify the discussion, we refer to all non-radiomic features as clinical features. We propose the use of clustering as a dimensionality reduction approach for both clinical and radiomic features. To circumvent the right-censoring of the survival outcomes, the residuals from a Cox proportional hazards model are used as a dependent variable [7].

We explore two main avenues. In the first one, we look at incorporating outcome information into the clustering algorithm in a Supervised Scaled Clustering (SSC) approach. The residuals as the outcome proxy is used to scale the features using a regression model and the patients are clustered using K-medians and consensus clustering. In the second one, we consider several subspaces or groups of features and apply clustering to each domain independently after learning the more relevant dissimilarity(ies) w.r.t. to the residuals. We call this method Supervised Domain Clustering (SDC).

Cluster labels are then used as covariates within a Cox model and/or other survival models. Several metrics are considered for model comparison and performance evaluation. The metrics evaluated are the Akaike Information Criterion (AIC), the log-likelihood ratio test (LRT), and additionally by evaluating Kaplan Meier (KM) curves. We further evaluate the predictive performance against a common technique in survival analysis, Random Survival Forest (rsf), and other Cox models with varying features. We compare these using the metrics of the area under the curve (AUC), Brier, concordance index C-Index) and calibration. The results show that the resulting clustering from both approaches are discriminative w.r.t. to the outcome and moreover, that they can be used to improve prognosis.

This work follows the proposed guidelines for evaluating predictive models in a clinical context [8]. These guidelines are intended to create a standardized approach when using machine learning methods in order to have a streamlined reliable structure to assess proposed, existing, and clinically-used methods.

The rest of this thesis is organized as follows. Chapter 2 presents background and related work. Chapter 3 introduces the Supervised Scaled Clustering. Chapter 4 describes the Supervised Domain Clustering approach. Finally, Chapter 5 concludes discussing limitation and future work directions.

CHAPTER 2
BACKGROUND AND RELATED WORK

In this section we present background and related work for survival analysis and multidimensional clustering, as well as the evaluation metrics used for predictive modeling.

2.1 Survival Analysis

Survival analysis refers to the methods for analyzing data where the outcome variable is the time-to-event. How much time a subject in the dataset remains in a study depends on the outcome (i.e. the event) under consideration. When the outcome is unknown, either because the starting point is not known and/or the event has not yet occurred as of the last follow up time, the outcome is said to be censored and is denoted by a flag, the censored status. One common reason why the event has not yet occurred for a large fraction of the subjects is because they may not have remained under study before having experienced the event. If the follow up time ends before having experienced the event, the time until the event occurs (ie. event time) is said to be right-censored. To simplify the discussion throughout the text we will use censored to refer to right-censored also.

2.1.1 Nomograms

For oropharyngeal cancer outcome prediction, as in many other medical fields, nomograms are routinely used and widely accepted as support tools. Nomo-

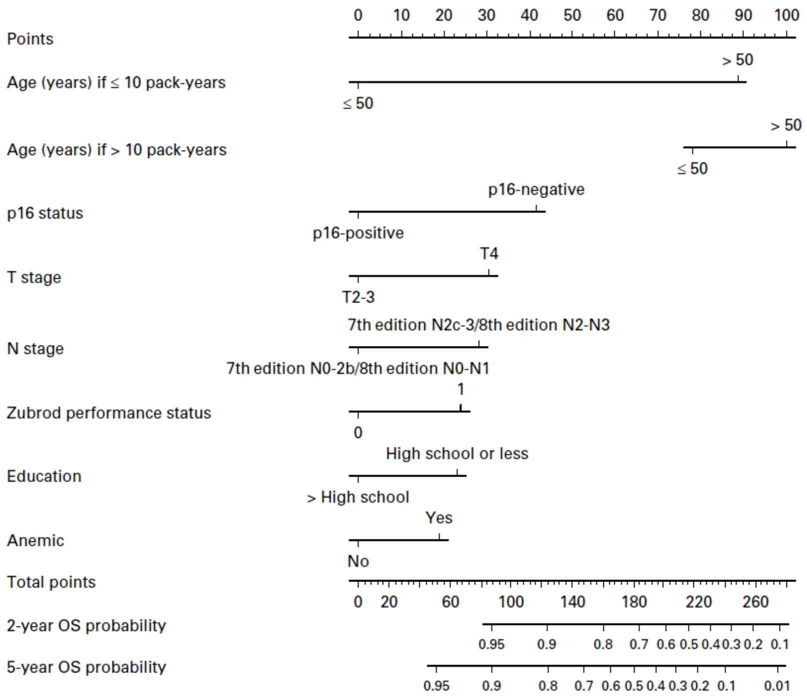


Figure 2.1: Nomogram example

grams are commonly used to estimate a survival probability for a patient of some outcome at a specific cutoff time and are derived as a visual aid from some model. Recently, online systems have been developed for clinical use to compute survival probablity based on nomograms created offline [9]. A nomogram from this work is shown in Figure 2.1. Additionally, they have created an online calculator for prognosis.

2.1.2 Survival Curves

A way to express the cumulative risk of an individual is through a survival function [10]. A widely used statistic to estimate the survival function is the Kaplan-Meier (KM) estimator. KM curves are non parametric as the procedure to

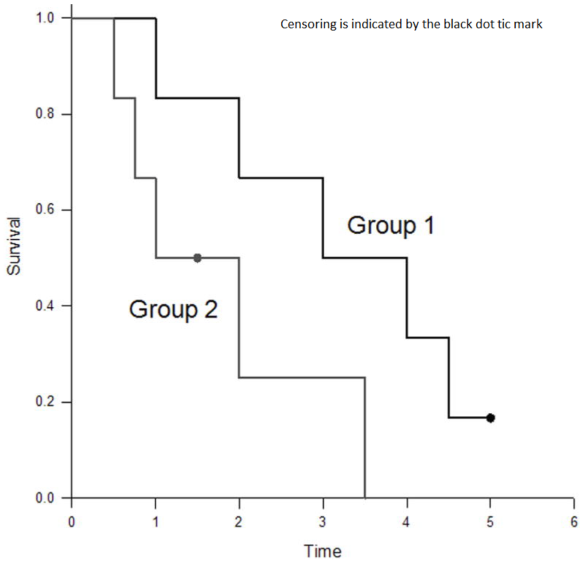


Figure 2.2: Kaplan Meier (KM) example with 2 groups. [11]

produce the curves makes no assumption about the shape of the underlying survival curve. Construction of the KM curves consist of using the event times where the event occurs (uncensored samples) such that at every point that it occurs it considers the previous survival probability and adjusts it to account for the outcomes at the event time and any censored outcomes since the last event time. Figure 2.2 shows an example of a curve with 2 groups. In this manner, the curves contain some information relating to the censored samples since these at least tell us that the event for these samples did not happen at least up to the last follow up time. KM estimators are limited in its ability to estimate survival adjusted for covariates.

Another way to express the same information but in terms of the risk's rate of change is through the hazard function. If we let S(t) and h(t), denote the survival function and hazard function respectively then their relationship can be expressed with

$$h(t) = \frac{dlog(S(t))}{d(t)}$$

Cox proportional hazards (Cox) regression models allows us to have simultaneous estimates in light of multiple covariates. Proportional hazards refers to the hazard ratio not changing over time. The Cox model can be expressed as:

$$\log h_i(t) = \alpha(t) + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}$$

i is the ith observation while *x* are the covariates and β are the corresponding coefficients. The baseline hazard function $\alpha(t)$ in this model remains unspecified. As there are no constraints on the form that the baseline hazard can take and it's a linear combination of the covariates, it is a semi-parametric model [12].

2.1.3 Martingale Residuals

Martingale Residuals are defined as follows:

$$M_i(t) = N_i(t) - \int_0^t Y_i(s) e^{\beta' Z_i(s)} d\Lambda_o(s)$$

$N_i(t)$ indicates the number of observed events at time t for subject. $Y_i(t)$ is a 0-1 process indicating whether the i^{th} subject is a risk at time t, is a vector of regression coefficients, $Z_i(t)$ is a p dimensional vector of covariate processes, and A_o is the baseline cumulative hazard function [7]. Each subject is associated with a martingale residual regardless of it's event status.

Martingale residuals may be interpreted as a measure of excess of deaths. Moreover in [7] they explored using these residuals as their dependent and empirically assessed viability on classification and regression. Residuals are bounded between $-\infty$ and $+1$.

2.2 Multidimensional Clustering

There are three general types of machine learning algorithms: 1) supervised learning; 2) unsupervised learning and 3) reinforcement learning. They are essentially classified on the basis of desired outcome of the algorithm [13], [14]. In supervised learning algorithms, a labeled set of training data or examples is used in order to classify a categorical variable, by probability or category, or predict some continuous variable. In supervised learning there are the more traditional statistical approaches such as for example Linear Regression, Naive Bayes, and Logistic Regression, and common machine learning ones such as for example Decision Trees, Random Forest, Support Vector Machines (SVM), and artificial neural networks (ANN).

In unsupervised learning, given a set of examples where no labels are provided the goal is to find the pattern or discover groups that are similar in some respect which can serve to summarize or identify some underlying salient characteristic. Here we come across the challenge of being able to define the similarity among samples (eg. patients, points, etc.) in a meaningful manner. To this extent we should first define the type of data for the features under consideration. The measuring scale types are termed nominal, ordinal, interval and ratio. These are defined with respect to the properties of numbers that correspond to the underlying properties of the attribute. The types mentioned previously are in order of increasing number of valid operations. Nominal has a equality operation, ordinal also has valid greater or less than operation, interval also has valid addition and subtraction operations, and finally ratio also has valid multiplication and division operations. Nominal and ordinal are referred to as categorical whereas interval and ratio are considered as numeric. Further we assess the value type, discrete, or continuous. When a feature takes a finite number of values or countably infinite number of values it is considered discrete, whereas, a continuous variable has real number values. For the purpose of defining similarities, the more relevant ones are the distinction between categorical (specially nominal) and numeric as the definitions are more sensible depending on these types.

Common dissimilarities for numeric types are minkowski (eg. euclidean, manhattan), cosine, mahalonobis, etc., and as for dissimilarities for categorical types there is hamming, jaccard, etc. Euclidean and manhattan belong to the more general L_p-metric with p=2 and p=1 respectively. Hamming is well suited for categoricals where the measure is the average of the number of features that agree. Jaccard may be well suited for binary data where it is defined as 1 minus the size of intersection over the size of the union. The mahalanobis distance is well suited for multivariate normal and elliptic distributions with fixed shape but varying location [15]. Cosine similarity does not consider the magnitude of the feature vectors but rather the angle between the feature vectors of two samples. More recent work on other similarities for numeric variables is the broad area of metric learning where the goal is to ”adapt some pairwise real-valued metric function” [16] such that the features may not all contribute in equal parts.

From a broader perspective, the outline of unsupervised learning can be categorized into two groups, association, and clustering. Here we sketch some of the more relevant outline to the context of the work considered here:

- **Clustering** = {Partitional, Hierarchical}. Clustering analysis is a type of unsupervised learning, where the goal is to find meaningful and or useful groups in the data [17]. Survey of clustering algorithms can be found in [18] and of clustering in high-dimensional data in [19]. Partitional is the more specific approach of being able to partition the space into mutually disjoint set, even if probabilistically. Hierarchical likewise can partition the space at a specific point of the hierarchy but the goal is to create a hierarchical structure at varying levels of the dissimilarity used to compare the pair of samples or points.

  - **Partitional** = {Graph, Density, Relocation}. Density approaches are concerned with primarily clustering points that are close to each other. The characterizing approach for density based is DBSCAN and it relies on two parameters, the min. number of points that need to be together to be considered a cluster and another parameter that indicates how far are two points close enough such that they are even considered as part of a cluster. It also doesn't exhaustively partition the space such that some points are not assigned a particular label. Graph based clustering is an extensive area that partitions based on comparing a set of graphs or clustering nodes and edges of a graph.

  - **Relocation** = {k-Centroids, Probabilistic}. Relocation approach rely on iterative approaches, where the cluster grouping changes in each iteration until it converges. Probabilistic approach as the name implies, give probabilities of belonging to one class or the other.
    - **k-Centroids** = {k-medians,k-means,k-medoids,k-modes}. Described further in section 2.2.1.

- **Hierarchical** = {Divisive, Agglomerative}. Whereas agglomerative combines the samples into different clusters at different dissimilarity levels, until a single cluster remains, divisive starts with a single cluster and breaks them down at every level until every sample is its own cluster.

In semi-supervised learning, the goal is the same as supervised learning, except we have many unknown/unlabeled outcomes to consider. This is why methods like Cox would be in such a category.

2.2.1 K-Centroids Clustering

If we generalize k-means and express it as a k-centroids clustering [20] problem then we consider the following approach. An iterative approach to performing k-centroids is to initially set k samples as the initial cluster centers and identify them with an arbitrary label (i.e. initial “centroids”). Then the samples are associated to it’s nearest cluster as established by the dissimilarity, eg. Euclidean for k-means and Manhattan for k-medians. After each iteration the centroids for each cluster are re-computed given some method, eg. mean for k-means, median for k-median, and modes for k-modes. A similar approach is used for k-medoids except it may use many dissimilarity measures but is further constrained by requiring to select an existing sample as cluster centroids at every iteration, although this may be suitable depending on the context. Eventually for these methods the iterations converge locally and these are ultimately the cluster labels assigned to the samples. Given the total number of possible partitions (or combinations of clusters) a global optimum solution is NP-hard (eg. k-median and k-means) [21], [22], which is why the previously described iterative solutions have been proposed with convergence guarantees. While these may find only local minima, in practice these methods have proven very effective [23]. These partitioning clustering techniques are very popular, conceptually well understood, and with a solid statistical basis [21], [22], [24].

The clustering method applied in Chapter 3 is k-medians [22]. There we decided to use k-medians given that many of the features are categorical, and the use of the median over the mean (as in k-means) is more robust to outliers [22]. Further, in order to reduce the effect of the starting seeds selection and avoid local minima, we use consensus clustering [25] as further described in section 2.2.2 to run k-medians 1000 times with different seeds and use kmeans++ initialization, in order to then find consensus among the different iterations. kmeans++ initialization finds initial cluster centers that more likely to be far from each other.

2.2.2 Ensemble Clustering

Ensemble clustering has shown to be an attractive option in order to increase the robustness and effectiveness of clustering to capture any underlying structure [26]. These structures are not known beforehand, in general, and selecting appropriate parameters for the clustering algorithms that additionally may be sensitive to initialization (as mentioned with k-centroids) is not trivial. Additionally, a common problem is to identify how many groups to partition the data in. Methods like DBSCAN have two parameters that can considerably affect the structures that are captured, and moreover may not exhaustively partition the space. As can be seen, for each clustering method, there is an additional potential challenge of selecting the appropriate parameters. Ensemble clustering aims, and has been shown to successfully achieve improved results over single clusterings.

In order to provide evidence of cluster stability regardless of the starting seed such as to reduce the effect of the local minima, one approach was consensus clustering [25] which attempts to find consensus among different clustering iterations. It is an approach that is agnostic to the clustering method and can therefore be leveraged by any method that does not produce a global solution to clustering. It first restarts the iteration at different starting locations, then samples without replacement [considering the entire dataset at each iteration] in order to ultimately formulate a co-association matrix, or consensus matrix. The cells of this matrix correspond to the pairwise agreement of cluster assignments, normalized over the number of times the samples were considered. It is ultimately hierarchically clustered to provide the actual clustering labels. Visualization of the consensus matrix is the initial aid in considering the stability while providing numerical stability indices such as a cluster-consensus and item-consensus matrix. The consensus matrix is defined as:

$$\mathcal{M}(i,j) = \frac{\sum_h M^{(h)}(i,j)}{\sum_h I^{(h)}(i,j)}$$

Where h is the h'th iteration of the chosen clustering algorithm. *I* and *M* are **N** x **N** matrices. *M* is the connectivity matrix where a cell is 1 if pair (i,j) appear together, 0 otherwise. And *I* is the indicator matrix where a cell is 1 if pair (i,j) are sampled for an iteration, 0 otherwise.Hierarchical clustering [23] is then used on the consensus matrix to extract the clusters.

A stability measure for cluster consensus is defined as

$$m(k) = \frac{1}{N_k(N_k - 1)/2} \sum_{i,j \in I_k, i<j} \mathcal{M}(i,j)$$

and more specifically, the item consensus is defined as

$$m_i(k) = \frac{1}{N_k - 1\{e_i \in I_k\}} \sum_{j \in I_k, j \neq i} \mathcal{M}(i,j)$$

where k is the cluster id, e_i is the ith sample in the dataset, I_k is the set of items belonging to cluster k, and N_k refers to the number of items in cluster k. Similarly other approaches that rely on the co-association matrix are [27]–[30] and recently [31] proposes an objective function that simultaneously decreases the time complexity using spectral clustering on the co-association matrix.

Further work in ensemble clustering or diversity ensemble clustering expands on the diversity of clusterings by including a variety of clustering algorithms. Ensemble clustering varies the clustering to be ensembled in various ways such as by sampling of the data, as with the previous discussed consensus clustering, varying the construction of the consensus matrix (termed similarity matrix), and the consensus function to determine the clustering, etc.. Many options [26] like simple majority voting, hierarchical clustering, k-modes, CSPA and LCE, etc provide the consensus objectives of determining the final clustering for some k. Furthermore, methods to derive a clustering that spans over multiple k have been explored such as the diceR [32] implementation where statistical transformations on the ensemble clusters is done.

2.3 Ensemble Methods

Ensemble methods can combine multiple base classifiers in order to get improved performance such as accuracy mainly by reducing the overall variance (but may also help in reducing bias) relative to any specific base classifier [33]. These multiple base classifiers can result after multiple resamplings of the training data, subsetting of the features considered, manipulating the outcome (eg. multiple different binnings), or changing a methods parameters. A naive approach to combining the multiple prediction from these classifiers is to average the predictions. Two conditions are necessary for an ensemble method to improve over base classifiers. 1) They should be independent of each other. Intuitively if all the classifiers are the same then as expected the performance will be the same. 2) Classifiers should be better than random guessing.

In the context of ensemble methods, two common techniques are used, bagging and boosting. Bagging, also known as Bootstrap AGGregating, samples with replacement up to the size of the dataset from a uniform distribution. After training k numbers of bootstramp samples, new instances are assigned a label based on some method like majority voting. Random Forests uses this idea and additionally considers a small subset of features at each internal node for the splitting criterion.

Boosting, unlike bagging, aims to focus on misclassified samples at every resampling or iteration such that either these have a higher probability of being sampled or it biases the classifiers or models being considered to increasingly weight them. Varying implementation of boosting have focused on modifying how to weight the samples or inform the distribution for sampling, and/or how the predictions among all these classifiers are combined.

Gradient boosting is an ensemble method that incorporates multiple weak learners in a sequential manner, using the previous step's more mispredicted samples into the next weak learner (such as Decision Trees with a single internal node, also known as decision stumps). It has been shown to be a powerful tool in

ML [34]. This approach allows us to capture the relative influence of features. This relative influence is a measure of how much the features explain when they are used in the trees. They are defined as follows:

$$Influence_j = \frac{1}{M}\sum_{i=1}^{M} Influence_j(T_i)$$

Where j refers to the variable , M refers to the numbers of times this variable is considered in a tree. And $Influence_j(T_i)$ is defined as follows:

$$Influence_j(T) = \sum_{i=1}^{L-1} I^2 i1(S_i = j)$$

Where I^2 is the empirical squared improvement. S_i is the current splitting variable and L is the number of trees.

2.4 Machine Learning in Survival Analysis

Machine learning is not new to cancer research. Numerous machine learning (ML) methods have been adapted for survival analysis, prognosis, and prediction [35]–[37]. Artificial neural networks (ANNs) and decision trees (DTs) have been used in cancer detection and diagnosis for over 30 years [38]–[40] and most recently random survival forests [41]. Initially, machine learning methods were used to identify, classify, detect, or distinguish tumors and other malignancies. In other words machine learning was primarily used as an aid to cancer diagnosis and detection [42]. More recently, cancer researchers have applied machine learning towards cancer prediction and prognosis.

2.4.1 Inverse Proportional Censor Weighting (IPCW)

IPCW incorporates censored samples such that uncensored samples are weighted in a manner that reflects the amount of censored samples that they ’shadow’ (ie. that have occurred prior to event time). Furthermore, a recent approach has shown how to incorporate IPCW unto multiple modeling methods [43].

2.4.2 Clustering in Survival Analysis

Clustering approaches specific in the context of leveraging right-censored outcomes have been previously considered in the literature. In [44], in the context of a gene dataset, the outcome information is considered by computing the univariate Cox score for all potentially relevant features, and then selected the top k of them as input to a nearest shrunken centroid clustering method. This method uses the Cox score for feature selection but performs clustering using equal weights. In our case, supervised scaling provides a mean to weight the features according to a particular outcome. A weighted approach has been also proposed in [45]. In this work, univariate Cox score is assessed for each feature, the score is then ordered, and ultimately the k largest features are selected. A weighted sparse clustering maximizes a weighted between-cluster sum of squares. This work uses the censored outcome directly which makes less effective for largely censored data as the one used in this study. In [46], the area under the curve between survival curves is considered as a measure of dissimilarity. The samples are initially grouped by considering all possible combinations of the features being considered. KM curves are formed by the groupings, the area between the curves would be the measure of dissimilarity and hierarchical clustering is applied over these dissimilarity values. In this study the number of cases considered was 110k and 4 factors. Given our vastly smaller sample size and the consideration of many more feature combinations, the KM curves would need to be initially constructed with very few samples, where most would be censored, such that the curves and by extension the area between the curves would not be meaningful.

Previous methods in the context of survival time and gene expression, have proposed approaches that enable dimensionality reduction in a high dimensional space in order to ultimately improve prognosis, by first allowing dimensionality reduction to be informed to some degree by the response of interest. In Supervised Principal Components [44] (Bair's SPC) this approach first determines a univariate cox score threshold and keeps all the features above that threshold. This threshold is determined from selecting the greatest χ^2-statistic determined from cross-validation of the training set. Then PCA is done on the training to ultimately fit a Cox model on the first principal components. To predict for a test set, the principal component for the test are first obtained by the proposed method. Empirical results show that Bair's SPC is more effective in reducing dimensionality with less error in survival time prediction [47].

Another approach is a Supervised Wavelet [48] method for classification and is similar to Bair's PC in that it first determines the features that are related to the outcome by selecting the top features based on the q value of a t test. It then performs PCA and uses the components to fit an SVM classifier for the event at a time cutoff. The unappealing component of these approaches however is their difficulty in identifying a clinically salient interpretation of any underlying characteristic. It is for this reason that we do not compare against these in this initial work. That is to say, one of the key driving motivations is to be able to create or maintain interpretable categorizations while reducing the dimensionality which in turn, and as we show empirically, may and does, improve prognosis.

2.5 Evaluation Metrics

We consider the traditional measures for performance evaluation of survival prediction models [49]:

C-index. The C-index (i.e. probability of concordance) is a unitless quantitive measure of the discriminative strength of a model. The C-index is identical to the area under ROC for binary outcomes [50]. It is the proportion of evaluable predicted pairs with the right survival order over all evaluable pairs. The evaluability of the pairs is determined from the known censored status. (Censored, Censored) is not evaluable, (Censored, Uncensored) is only evaluable if censored event time is greater than the uncensored known time [51].

Calibration. Calibration index is considered an important validation. In the guidelines documentation, for example, the calibration was indicated as part of required metrics to report in any approach [8]. Moreover, it is also given in the most recent related work [9] in the form of a calibration plot. The purpose is to establish agreement between the number of individuals that are predicted with a certain probability and the actual proportion of individuals [52].

Brier. This measure serves as an indication of overall performance. It is a quadratic scoring rule that ranges from a very informative model at 0 to 0.25 for a non informative model when the probability for the event is 50% [49]. For survival probabilities we can use a weight function to account for the censored samples [52].

ROC. The Receiver Operating Characteristic (ROC) curve plots sensitivity against specificity for consecutive cutoffs of the survival probability.

Log Rank Test. The log rank test or chi-square statistic allows us to compare n KM curves. The p-value associated compares against the null hypothesis that no curve is different (the null is a chi-square distribution with n - 1 degrees of freedom). This p-value is displayed on the KM plots.

AIC [53] and AICc [54]. AIC is a unitless quantity that can be used to compare fits between different parametric models using the same data. It estimates the Kullback Leibler divergence which means lower values are better for AIC. $AIC = 2p - 2ln(\hat{L})$

AICc was used to overcome overfitting due to small sample size and its formula is given by: $AICc = AIC + \frac{2p^2+2p}{n-p-1}$

$\hat{L}$ is the model evaluated at the most likely set of parameters, n is the number of samples, and p is the number of estimated parameters.

An AIC value of +3 is roughly considered to be a better model.

Log-Likelihood Ratio Test (LRT). The ratio between the log-likelihood of the simpler model to the model with more parameters [55]. The anova.Cox [56] function was used for the test.

$$LRT = -2log_e(\frac{L_{null}(\hat{\theta})}{L_{alternative}(\hat{\theta})})$$

The test statistic approximates a chi-squared random variable with degrees of freedom equal to the difference in the number of parameters of the null vs alternative model.

Adjusted Rand Index [57]. This index measures the agreement for every pair between the labels assigned by the AJCC stage and the labels of the cluster. The adjusted refers to a correction for chance assignment.

CHAPTER 3
SUPERVISED SCALED CLUSTERING

3.1 Introduction

In this chapter the goal of the work is to identify and exploit any underlying latent characteristics that may help stratify the feature space meaningfully towards some outcome. The proposed approach combines supervised and unsupervised methods such that ultimately clustering can be used to improve prediction of our outcomes of interest in the context of right-censored oropharyngeal head and neck cancer data. Since clustering is agnostic to the outcome, we first transform our feature space in order to relate the discovery towards the outcome. To achieve this we first create a proxy dependent variable, the martingale residuals, then train a supervised model (such as linear regression) and ultimately use it's fitted feature coefficients to scale the feature space towards the outcome. We evaluate the resulting groups through model comparisons of using its group label as a feature in a Cox Proportional Hazards (Cox) model considering AIC and LRT, and additionally by evaluating KM curves. Finally, we further evaluate the predictive performance against a common technique in survival analysis, rsf, and other Cox models with varying features. We compare these using the metrics of AUC, Brier, C-Index and calibration.

To summarize, the aims of this chapter are as follows: 1) incorporate outcome information to influence cluster analysis; 2) identify discriminative clusters using patient characteristics available at the time of diagnosis and radiomic signatures; 3) use the cluster labels to stratify the patients and generate KM curves for each cluster, and compare to AJCC stage; and 4) evaluate the predictive performance of including the cluster label as a feature in a Cox model for OS and RFS outcomes.

3.2 Methods

All analyses were conducted using R version 3.4.1 (R Foundation for Statistical Computing, Vienna, Austria). All statistical tests are two-sided with statistical significance defined as a p¡ 0.05.

3.2.1 Data

The dataset consists of 644 of oropharyngeal cancer (OPC) patients who were treated at MD Anderson Cancer Center between the periods of (2005-2013). Following IRB approval, clinical features including age at diagnosis, sex, ethnicity, HPV status, smoking status and frequency, subsite within the oropharynx, T category, N category, therapeutic combination and AJCC stage (7th and 8th edition) were extracted from electronic medical records. Table 3.1 shows the demographics of patients for the clinical features and survival outcomes considered. Response variable units are given in months and the breakdown is given on Tables 3.2 for the response distribution and censored proportions.

A more detailed description of these data can be found in [58].

3.2.2 Data Preprocessing

Missing Data was imputed using the Multivariate Imputation by Chained Equations (MICE) approach [59]. This is a standard widely used approach in survival analysis and the one used here. Imputation of each validation sample was performed individually and only considering training after the training had been imputed, per fold. Features used in clustering and training with missing values (radiomic features and Smoking Packs Per Year) that were ultimately imputed used Predictive Mean Matching with $k = 5$. As we are comparing against AJCC stage, the 2 patients with missing values for it were discarded. The 2 patients with missing age were never considered as patients with missing response (2 for OS, 6 for RFS) were discarded and overlapped with the missing age.

Min-Max normalization was used to standardize each attribute's range into the interval $[0, 1]$. This was done as a pre-processing step for feature selection, model training, and clustering. This prevents features from dominating the dissimilarity value (e.g. L_p-norm) when clustering which in our case was Manhattan distance (L_1-norm).

Out of the initial 3831 radiomic features, we removed those with zero variance and those with a correlation above 99%. Previous studies have identified tumor volume and intensity as relevant features for local control [60]. To further reduce redundancy, we also removed any radiomic features that were highly correlated ($> 80\%$) to F25.ShapeVolume and F29.IntensityDirectGlobalMean. Finally the RReliefF feature selector was applied over the remaining 542 radiomic features.

The Relief family of algorithms calculate a feature importance value for each feature by calculating the distance between pairs of near observations which fall in the same and different classes [61]. Features with more similar values for observations having the same class get higher importance values and likewise features with more different values for observations not having the same class get higher importance values. RReliefF calculates feature importance based on a continuous outcome, in this case, the martingale residuals resulting from using a Cox model considering the clinical features. It achieves this by probabilistically determining whether the instances are different and is based on the relative difference between the outcomes. Feature importance for the Relief algorithms in general is expressed by the following equation:

$$\begin{aligned} W[A] = P(&\text{diff. value of A}|\text{nearest inst. from diff. class}) - \\ P(&\text{diff. value of A}|\text{nearest inst. from same class}) \end{aligned}$$

The radiomic signature of 4 features, described later in Results, obtained through this feature selection was then included together with the clinical features for clustering. Given our evaluation of using the Cox model to assess the ultimate clustering, and comparing against this model using the original features, a reduced space of the entire radiomic feature space is necessary as otherwise there would be too many parameters for the Cox model to reasonably estimate.

3.2.3 Novel Supervised Scaling for Clustering

Clustering without any considerations can certainly capture latent characteristics, but nevertheless these may not be related to the outcome of interest.

The challenge then is to incorporate the outcome information in a meaningful way that can help identify discriminative groups for a particular outcome. Previous studies have explored using residuals as the dependent variable and empirically assessed viability on classification and regression [7]. For largely censored samples, the use of residuals has the advantage that each subject would be associated with a residual regardless of it's event status. This allows us to incorporate all data available into the training process. Martingale residuals [7] in particular can be interpreted as a measure of excess of deaths.

The Supervised Scaling processing pipeline is illustrated in Figure 3.1. First, a null Cox model is trained for a particular outcome in order to obtain a proxy dependent variable, the martingale residuals (1). Then, these residuals are used to train a regression model such that the fitted coefficients are used to scale the feature space (2). This effectively produce features weights associated to the outcome. Finally, the scaled feature space is clustered using a machine learning algorithm, e.g. consensus clustering over 1000 runs of k-medians (3). Validation sample assignment of cluster labels is done by computing the Manhattan distance to the centroids of the formed clusters and assigning the label of the closest centroid. Cluster assignment per fold is arbitrary but may relate to the same underlying characteristic. Therefore, in order to visualize clusters and assess the cluster label assignment

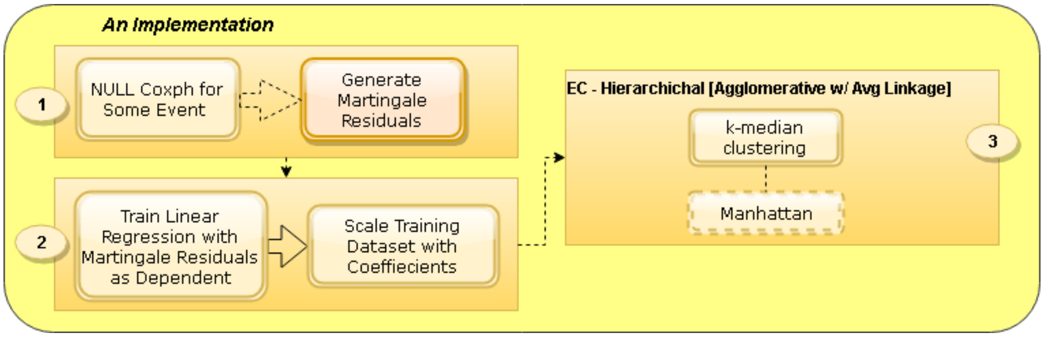

An Implementation
1
NULL Coxph for Some Event
Generate Martingale Residuals
2
Train Linear Regression with Martingale Residuals as Dependent
Scale Training Dataset with Coeffiecients
EC - Hierarchichal [Agglomerative w/ Avg Linkage]
k-median clustering
Manhattan
3

Figure 3.1: Supervised Scaled Clustering (SSC) approach. A null Cox model is trained in order to obtain a proxy dependent variable (1), e.g. martingale residuals. The fitted coefficients obtained from training a supervised learning method, e.g. linear regression, are used to scale our feature space (2). A clustering method is applied over the scaled feature space (3). The clustering implementation here shown is consensus clustering over 1k runs of the k-median (k=2) clustering method using different initial seeds and Manhattan distance as the dissimilarity measure.

across folds, clusters at every fold are matched to fold 1 (arbitrarily selected). That is, if the training labels at a fold correspond with the training labels at fold 1 more than they don't then the labels are kept the same, otherwise they are inverted. The validation samples are then assigned to these clusters. Given that the labels are arbitrary, this would just provide consistency of label assignment.

Through the remainder of this paper, scaling or scaled refers to applying these feature weights in addition to first standardizing the features with min max normalization.

Once we have clustered the data with Supervised Scaling we proceed to use these cluster labels as a feature in the prediction method.

3.2.4 Survival Models

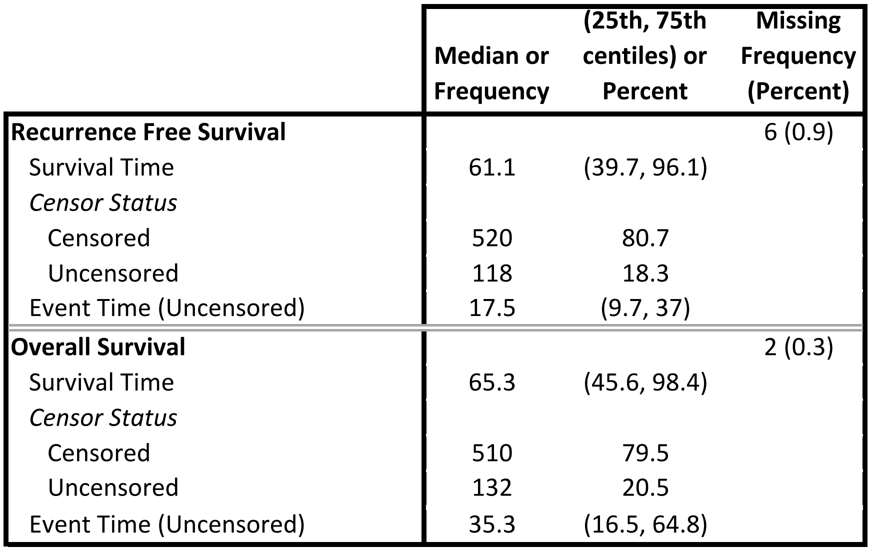
Since Cox proportional hazards (Cox) models are generally used to model survival and meaningful comparisons among models with various metrics can be made, we construct several Cox models using different features, including the cluster label where indicated, as described below.

- *AJCC Only* - Only 4 AJCC categories are considered in the model.
- *[Sc.] Cluster Only* - only the cluster label as a feature after standardizing and scaling of the feature space.
- *[Stand.] Cluster Only* - only the cluster label as a feature without scaling the feature space (only standardization).
- *Only AJCC & [Sc.] Cluster* - Only 4 AJCC categories and scaled feature space cluster labels are considered in the model.
- *Clin. Only* - only the clinical features.
- *Clin & X* - Clinical features and, in addition, what X describes (eg. *Rad.* for the 4 radiomic feature signature, *[SC.] Cluster Only* for the scaled feature space cluster labels, etc).

In addition to these Cox models, we also evaluated Random Survival Forest (rsf) implemented in the randomForestSRC(v2.7) package [62]. We grow 100 trees, choosing the default $\sqrt{p}$ of the features, where p is the number of features. These trees consider the clinical features and the radiomic signatures.

	Median or Frequency	(25th, 75th centiles) or Percent	Missing Frequency (Percent)
Female			0
No	566	87.9	
Yes	78	12.1	
Age	58	(52.3, 65.3)	2 (0.3)
HPV Status			0
Negative	50	7.8	
Positive	393	61	
Unknown	201	31.2	
T Category			0
T1,T2,Tis,Tx	410	63.7	
T3,T4	234	36.3	
N Category			0
N0, N1	341	53	
N2, N3	303	47	
Smoking Status			0
Current	139	21.6	
Former	238	37	
Never	267	41.5	
Smoking Pack Per Year (Current)	35	(20, 50)	13 (2)
Tumor Subsite			0
BOT	328	50.9	
Tonsil	259	40.2	
GPS, NOS, Soft Palate	57	8.9	
White/Caucasian			0
No	57	8.9	
Yes	587	91.1	
Therapeutic			0
CC	340	52.8	
IC_and_CC	160	24.8	
IC_and_Radiation	61	9.5	
Radiation	83	12.9	
F25.ShapeVolume	7.7	(3.8, 14.8)	86 (13.4)
F29.IntensityDirectLocalRangeMax	1136	(1103, 1195.8)	86 (13.4)
F5.IntensityDirectGlobalMax	1199	(1165, 1341.8)	86 (13.4)
F29.IntensityDirectGlobalMax	1190.5	(1152, 1369.5)	86 (13.4)
AJCC 8th			2 (0.3)
I	238	37	
II	109	16.9	
III	74	11.5	
IV	221	34.3	

Table 3.1: Characteristics of population. Following AJCC standard definitions, T1 - T4: ”Size and/or extent of the primary tumor”, Tx: ”Primary tumor cannot be evaluated”, Tis: ”Early cancer that has not spread to neighboring tissue”, and N0-N4: ”Involvement of regional lymph nodes”. BOT: Base of Tongue. NOS: Not otherwise specified. GPS: Glossopharyngeal Sulcus. CC: Concurrent Chemotherapy. IC: Induction Chemotherapy.



	Median or Frequency	(25th, 75th centiles) or Percent	Missing Frequency (Percent)
Recurrence Free Survival			6 (0.9)
Survival Time	61.1	(39.7, 96.1)	
Censor Status			
Censored	520	80.7	
Uncensored	118	18.3	
Event Time (Uncensored)	17.5	(9.7, 37)	
Overall Survival			2 (0.3)
Survival Time	65.3	(45.6, 98.4)	
Censor Status			
Censored	510	79.5	
Uncensored	132	20.5	
Event Time (Uncensored)	35.3	(16.5, 64.8)	

Table 3.2: Outcomes summary

3.3 Results

Two clusters were identified and evaluated using 10-fold cross validation for OS and RFS.

Radiomic Feature Selection

The top 4 radiomic features selected from RReliefF for both OS and also for RFS were:

- F25.ShapeVolume
- F29.IntensityDirectLocalRangeMax
- F5.IntensityDirectGlobalMax
- F29.IntensityDirectGlobalMax

Clustering with Supervised Scaling

Figure 3.2 shows the KM curves for the cluster assignments over the validation samples across folds for the OS outcome.

The KM curves for the two clusters differ significantly (p-val<0.0001). They are also significantly different (p-val< 0.01) for RFS. The demographic breakdown per cluster is given in Table 3.3 for OS and Table 3.4 for RFS. Albeit omitted for conciseness of figures and tables, for standardization only, the p-values associated to the KM curve comparison were not significant for either outcome.

Comparison with AJCC Staging System (8th edition)

We compare the KM plots for to AJCC stage against the clustering label results mentioned previously as indicated in the same Figure 3.2. To aid this comparison, Stages I and II were grouped together, likewise Stages III and IV were grouped together. The Adjusted Rand Index comparing the 2 clusters in these figures for OS vs the AJCC groupings is 0.193, and 0.104 for RFS. When comparing the cluster labels vs all the 4 stages of AJCC considering the unknown HPV, it is 0.028 for OS and 0.023 for RFS. Given that this pairwise agreement measure is low, but we know that both (1) AJCC is clinically informative and moreover (2) that the clusters have a strong discrimination on the outcome, in the model comparison we compare how adding both the label and the AJCC status affects the model.

Model Comparisons and Prediction

We compare how meaningfully the cluster labels are by quantitatively assessing them (AIC/AICc and LRT) as an additional feature in the Cox model as shown in Table 3.5. We consider the entire dataset and the cluster labels are those assigned to the validation samples at every fold.

This table compares against two reduced models. To display results in an intuitive manner, AIC and AICc values are the negated difference to these reduced models such that negative values indicate a worse model and positive values a better one relative to the reduced models. Table 3.5 compares against the these reduced models, *Vs Clinical* considers a Cox model with only the clinical features

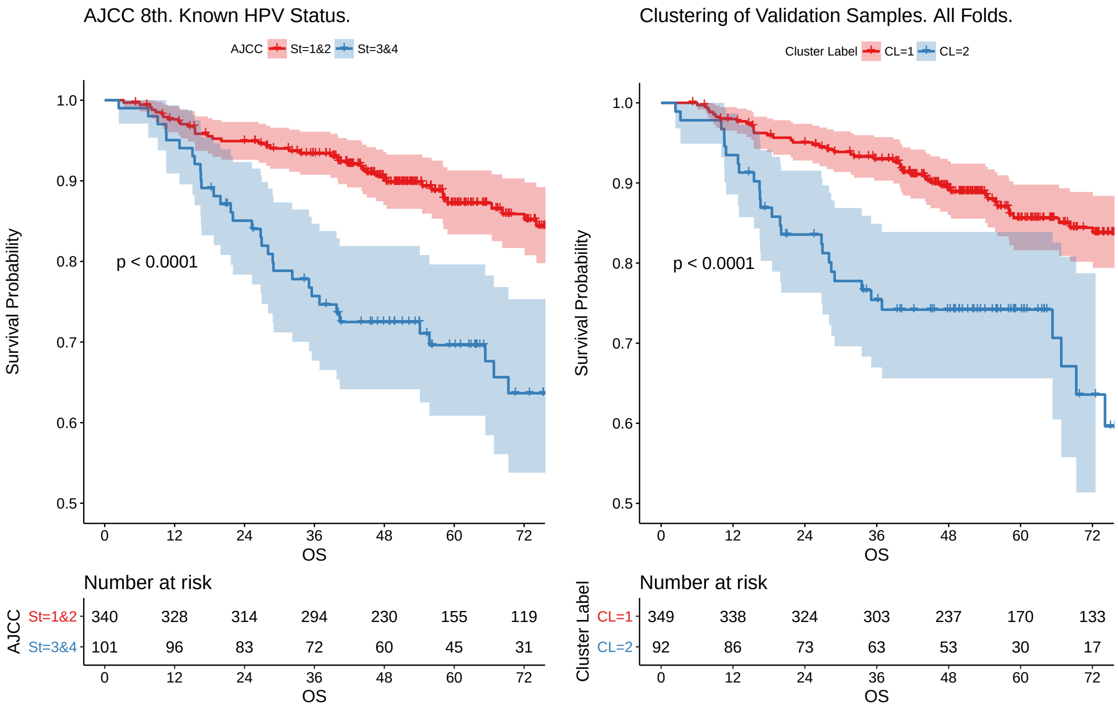


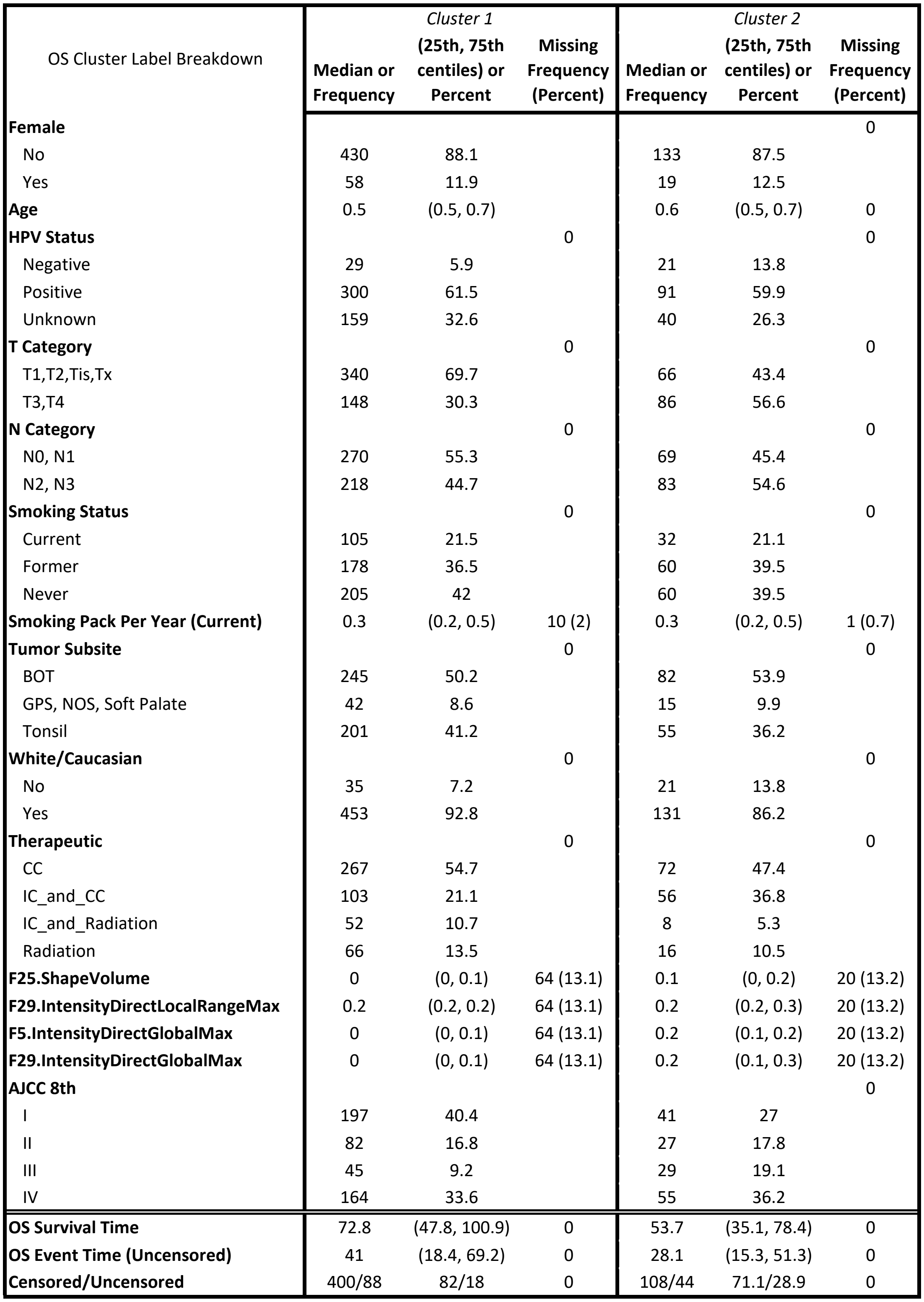
Figure 3.2: OS outcome considering known HPV Status. AJCC 8th KM curves are formed by aggregating AJCC stage categories as indicated by legend (Stage I and II vs Stage III and IV). The clustering of validation samples across folds likewise is only for known HPV Status in this comparison.

and *Vs NULL* against the null Cox model (Cox model with no covariates). As the clinical features are known features that are relevant to the diagnosis and prognosis, we consider this model as a baseline. Moreover, since we know that AJCC is a clinically relevant categorization we consider it as a feature against both reduced models and compare it against our quantitative approach to categorization.

When considering standardization ([Stand.]) only cluster, LRT and AIC indicate that these labels are not informative as features against either of the reduced models.

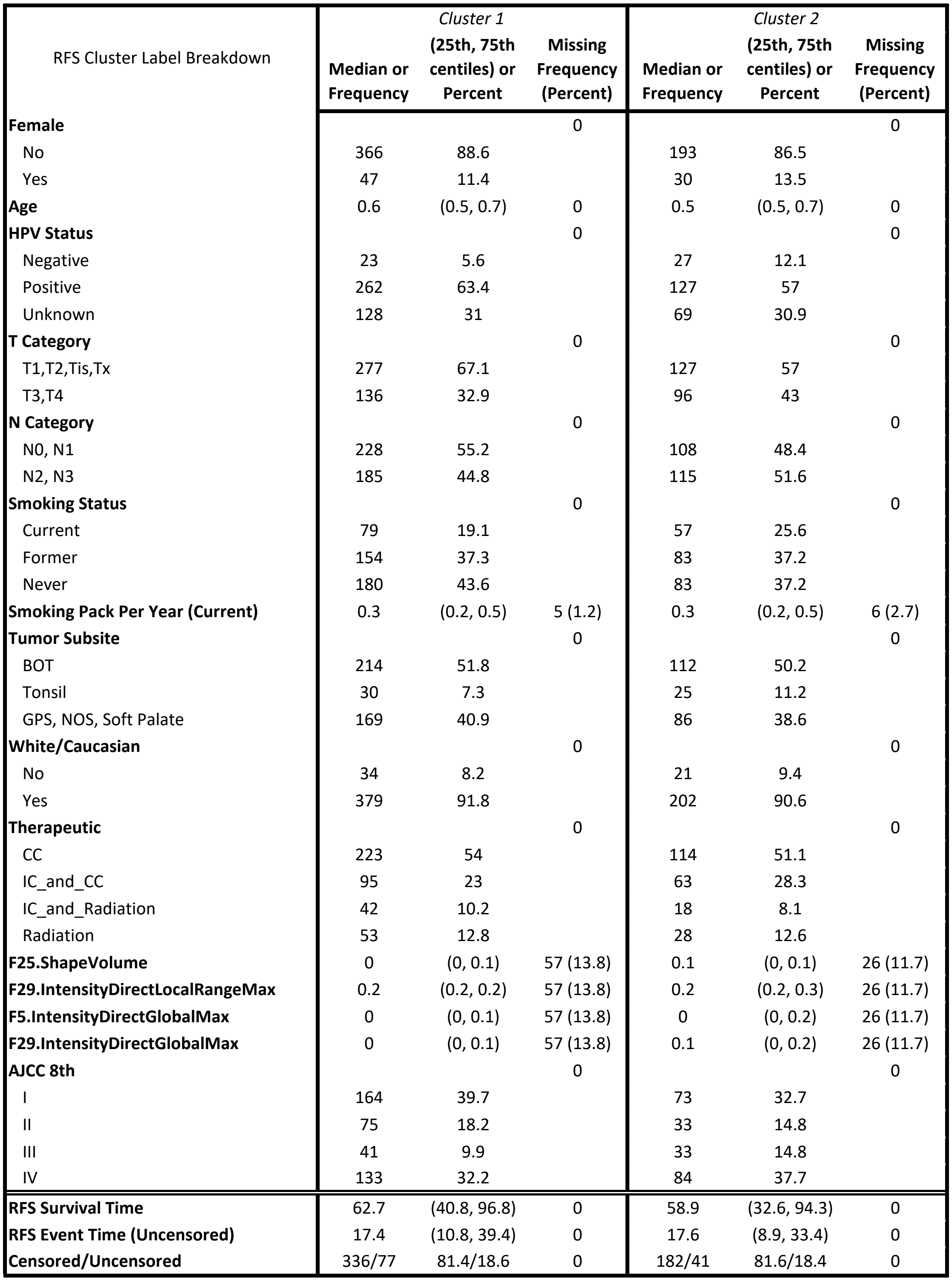
The models with overall better AICs (> 3) vs the clinical model were Clin & Rad. and the models using scaled clusters ([Sc.] Cluster) as features. This is expectedly moreso against the reduced null model. For the models with the [Sc.] Cluster as feature, the 95% CI for the estimated hazard ratio of the non reference label was [2.22,4.66] for OS and [0.30,0.66] for RFS. Similarly, when considering the clinical features and the cluster label, the interval for the cluster label was [1.64,3.64] for OS and [0.34,0.78] for RFS. All hazard ratios for clusters with standardized only are non signifcant. As expected from the fact that the AJCC labels dont match with the cluster labels yet both could be informative, when comparing against the null model we note that the inclusion of both AJCC and the [Sc.] Cluster reflects a better model with AIC rather than either [SC.] Cluster or AJCC alone. However, once we control for the clinical variables, AJCC doesnt indicate any significant improvement. Additionally, even when controlling for AJCC and Clinical, the [Sc.] Cluster feature still provides significant hazard ratios for the non reference label, which are [2.01,3.59] for OS and [0.35,0.80] for RFS.

Table 3.6 shows the main model prediction evaluation using four of metrics described in the Evaluation Metrics section. With our proposed method, when evaluating the labels as a feature in Clin.&[Sc.] Cluster, for OS we see better values for AUC, Brier and C-Index, and a well calibrated model. As for RFS, using the 4 radiomic signature features shows the better AUC, Brier and C Index despite not being as well calibrated as the other models Clin.&[Sc.]. Clusters with standardization only, as expected from AIC and LRT evaluation, considerably underperform against the radiomics or scaled clusters.



OS Cluster Label Breakdown	*Cluster 1*			*Cluster 2*		
	Median or Frequency	(25th, 75th centiles) or Percent	Missing Frequency (Percent)	Median or Frequency	(25th, 75th centiles) or Percent	Missing Frequency (Percent)
Female						0
No	430	88.1		133	87.5	
Yes	58	11.9		19	12.5	
Age	0.5	(0.5, 0.7)		0.6	(0.5, 0.7)	0
HPV Status			0			0
Negative	29	5.9		21	13.8	
Positive	300	61.5		91	59.9	
Unknown	159	32.6		40	26.3	
T Category			0			0
T1,T2,Tis,Tx	340	69.7		66	43.4	
T3,T4	148	30.3		86	56.6	
N Category			0			0
N0, N1	270	55.3		69	45.4	
N2, N3	218	44.7		83	54.6	
Smoking Status			0			0
Current	105	21.5		32	21.1	
Former	178	36.5		60	39.5	
Never	205	42		60	39.5	
Smoking Pack Per Year (Current)	0.3	(0.2, 0.5)	10 (2)	0.3	(0.2, 0.5)	1 (0.7)
Tumor Subsite			0			0
BOT	245	50.2		82	53.9	
GPS, NOS, Soft Palate	42	8.6		15	9.9	
Tonsil	201	41.2		55	36.2	
White/Caucasian			0			0
No	35	7.2		21	13.8	
Yes	453	92.8		131	86.2	
Therapeutic			0			0
CC	267	54.7		72	47.4	
IC_and_CC	103	21.1		56	36.8	
IC_and_Radiation	52	10.7		8	5.3	
Radiation	66	13.5		16	10.5	
F25.ShapeVolume	0	(0, 0.1)	64 (13.1)	0.1	(0, 0.2)	20 (13.2)
F29.IntensityDirectLocalRangeMax	0.2	(0.2, 0.2)	64 (13.1)	0.2	(0.2, 0.3)	20 (13.2)
F5.IntensityDirectGlobalMax	0	(0, 0.1)	64 (13.1)	0.2	(0.1, 0.2)	20 (13.2)
F29.IntensityDirectGlobalMax	0	(0, 0.1)	64 (13.1)	0.2	(0.1, 0.3)	20 (13.2)
AJCC 8th						0
I	197	40.4		41	27	
II	82	16.8		27	17.8	
III	45	9.2		29	19.1	
IV	164	33.6		55	36.2	
OS Survival Time	72.8	(47.8, 100.9)	0	53.7	(35.1, 78.4)	0
OS Event Time (Uncensored)	41	(18.4, 69.2)	0	28.1	(15.3, 51.3)	0
Censored/Uncensored	400/88	82/18	0	108/44	71.1/28.9	0

Table 3.3: Demographic breakdown per cluster for OS. Following AJCC standard definitions, T1 - T4: ”Size and/or extent of the primary tumor”, Tx: ”Primary tumor cannot be evaluated”, Tis: ”Early cancer that has not spread to neighboring tissue”, and N0-N4: ”Involvement of regional lymph nodes”. BOT: Base of Tongue. NOS: Not otherwise specified. GPS: Glossopharyngeal Sulcus. CC: Concurrent Chemotherapy. IC: Induction Chemotherapy.



RFS Cluster Label Breakdown	*Cluster 1*			*Cluster 2*		
	Median or Frequency	(25th, 75th centiles) or Percent	Missing Frequency (Percent)	Median or Frequency	(25th, 75th centiles) or Percent	Missing Frequency (Percent)
Female			0			0
No	366	88.6		193	86.5	
Yes	47	11.4		30	13.5	
Age	0.6	(0.5, 0.7)	0	0.5	(0.5, 0.7)	0
HPV Status			0			0
Negative	23	5.6		27	12.1	
Positive	262	63.4		127	57	
Unknown	128	31		69	30.9	
T Category			0			0
T1,T2,Tis,Tx	277	67.1		127	57	
T3,T4	136	32.9		96	43	
N Category			0			0
N0, N1	228	55.2		108	48.4	
N2, N3	185	44.8		115	51.6	
Smoking Status			0			0
Current	79	19.1		57	25.6	
Former	154	37.3		83	37.2	
Never	180	43.6		83	37.2	
Smoking Pack Per Year (Current)	0.3	(0.2, 0.5)	5 (1.2)	0.3	(0.2, 0.5)	6 (2.7)
Tumor Subsite			0			0
BOT	214	51.8		112	50.2	
Tonsil	30	7.3		25	11.2	
GPS, NOS, Soft Palate	169	40.9		86	38.6	
White/Caucasian			0			0
No	34	8.2		21	9.4	
Yes	379	91.8		202	90.6	
Therapeutic			0			0
CC	223	54		114	51.1	
IC_and_CC	95	23		63	28.3	
IC_and_Radiation	42	10.2		18	8.1	
Radiation	53	12.8		28	12.6	
F25.ShapeVolume	0	(0, 0.1)	57 (13.8)	0.1	(0, 0.1)	26 (11.7)
F29.IntensityDirectLocalRangeMax	0.2	(0.2, 0.2)	57 (13.8)	0.2	(0.2, 0.3)	26 (11.7)
F5.IntensityDirectGlobalMax	0	(0, 0.1)	57 (13.8)	0	(0, 0.2)	26 (11.7)
F29.IntensityDirectGlobalMax	0	(0, 0.1)	57 (13.8)	0.1	(0, 0.2)	26 (11.7)
AJCC 8th			0			0
I	164	39.7		73	32.7	
II	75	18.2		33	14.8	
III	41	9.9		33	14.8	
IV	133	32.2		84	37.7	
RFS Survival Time	62.7	(40.8, 96.8)	0	58.9	(32.6, 94.3)	0
RFS Event Time (Uncensored)	17.4	(10.8, 39.4)	0	17.6	(8.9, 33.4)	0
Censored/Uncensored	336/77	81.4/18.6	0	182/41	81.6/18.4	0

Table 3.4: Demographic breakdown per cluster for RFS. Following AJCC standard definitions, T1 - T4: ”Size and/or extent of the primary tumor”, Tx: ”Primary tumor cannot be evaluated”, Tis: ”Early cancer that has not spread to neighboring tissue”, and N0-N4: ”Involvement of regional lymph nodes”. BOT: Base of Tongue. NOS: Not otherwise specified. GPS: Glossopharyngeal Sulcus. CC: Concurrent Chemotherapy. IC: Induction Chemotherapy.

Vs Clinical	**OS**			**RFS**		
Model	**AIC**	**AICc**	**LRT**	**AIC**	**AICc**	**LRT**
Clin. & Rad.	+21.80	+21.35	5.36e-06	+17.82	+17.37	3.43e-05
Clin. & [Sc.] Cluster	+15.60	+15.50	2.72e-05	+7.03	+6.92	2.66e-03
Clin. & [Stand.] Cluster	+0.52	+0.42	1.12e-01	-1.88	-1.99	7.34e-01
Clin. & AJCC	-1.01	-1.34	1.73e-01	+2.05	+1.72	4.49e-02
Clin. & AJCC & [Sc.] Cluster	+13.47	+13.02	2.55e-04	+8.65	+8.19	2.26e-03
Vs NULL	**OS**			**RFS**		
[Sc.] Cluster Only	+30.69	+30.68	1.08e-08	+12.89	+12.88	1.14e-04
[Stand.] Cluster Only	+1.77	+1.76	5.22e-02	-0.19	-0.20	1.79e-01
AJCC Only	+11.48	+11.44	5.64e-04	+8.93	+8.89	1.88e-03
Only AJCC & [Sc.] Cluster	+36.54	+36.48	4.96e-09	+19.50	+19.43	1.58e-05

Table 3.5: Model comparisons of Cox models varying the features including with and without AJCC. The reduced model for *Vs Clinical* is the Cox model using clinical covs whereas for *Vs NULL* it was the null Cox model. Models were fitted on the entire dataset and the cluster labels (for the models the labels were used, ie. denoted by **Cluster**) were those assigned to the validation samples at every fold. AIC/AICc values are given relative to the reduced model as the negated difference. **[Stand.]** Refers to min max standardization only. **[Sc.]** Refers to scaling features prior to clustering.

	OS			
Method	**AUC**	**Brier**	**C-Index**	**Calibration**
Clin. Only	0.6029 ± 0.0299	0.1349	0.6616 ± 0.0254	12.11
Clin. & Rad.	0.6203 ± 0.0302	0.1325	0.6785 ± 0.0259	15.25
Clin. & [Sc.] Cluster	0.6335 ± 0.0298	0.1298	0.6851 ± 0.0252	13.80
Clin. & [Stand.] Cluster	0.6061 ± 0.0297	0.1344	0.6645 ± 0.0254	10.47
Random Surv Forest	0.6267 ± 0.0302	0.1338	0.6844 ± 0.0257	24.48
Clin. & AJCC	0.6056 ± 0.0299	0.1347	0.6643 ± 0.0256	17.00
Clin. & AJCC & [Sc.] Cluster	0.6359 ± 0.0298	0.1302	0.6881 ± 0.0254	26.15
	RFS			
Method	**AUC**	**Brier**	**C-Index**	**Calibration**
Clin. Only	0.6111 ± 0.0308	0.1378	0.6044 ± 0.0276	12.58
Clin. & Rad.	0.6639 ± 0.0302	0.1335	0.6408 ± 0.0278	25.60
Clin. & [Sc.] Cluster	0.6377 ± 0.0302	0.1354	0.617 ± 0.0274	18.39
Clin. & [Stand.] Cluster	0.6008 ± 0.0312	0.1387	0.5902 ± 0.0281	11.48
Random Surv Forest	0.6177 ± 0.0321	0.1361	0.6061 ± 0.0287	34.37
Clin. & AJCC	0.6185 ± 0.0312	0.1359	0.6103 ± 0.028	11.29
Clin. & AJCC & [Sc.] Cluster	0.6483± 0.0306	0.1340	0.6279 ± 0.0278	19.19

Table 3.6: Validation metric summary for OS and RFS outcomes. Using 10-fold cross validation. Cox model was used for all methods except Random Surv Forest. Description of methods given in the Survival Models section

3.4 Discussion

As our driving motivation is to find discriminative groups of oropharyngeal head and neck cancer patients, we evaluate the performance of the proposed approach (Supervised Scaled Clustering) in terms of the KM curves it generates, the model performance under AIC and LRT metrics, and the predictive performance in terms of AUC, C-index, Calibration, and Brier scores.

Figure 3.2 compares the KM curves for the cluster groups against the latest edition of the AJCC staging (8^{th} edition) for patients with known HPV status. As can be seen in Figure 3.2, both AJCC staging and the proposed Supervised Scaling, significantly discriminates w.r.t. to the patient's time to event. Moreover, when evaluating the predictive performance of these classification schemes, the proposed Supervised Scaling clustering method outperfoms AJCC staging. As can be seen in Table 3.5, the addition of AJCC staging has significant LRTs for all comparisons except for OS when compared to the model with clinical features. For AIC, however, including the AJCC staging only improves when compared ($-\Delta(AIC) > 3$) against the $null$ model. Compared to the Cox model with clinical features only, the scaled cluster labels have high significance in LRT for the OS outcome whereas AJCC is not significant. The AIC values for the additional scaled cluster labels over only clinical are much greater than 3, which indicates an improved model.

Additionally, given the low pairwise agreement between AJCC staging and the cluster labels (rand index < 0.2), we notice that when we include both AJCC and the scaled cluster label, the information might be complimentary. Namely, when compared to the *null* model, considering both AJCC and the scaled cluster label, the resulting AIC shows considerably better values than either AJCC staging or the cluster labels alone. However, once we control for all the clinical features, we notice that including the scaled cluster labels only also shows an improvement. These lead us to conclude that the proposed approach does indeed find a clinically meaningful categorization, complementary to AJCC staging, that can be further explored in future analyses.

As can be seen in Table 3.6, the cluster labels resulting from the proposed approach (ie. [Sc.] Cluster or scaled cluster labels) shows improved performance over AJCC staging across all metrics, except [Sc.] Cluster is only well calibrated (Calibration ¡ 15.5) for OS, whereas Clin. & AJCC is only well calibrated for RFS.

The proposed approach summarizes a high dimensional space into a single covariate. Machine learning approaches for feature selection identify a small subset of highly predictive features given an outcome variable. For these experiments, we use RReliefF and selected four radiomic features. When comparing the model performance of the scaled cluster labels to the radiomic signature, we see better AIC and LRT values for the radiomic signatures, but better values for AUC, Brier and C-Index for the scaled clustering for the OS outcome. For OS, Clin & Rad and Clin & [Sc.] are both well calibrated. These are encouraging results given the fact we performed feature selection using the whole dataset (and the outcome information) as the training set. The proposed approach generates a single covariate that represents the entire radiomic feature space avoiding the need to limit the number of selected features.

Cox proportional hazard models are widely interpretable and commonly used in the oncologic community for survival analysis. We evaluate the proposed approach when the cluster labels are incorporated into a Cox model. However this approach is potentially extensible to parametric approaches with minor modifications and could represent an additional step, albeit one not heavily investigated in the current study. The utility of a future space reduction has the added value of avoiding significant overfitting, and this also has potential applications across a wider range of machine learning style approaches which incorporate right-censored variables.

A further advantage of using the scaled clustering approach is that missing data can be handled without imputation nor removal by computing the distance between the patient and cluster centroids using the known available features. However, a thorough evaluation of missing data's effect and performance comparison with established methods for data imputation are needed.

Clustering approaches specific in the context of leveraging right-censored outcomes have been previously considered in the literature. In [44], for a gene dataset, the outcome information is considered by computing the univariate Cox score for all potentially relevant features, and then selected the top k of them as input to a nearest shrunken centroid clustering method. This method uses the Cox score for feature selection but performs clustering using equal weights. In our case, supervised scaling provides a mean to weight the features according to a particular outcome. A weighted approach has been also proposed in [45]. In this work, univariate Cox score is assessed for each feature, the score is then ordered, and ultimately the k largest features are selected. A weighted sparse clustering maximizes a weighted between-cluster sum of squares. This work uses the censored outcome directly which makes less effective for largely censored data as the one used in this study. In [46], the area under the curve between survival curves is considered as a measure of dissimilarity. The samples are initially grouped by considering all possible combinations of the features being considered. KM curves are formed by the groupings, the area between the curves would be the measure of dissimilarity and hierarchical clustering is applied over these dissimilarity values. In this study the number of cases considered was 110k and 4 factors. Given our vastly smaller sample size and the consideration of many more feature combinations, the KM curves would need to be initially constructed with very few samples, where most would be censored, such that the curves and by extension the area between the curves would not be meaningful.

For many parametric and semi parametric methods such as Cox, the amount of features that can be considered, specially given the limitation on sample size, is constrained despite the availability of increasing number of potentially relevant features. A limitation for the generalization of this study is that even after vastly reducing the feature space of potential radiomic features to four or one (the cluster label), the number of features used within the Cox model exceeds the rule of thumbs of ten events per covariate in the model.

From a clinical perspective, a limitation of the current study is the dearth of real-time collected human papillomavirus data status on historical patients with the data set; we circumvented this by incorporating the previous corresponding staging categories where there was uncertainty about HPV status. However it should be noted that this is a major etiologic feature of head and neck cancers, and necessarily meant that the robustness of our analyses which incorporated HPV data was reduced by this. We hope in future iterations to increase the size of our HPV data set, and include external validation in these larger data sets which would be of significant value. We attempted to correct for this by using a rigorous cross validation approach which we hope should demonstrate the robustness of our findings across potentially generalizable clinical scenarios. However nonetheless, as with any radiomics approach, the extensibility or generalizability of our data to other head neck cancer databases is contingent upon their similarity to the patient characteristics, treatment profiles, and demographic information contained herein.

A natural extension of our approach would be to use clustering as a way to represent other high dimensional spaces related to the outcome such as genomics and other omics spaces, and then using these labels as potentially useful features in prognosis. Other directions for future work include further evaluation to identify the attribute-values that characterize the clusters, and the evaluation of different parameters or algorithms considered in the different stages of the proposed processing pipeline. For example, the type of model fitted that can scale the feature space, the type of clustering and dissimilarity measures considered, and moreover, other ways to incorporate or leverage these discriminating clusters beyond as an additional feature used in a Cox model.

CHAPTER 4
SUPERVISED DOMAIN CLUSTERING

4.1 Introduction

In this chapter we propose a novel approach for feature extraction through clustering of multiple feature groupings. Similar to previously, we fit a Cox model for the survival outcome and then use the residuals as a dependent variable. We then identify semantically related groups of features or subspaces for the available head and neck cancer data and compute the dissimilarity between every pair of patients. Then we determine the most relevant dissimilarities using the relative influence in gradient boosting on each subspace. The patients are then clustered using the more relevant dissimilarity(ies) per subspace. The cluster labels are then used as categorical factors to fit a Cox model.

The advantage of the proposed approach in this second chapter is threefold, 1) we reduce the dimensionality, 2) we maintain clinically meaningful interpretability, and 3) we leverage these categorizations to improve prognosis.

4.2 Proposed Approach

Figure 4.1 illustrates the processing of the proposed approach. In order to achieve these locally meaningful categories we should first define the domain subspaces that are sensible for this domain (eg. Domain Subspace 1) which we have and will refer to as just subspace throughout. Defining feature groups that dont span too many features will allow the dissimilarity measures to maintain informative value. Dissimilarity measues like euclidean, manhattan, hamming etc, can be affected by the curse of dimensionality such that as $d \rightarrow \infty$ the distances between pairs are increasingly similar, which this is exacerbated when we are considering categorical features with few unique values. The previously defined domain subspaces are then further divided by the type of data, eg. nominal, ordinal or continuous. For each of these SS-datatypes we can then define, for each, a dissimilarity or set of dissimilarities that will describe how similar patients are within these groups (shown as data type 1, data type 2, etc). Most of time we are unsure what the best dissimilarity measure to use is that can properly define and ultimately be used to find the structures with clustering. With that in mind and as indicated previously our approach can use a set of possible dissimilarities for each of these SS-datatype combinations (eg. diss. 1 , diss. 2). That is you can have more than one dissimilarity, say for example correlation and angular for SS1-numeric. Once all the SS-datatype combinations are selected we proceed to create all the dissimilarity matrices, which are then normalized between 0 and 1.

At this point we would like to select for each SS, the dissimilarity(ies) that more closely approximates the change in outcome for every pair of patients. However, since the outcome is right-censored for the majority of the patients, we can not use the time-to-event directly. In order to incorporate the outcome information we create a proxy dependent variable using the martingale residual of a null (considering no features) Cox model [12] .

Applying the Cox regression model to the data, we compute a residual, and

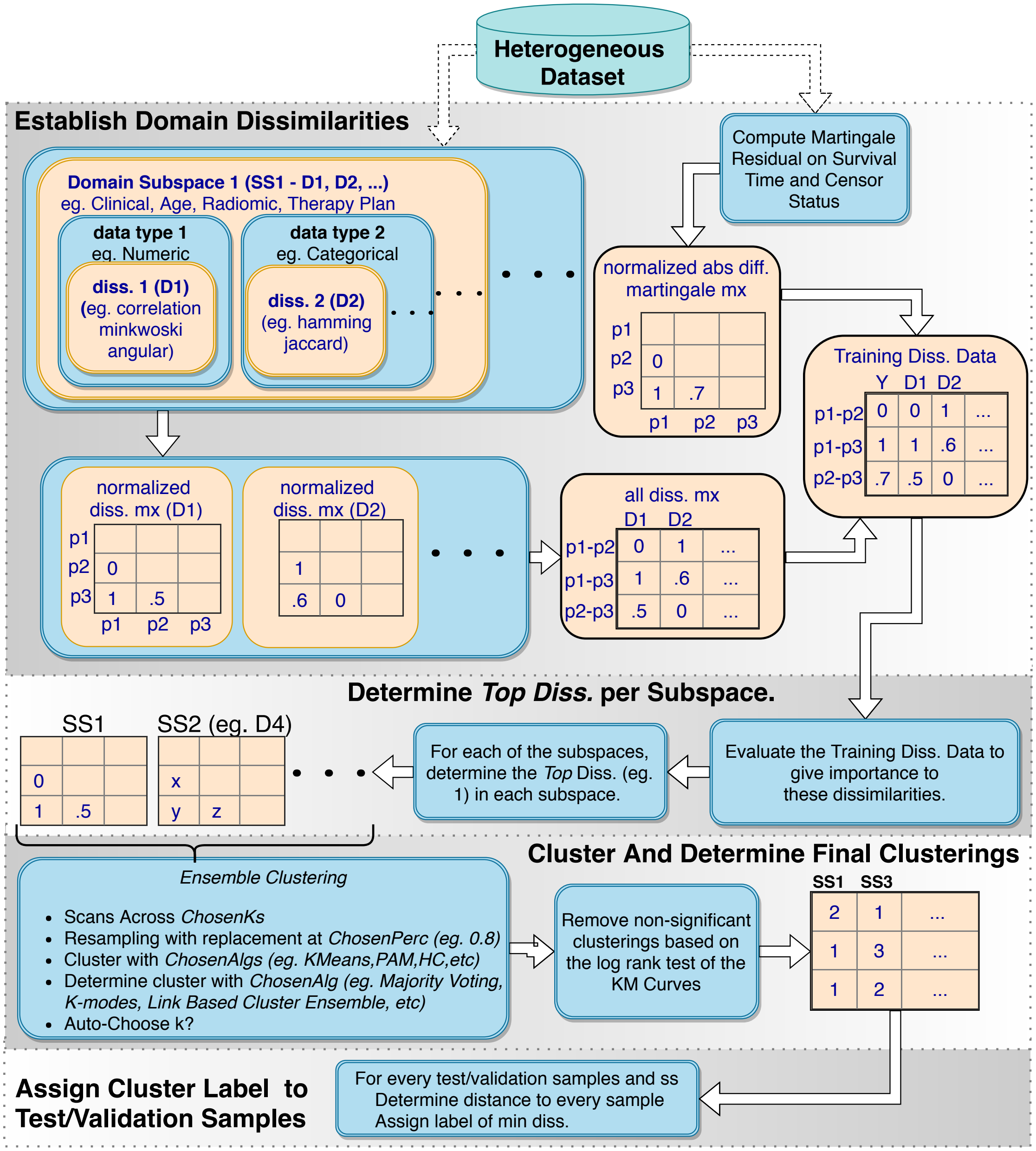

Heterogeneous Dataset
Establish Domain Dissimilarities
Compute Martingale Residual on Survival Time and Censor Status
Domain Subspace 1 (SS1 - D1, D2, ...)
eg. Clinical, Age, Radiomic, Therapy Plan
data type 1
eg. Numeric
diss. 1 (D1)
(eg. correlation minkwoski angular)
data type 2
eg. Categorical
diss. 2 (D2)
(eg. hamming jaccard)
normalized abs diff. martingale mx
p1
p2
p3
0
1
.7
Training Diss. Data
Y D1 D2
p1-p2 0 0 1 ...
p1-p3 1 1 .6 ...
p2-p3 .7 .5 0 ...
normalized diss. mx (D1)
0
1 .5
p1 p2 p3
normalized diss. mx (D2)
1
.6 0
all diss. mx
D1 D2
p1-p2 0 1 ...
p1-p3 1 .6 ...
p2-p3 .5 0 ...
Determine Top Diss. per Subspace.
SS1
SS2 (eg. D4)
0
1 .5
x
y z
For each of the subspaces, determine the Top Diss. (eg. 1) in each subspace.
Evaluate the Training Diss. Data to give importance to these dissimilarities.
Cluster And Determine Final Clusterings
Ensemble Clustering
Scans Across ChosenKs
Resampling with replacement at ChosenPerc (eg. 0.8)
Cluster with ChosenAlgs (eg. KMeans,PAM,HC,etc)
Determine cluster with ChosenAlg (eg. Majority Voting, K-modes, Link Based Cluster Ensemble, etc)
Auto-Choose k?
Remove non-significant clusterings based on the log rank test of the KM Curves
SS1 SS3
2 1 ...
1 3 ...
1 2 ...
Assign Cluster Label to Test/Validation Samples
For every test/validation samples and ss
Determine distance to every sample
Assign label of min diss.

Figure 4.1: Diagram of the proposed approach: Supervised Domain Clustering

use the abs difference between patients as a proxy for the survival outcome.

A new training matrix is then generated by transforming the n-by-n dissimilarity matrices into a n^2-element vector representing a column of the training matrix. The training matrix is a n^2-by-$(d+1)$ matrix, where n is the number of patients, and d is the number of dissimilarities considered. The extra column corresponds to the martingales residuals differences.

Second we run gradient boosting [63] considering the martingale residual column as the dependent and all other columns as independent features in order to capture the relative influence of the dissimilarities considered. In our context this influence means how much the dissimilarities can explain the martingale residual difference. With this method our goal is to be able to identify the most relevant dissimilarities w.r.t. to the outcome of interest. Then for each subspace we select the most important dissimilarities to be considered for clustering. Now that we have determined the most important dissimilarities that are related to the outcome of interest we proceed to find any salient structures via clustering.

To obtain the clusters for each subspace we incorporate the use of ensemble clustering which has shown to be more robust indentifying underlying characterists than single algorithms. In addition to clustering on subsamples we can also incorporate multiple clustering algorithms [18] to increase the diversity in the induced partitions. Eg. Partition around Medoids (PAM) is understood to be more robust to outliers than k-means as it selects real points to be the cluster prototype as it iterates to convergence and as describing the final clustering. Additionally, it is more sensible in terms of interpreting what the clustering is capturing. Likewise we can use Hierarchical Clustering with any of the common linkage options like average, single and complete. Or we can use others like Ward's which is the hierarchical analogue of k-means but without the potential problem of having merged clusters be less distant than the pair of clusters merged. Once these clusters are obtained then we prune away the clusters that do not show any discriminative capacity by assessing the Kaplan-Meier (KM) Curves. These curves incorporate the right-censored patients and are a ubiquitous survival analysis tool that allows for the visualization and comparison of risk stratification. For something to be non-significant for an arbitrary number of curves means that there is no difference among all the curves. In order to account for the multiple comparison tests being done here, one for each subspace, we adjust the p-value considering the false discovery rate. For this we use the Benjamini & Hochberg (BH) [64] method which is a common approach that has greater power than other alternatives like Bonferroni. Therefore our approach has two steps at which the outcome information is evaluated. One by using the proxy dependent martingale residual in gradient boosting, and secondly the log rank test in order to obtain the final clusterings per subspace. As implied by the missing SS2 in the Cluster and Determine Final Clusterings in the diagram, every subspace is not necessarily going to have it's own clustering if the clustering is not significant. These significant supervised subspace clusterings are then deemed to form our new feature space where each subspace corresponds to a clinically meaningful space and where the discriminative groupings can allow inspection of how risk is stratified. Motivating then potential further inquiry in the domain space. When new patients are to be assigned, we suggest using the average distance of these patients to the existing patients stratified by clusters and selecting the smallest of these average distances to assign the label.

4.3 Results and Discussion

In this section we present experimental results after applying the proposed approach to head and neck cancer patient data. We first describe the dataset and pre-processing done, then briefly summarize the metrics used for evaluation, and finally present the results. Our experiment were run in R, version 3.4.1.

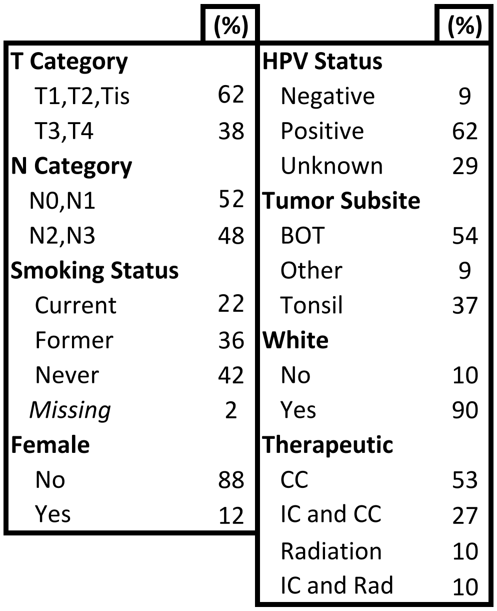
4.3.1 Dataset

The dataset consists of 525 oropharyngeal cancer (OPC) patients who were treated at MD Anderson Cancer Center between the periods of (2005-2013). Following IRB approval, clinical features including age at diagnosis, sex, ethnicity, HPV status, smoking status and frequency, subsite within the oropharynx, T category, N category, therapeutic combination and AJCC stage (7^{th} and 8^{th} edition) were extracted from electronic medical records. These clinical features are a subset of all the available clinical features and were used as they are understood in the domain to be relevant features in prognosis. We limit the number of clinical features to these 10 relevant features in head and neck cancer with the goal to compare against highly interpretable modeling strategies for prognosis. Tableřeffig:demo shows a breakdown for the categorical clinical information. For continuous data:

1. Age. mean-59.14, median-58.33, 25th-52.66, 75th-65.58

2. Packs per year for current smokers. mean-38.46, median-35, 25th-20, 75th-50

Missing Data was imputed using the Multivariate Imputation by Chained Equations (MICE) approach [59]. This is a standard widely used approach in sur-



	(%)		(%)
T Category		**HPV Status**	
T1,T2,Tis	62	Negative	9
T3,T4	38	Positive	62
N Category		Unknown	29
N0,N1	52	**Tumor Subsite**	
N2,N3	48	BOT	54
Smoking Status		Other	9
Current	22	Tonsil	37
Former	36	**White**	
Never	42	No	10
Missing	2	Yes	90
Female		**Therapeutic**	
No	88	CC	53
Yes	12	IC and CC	27
		Radiation	10
		IC and Rad	10

Table 4.1: Breakdown of categorical clinical data

vival analysis and the one used here. In our dataset only the number of smoking packs per year small fraction of number of smoking packs per year were imputed.

As part of our initial assessment of this approach and its viability, the initial radiomic space of over 3800 was preprocessed by first removing those with many missing values ($> 20\%$), and then removing any samples with any missing radiomics. Following this we removed those features with zero variance and features that were highly correlated. Finally the RReliefF feature selector was applied over the remaining over 500 radiomic features. The Relief family of algorithms calculate a feature importance value for each feature by calculating the distance between pairs of near observations which fall in the same and different classes [61]. To keep the comparisons against cox and logistic fair, we limit the number of

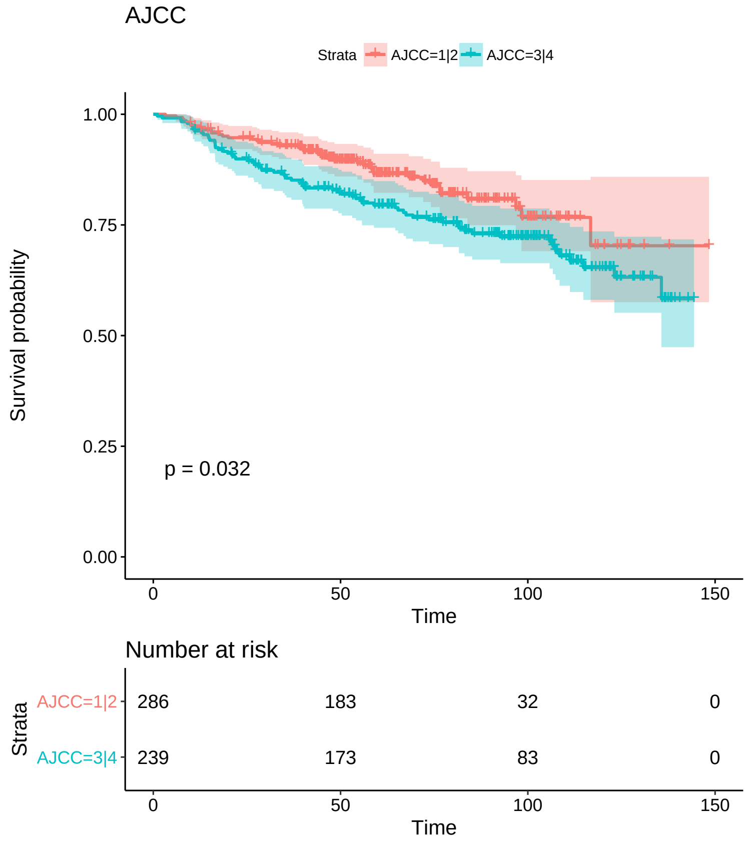


Figure 4.2: KM curves for AJCC. Grouped into 2 clusters. Stage I and II as 1|2 , and Stage III and IV as 3|4

radiomic features used for modeling.

4.3.2 Cluster Evaluation

We illustrate the applicability of the proposed approach by considering overall survival (OS). The details of the particular implementation of our approach are described and we compare the KM curves for the clusters that summarize the subspace. As a baseline of the discrimination of these curves we consider an important current clinically meaningful categorization AJCC Stage shown in Figure 4.2. We then evaluate the trained clusters with a 10-fold cross-validation over the metrics.

First we define 5 subspaces:

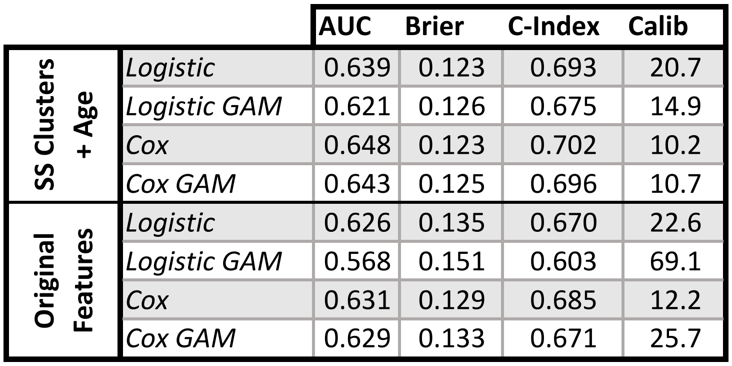
1. SS AGE: Age
2. SS SMOKING: Smoking Status (3 categories). Packs Per Year was initially continuous but binned into 3 categories of 0, Low and High using the mean of current smokers as threshold for Low and High.
3. SS STANDARD: T Category (4 categories), N Category (4 categories)
4. SS OTHER: Ethnicity (2 categories), Female (2 categories), Tumor Subsite (3 categories), Therapeutic Combination (4 categories), HPV Status (3 categories)
5. SS RADIOMICS: All Continuous.
    - F25.ShapeVolume
    - F29.IntensityDirectGlobalMean
    - F29.IntensityDirectLocalRangeMax
    - F52.NeighborIntensityDifference25Complexity

The dissimilarities considered for the categorical features were hamming and jaccard. For the ordinal features (ie. STANDARD subspace) we also considered manhattan. Whereas for the continuous, it was angular, correlation, absolute correlation, manhattan, and euclidean. Gradient boosting had selected a low (relative to the others) influence dissimilarity for the OTHER subspace with the 5 categorical features. The implementation of gbm used is here [65] with the params set for shrinkage to 0.01 and number n.trees 3000. The shrinkage value was used from an understanding that shrinkage values < 0.1 have better performance. Once all the top dissimilarities per subspace were determined the ensemble clustering performed over each of these top dissimilarities was using the implementation at [32]. Here the clustering algorithms considered were Partition around Medoids (PAM) and hierarchical clustering with the ward's minimum variance linkage method. These two methods were picked as they are distinctly different from each other. We selected 25 iterations per cluster as a recommendation of k-modes we later use [66] of no more than 50 and a scan over k's from 2 to 3 at 80% resampling. As we increased the k to scan over, larger k's ultimately selected per subspace increase the numbers of parameters to later fit. The consensus function used to create the clusters for each k was k-modes [66] due to its nature of applicability over categorical data. The final single clustering assignment from all the k's chosen in the implementation of this ensemble is computed using statistical transformations on the ensemble cluster.

Once all the clusters per subspace are computed we further evaluate that the log rank test among the curves is significant. The OTHER subspace resulted in a p-val of .54 and was therefore rejected when using the BH adjustment for multiple log rank test comparison. This suggest that this subspace cluster is not very discriminative towards the outcome and as such the clustering is removed from the final new SDC feature space. The three subspaces that were ultimately clustered are shown in Figures 4.3a, 4.4a and 4.5 which visualize the discrimination of these clusters. The p-values in these Figures reject that there is no difference among/between the curves. Here we clearly see a low and high risk group for two of the clusterings, and for the three group clustering, an additional medium risk group. Figure 4.4a results from a clustering of the continuous 4 radiomic feature signature over a euclidean dissimilarity matrix. Figure 4.3a results from a clustering of the 2 categorical smoking-related features and Figure 4.5 from a clustering of the T and N categories. For the the ones that compare two groups, we notice how these are visually even more discriminative than the standard categorization shown in Figure 4.2.

In the case that a subspace consists of a single feature such as our AGE subspace, we do not modify it or cluster it. Rather we use it directly with the subspace clusters. Therefore the new feature space resulting from these clusterings and age is the new feature space.

The AIC value against using the original features that composed the entire space was +1 (less is better for AIC). In AIC there is no hard rule for the threshold at which a model is better but 3 is a generally acceptable rule of thumb. This indicates that the model despite having a significantly reduced feature space, and even discarding 5 features entirely (from OTHER), it is not worse than using all the features. The initial feature space consisted of 14 features whereas after SDC, we only have 4 remaining (28%). The difference can be much greater if we account for the number of parameters fit specially as most of our clinicals were categoricals.



		AUC	Brier	C-Index	Calib
SS Clusters + Age	*Logistic*	0.639	0.123	0.693	20.7
	Logistic GAM	0.621	0.126	0.675	14.9
	Cox	0.648	0.123	0.702	10.2
	Cox GAM	0.643	0.125	0.696	10.7
Original Features	*Logistic*	0.626	0.135	0.670	22.6
	Logistic GAM	0.568	0.151	0.603	69.1
	Cox	0.631	0.129	0.685	12.2
	Cox GAM	0.629	0.133	0.671	25.7

Table 4.2: Prediction modeling evaluation

4.3.3 Survival Prediction

To provide further robust findings we evaluate the performance using 10-fold cross validation. The time under consideration is 60 months (needed for evaluating Brier for example). Table 4.2 shows that for the parametric logistic and logistic Generalized Additive Model (GAM), and the semi-parametric cox and cox GAM, when considering the reduced feature space of SDC, it is either relatively better or at least as good across all metrics for all 4 methods. All methods except Logistic, using a rule of thumb of 15 calibration threshold, we consider well calibrated. Logistic GAM particularly sees the most improvement for all metrics, AUC (+9.5%), Brier (+16.4%), C-Index(+12%) and Calibration (+78.4%).

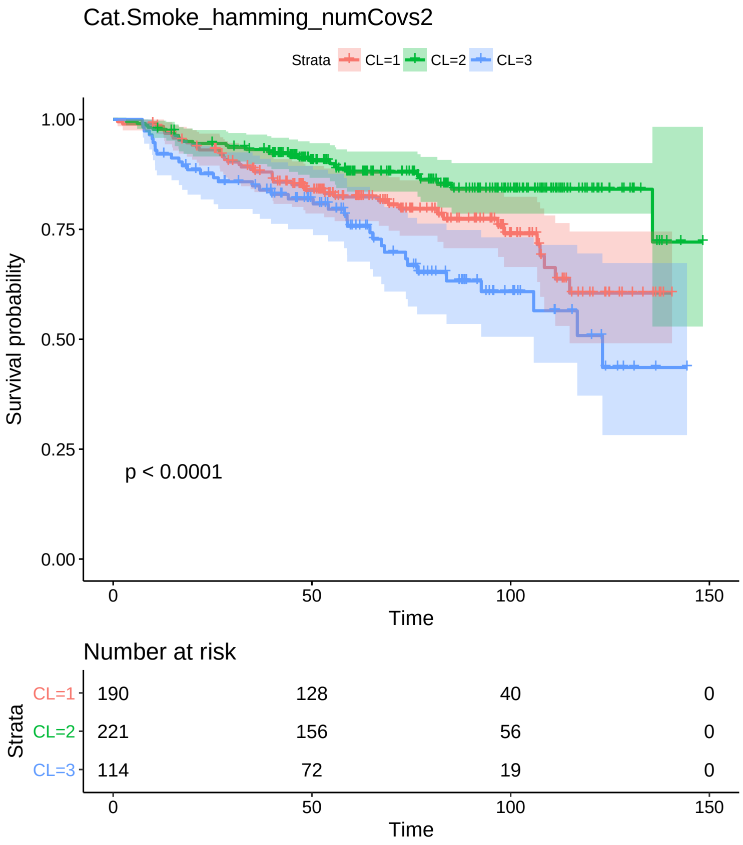


Figure 4.3: SS SMOKING clustering

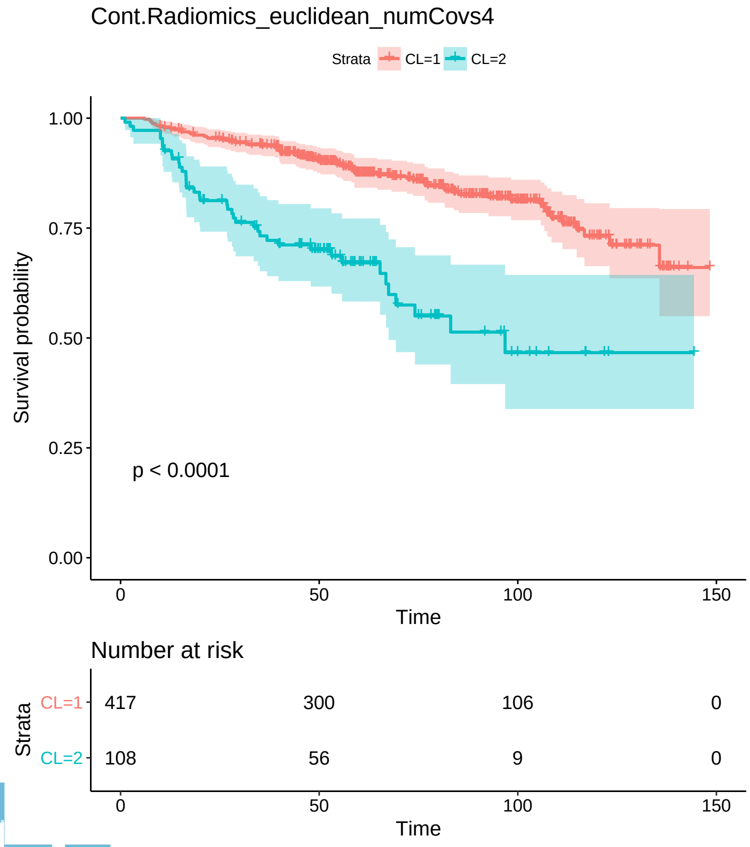


Figure 4.4: SS RADIOMICS clustering

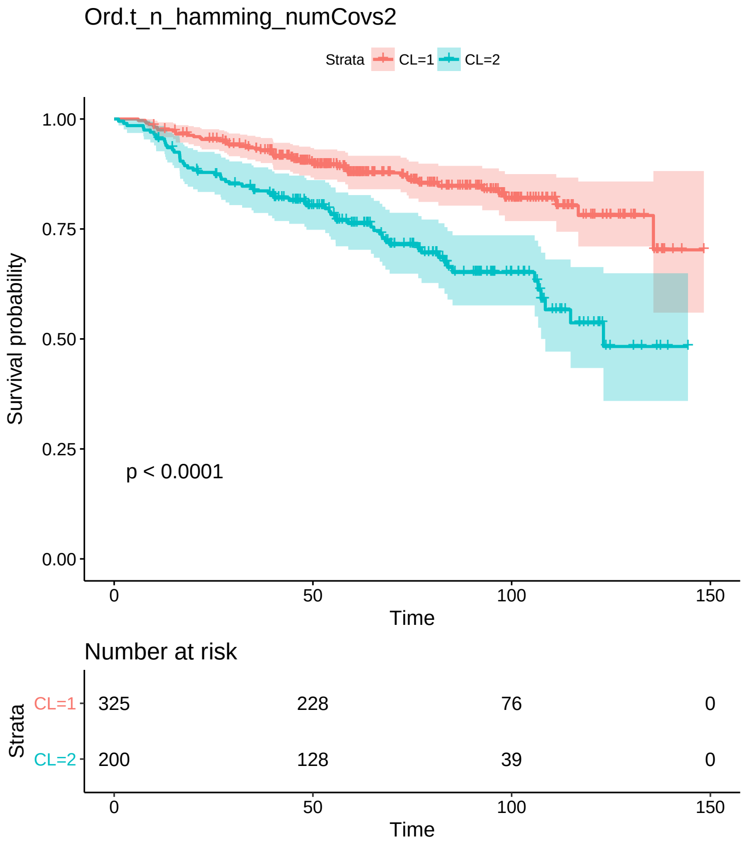


Figure 4.5: SS STANDARD clustering

CHAPTER 5
CONCLUSION

This work was able to leverage clustering in head and neck cancer to improve prognosis across various metrics for the outcomes of OS and/or RFS.

In Chapter 3 we showed how for the OS and RFS outcome we improve prognosis as indicated by the metrics assessed of AUC, Brier and C-Index. Cox proportional hazard models are widely interpretable and commonly used in the oncologic community for survival analysis. We evaluate the proposed approach when the cluster labels are incorporated into a Cox model. However this approach is potentially extensible to parametric approaches with minor modifications and could represent an additional step, albeit one not heavily investigated in the current study. The utility of a future space reduction has the added value of avoiding significant overfitting, and this also has potential applications across a wider range of machine learning style approaches which incorporate right-censored variables.

The second work, Chapter 4 proposed an approach that can indeed find categorizations within domain defined subspaces which are related to the outcome of interest. Moreover, our assessment of the metrics AUC, Bried, C-Index and Calibration indicate that not only can this approach significantly reduce the feature space, but may also improve prognosis. This can be a very attractive option in reducing the dimensionality for highly dimensional spaces while still retaining the ability of interpretability in the various domains. These groupings serve as a summary of the defined subspaces which can be readily inspected to gain further insights. Similar to Chapter 3 we understand that the method used here, with minor adjustments, such as finding or using a suitable dependent or proxy dependent variable, and a way to ensure discriminative groupings in the final clustering to be used, could be explored in other domains. The ability to define subspaces, formation of distance vectors, gradient boosting and ensemble clustering are not exclusive to this specific domain either.

A further advantage of using clustering as with the approaches presented is that missing data can be handled without imputation nor deletion by either computing the distance between the patient and cluster centroids using the known attributes as was done in Chapter 3 or the min distance to the cluster as in Chapter 4 using the known available features. However, a thorough evaluation of missing data's effect and performance comparison with established methods for data imputation are needed.

5.1 Limitations

As can be expected from the clustering approaches used here, there can be some information loss when a single feature, the cluster label, is used to represent an entire set of radiomic features, as is evident in Clin. & Rad from Table 3.6 for RFS. However despite not outperforming the raw features for RFS, the information loss can be a tradeoff in order to incorporate too high dimensional spaces.

For many parametric and semi parametric methods such as Cox, the amount of features that can be considered, specially given the limitation on sample size, is constrained despite the availability of increasing number of potentially relevant features. A limitation for the generalization of the work in Chapter 4 is that even after vastly reducing the feature space of potential radiomic features to four or one (the cluster label), the number of features used within the Cox model exceeds the rule of thumbs of ten events per covariate in the model.

From a clinical perspective, a limitation of the current study is the dearth of real-time collected human papilloma virus data status on historical patients in the data set. However it should be noted that this is a major etiologic feature of head and neck cancers, and necessarily meant that the robustness of our analyses which incorporated HPV data was reduced by this. We hope in future iterations to include external validation in larger data sets which would be of significant value. We attempted to correct for this by using a rigorous cross validation approach which we hope should demonstrate the robustness of our findings across potentially generalizable clinical scenarios. However nonetheless, as with any radiomics approach, the extensibility or generalizability of our data to other head neck cancer databases is contingent upon their similarity to the patient characteristics, treatment profiles, and demographic information contained herein.

5.2 Future Work

A natural extension of our approach would be to use clustering as a way to represent other high dimensional spaces related to the outcome such as genomics and other omics spaces, and then using these labels as potentially useful features in prognosis.

There are multiple other directions in which our work can continue. First is to enable clustering of the large radiomic space in a meaningful way such as clustering. A colleagues work, Luka Zdilar, in which I collaborated has sought the use of a proximity matrix derived from random regression forest and random survival forest where the outcomes where a martingale residuals from a cox model with clinical covariates, and the raw right-censored outcome, respectively. From each random forest, a proximity matrix is constructed such that a dissimilarity is generated on which hierarchical clustering is performed. This ultimately collapses the entire radiomic space into a single extracted feature without the use of any feature selection such that it is ultimately used in a cox model. The results there were promising specially when using random regression forest. Considering this work, extending it or following its conceptual goal, we can pursue finding novel approaches to defining meaningful similarities over high dimensional spaces such that one or few features can be extracted that can ultimately aid in improving prognosis.

Although our work here has focused on creating features used primarily on Cox or logistic, it is not the case that our clustering approach needs to be constrained as a feature extraction technique. Other ways to approach it would be to use these clusters to weight the samples used in an ensemble or use the dissimilarity measure directly, for example.

Other directions for future work of more clinical interest include further evaluation of the clusters per se to identify the attribute-values that characterize them. A systematic approach to extract what characterizes these cluster may allow for further decision support or as another hypothesis generating method.

As further evaluation or extensions to the work in Chapter 3 is to consider the ensemble used in Chapter 4 to first determine the similarities to use and modify how to validate the number of clusters used, change the type of model fitted that can scale the feature space after exhaustive evaluation of model assumptions, and vary the clustering approach considered (or a diversity ensemble) and the dissimilarity measures considered. We could also combine the approach in Chapter 4 with the scaling done in Chapter 3 where in addition to identifying the more relevant dissimilarities for each subspace, we train one or multiple models to generate the scaling factors at each subspace.

Although in Chapter 4 we aimed to find groupings that were significantly discriminative w.r.t. the outcome such that they could be readily inspected, evaluation of confounding effects from locally non-significant groupings that may improve the ultimate prognosis would be a valuable endeavor.

A limiting factor in obtaining results was certainly the time complexity bottleneck of ensemble clustering. There are many different variations that can be explored (the similarity matrix, or consensus matrix, the consensus function, etc) and/or a variety of clustering with various parameters that within an ensemble could improve the resulting clustering. Considering distributed programming for the implementation of these and similar approaches, straightforwardly by the ubiquitous MPI or more specialized frameworks such as Spark may be able to considerably cut back on the time taken to execute the results.

REFERENCES

[1] Lola Rahib, Benjamin D Smith, Rhonda Aizenberg, Allison B Rosenzweig, Julie M Fleshman, and Lynn M Matrisian. Projecting cancer incidence and deaths to 2030: the unexpected burden of thyroid, liver, and pancreas cancers in the united states. *Cancer research*, 74(11):2913–2921, 2014.

[2] The American Joint Committee on Cancers. Cancer staging system. https://cancerstaging.org/references-tools/Pages/What-is-Cancer-Staging.aspx. Online; accessed Sept 2017.

[3] G Castellano, L Bonilha, LM Li, and F Cendes. Texture analysis of medical images. *Clinical radiology*, 59(12):1061–1069, 2004.

[4] Chintan Parmar, Patrick Grossmann, Derek Rietveld, Michelle M Rietbergen, Philippe Lambin, and Hugo JWL Aerts. Radiomic machine-learning classifiers for prognostic biomarkers of head and neck cancer. *Frontiers in oncology*, 5, 2015.

[5] Stefan Leger, Alex Zwanenburg, Karoline Pilz, Fabian Lohaus, Annett Linge, Klaus Zöphel, Jörg Kotzerke, Andreas Schreiber, Inge Tinhofer, Volker Budach, et al. A comparative study of machine learning methods for time-to-event survival data for radiomics risk modelling. *Scientific Reports*, 7(1):13206, 2017.

[6] Martin Vallières, Emily Kay-Rivest, Léo Jean Perrin, Xavier Liem, Christophe Furstoss, Hugo JWL Aerts, Nader Khaouam, Phuc Felix Nguyen-Tan, Chang-Shu Wang, Khalil Sultanem, et al. Radiomics strategies for risk assessment of tumour failure in head-and-neck cancer. *arXiv preprint arXiv:1703.08516*, 2017.

[7] Terry M. Therneau, Patricia M. Grambsch, and Thomas R. Fleming. Martingale-based residuals for survival models. *Biometrika*, 77(1):147–160, 1990.

[8] Wei Luo, Dinh Phung, Truyen Tran, Sunil Gupta, Santu Rana, Chandan Karmakar, Alistair Shilton, John Yearwood, Nevenka Dimitrova, Tu Bao Ho, et al. Guidelines for developing and reporting machine learning predictive models in biomedical research: A multidisciplinary view. *Journal of medical Internet research*, 18(12), 2016.

[9] Carole Fakhry, Qiang Zhang, Phuc Felix Nguyen-Tân, David I. Rosenthal, Randal S. Weber, Louise Lambert, Andy M. TrottiIII, William L. Barrett, Wade L. Thorstad, Christopher U. Jones, Sue S. Yom, Stuart J. Wong, John A. Ridge, Shyam S.D. Rao, James A. Bonner, Eric Vigneault, David Raben, Mahesh R. Kudrimoti, Jonathan Harris, Quynh-Thu Le, and Maura L. Gillison. Development and validation of nomograms predictive of overall and progression-free survival in patients with oropharyngeal cancer. *Journal of Clinical Oncology*, 0(0):JCO.2016.72.0748, 0. PMID: 28777690.

[10] Kate Bull and David J Spiegelhalter. Tutorial in biostatistics survival analysis in observational studies 1997. *Statistics in medicine*, 16(9):1041–1074, 1997.

[11] Jason T Rich, J Gail Neely, Randal C Paniello, Courtney CJ Voelker, Brian Nussenbaum, and Eric W Wang. A practical guide to understanding kaplan-meier curves. *Otolaryngology-Head and Neck Surgery*, 143(3):331–336, 2010.

[12] John Fox. Cox proportional-hazards regression for survival data. *An R and S-PLUS companion to applied regression*, 2002, 2002.

[13] Tom M Mitchell. Machine learning. 1997. *Burr Ridge, IL: McGraw Hill*, 45(37):870–877, 1997.

[14] Richard O Duda, Peter E Hart, and David G Stork. Pattern classification. 2nd. *Edition. New York*, page 55, 2001.

[15] Ann FS Mitchell and Wojtek J Krzanowski. The mahalanobis distance and elliptic distributions. *Biometrika*, 72(2):464–467, 1985.

[16] Aurélien Bellet, Amaury Habrard, and Marc Sebban. A survey on metric learning for feature vectors and structured data. *arXiv preprint arXiv:1306.6709*, 2013.

[17] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. *Introduction to Data Mining, (First Edition)*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 2005.

[18] Rui Xu and Donald Wunsch. Survey of clustering algorithms. *IEEE Transactions on neural networks*, 16(3):645–678, 2005.

[19] Hans-Peter Kriegel, Peer Kröger, and Arthur Zimek. Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 3(1):1, 2009.

[20] Friedrich Leisch. A toolbox for k-centroids cluster analysis. *Comput. Stat. Data Anal.*, 51(2):526–544, November 2006.

[21] David Arthur and Sergei Vassilvitskii. K-means++: The advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '07, pages 1027–1035, Philadelphia, PA, USA, 2007. Society for Industrial and Applied Mathematics.

[22] Paul S Bradley, Olvi L Mangasarian, and W Nick Street. Clustering via concave minimization. In *Advances in neural information processing systems*, pages 368–374, 1997.

[23] Anil K. Jain and Richard C. Dubes. *Algorithms for Clustering Data*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1988.

[24] Pavel Berkhin et al. A survey of clustering data mining techniques. *Grouping multidimensional data*, 25:71, 2006.

[25] Stefano Monti, Pablo Tamayo, Jill Mesirov, and Todd Golub. Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Machine learning*, 52(1):91–118, 2003.

[26] Tossapon Boongoen and Natthakan Iam-On. Cluster ensembles: A survey of approaches with recent extensions and applications. *Computer Science Review*, 28:1–25, 2018.

[27] Xi Wang, Chunyu Yang, and Jie Zhou. Clustering aggregation by probability accumulation. *Pattern Recognition*, 42(5):668–675, 2009.

[28] Fei Wang, Xin Wang, and Tao Li. Generalized cluster aggregation. In *IJCAI*, pages 1279–1284, 2009.

[29] Ana LN Fred and Anil K Jain. Combining multiple clusterings using evidence accumulation. *IEEE transactions on pattern analysis and machine intelligence*, 27(6):835–850, 2005.

[30] André Lourencco, Samuel Rota Bulò, Nicola Rebagliati, Ana LN Fred, Mário AT Figueiredo, and Marcello Pelillo. Probabilistic consensus clustering using evidence accumulation. *Machine Learning*, 98(1-2):331–357, 2015.

[31] Hongfu Liu, Junjie Wu, Tongliang Liu, Dacheng Tao, and Yun Fu. Spectral ensemble clustering via weighted k-means: Theoretical and practical evidence. *IEEE Transactions on Knowledge and Data Engineering*, 29(5):1129–1143, 2017.

[32] Derek S Chiu and Aline Talhouk. dicer: an r package for class discovery using an ensemble driven approach. *BMC bioinformatics*, 19(1):11, 2018.

[33] Pang-Ning Tan et al. *Introduction to Data Mining, (Second Edition)*. 2019.

[34] Alexey Natekin and Alois Knoll. Gradient boosting machines, a tutorial. *Frontiers in neurorobotics*, 7:21, 2013.

[35] Joseph A Cruz and David S Wishart. Applications of machine learning in cancer prediction and prognosis. *Cancer informatics*, 2:59, 2006.

[36] Konstantina Kourou, Themis P Exarchos, Konstantinos P Exarchos, Michalis V Karamouzis, and Dimitrios I Fotiadis. Machine learning applications in cancer prognosis and prediction. *Computational and structural biotechnology journal*, 13:8–17, 2015.

[37] Faisal M Khan and Valentina Bayer Zubek. Support vector regression for censored data (svrc): a novel tool for survival analysis. In *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*, pages 863–868. IEEE, 2008.

[38] R John Simes. Treatment selection for cancer patients: application of statistical decision theory to the treatment of advanced ovarian cancer. *Journal of chronic diseases*, 38(2):171–186, 1985.

[39] Philip S Maclin, Jack Dempsey, Jay Brooks, and John Rand. Using neural networks to diagnose cancer. *Journal of medical systems*, 15(1):11–19, 1991.

[40] DV Cicchetti. Neural networks and diagnosis in the clinical laboratory: state of the art. *Clinical chemistry*, 38(1):9–10, 1992.

[41] Hemant Ishwaran, Udaya B Kogalur, Eugene H Blackstone, and Michael S Lauer. Random survival forests. *The annals of applied statistics*, pages 841–860, 2008.

[42] John F Mccarthy, Kenneth A Marx, Patrick E Hoffman, Alexander G Gee, Philip O'neil, M L Ujwal, and John Hotchkiss. Applications of machine learning and high-dimensional visualization in cancer detection, diagnosis, and management. *Annals of the New York Academy of Sciences*, 1020(1):239–262, 2004.

[43] David M Vock, Julian Wolfson, Sunayan Bandyopadhyay, Gediminas Adomavicius, Paul E Johnson, Gabriela Vazquez-Benitez, and Patrick J O'Connor. Adapting machine learning techniques to censored time-to-event health record data: A general-purpose approach using inverse probability of censoring weighting. *Journal of biomedical informatics*, 61:119–131, 2016.

[44] Eric Bair and Robert Tibshirani. Semi-supervised methods to predict patient survival from gene expression data. *PLoS biology*, 2(4):e108, 2004.

[45] Sheila Gaynor and Eric Bair. Identification of relevant subtypes via preweighted sparse clustering. *Computational Statistics & Data Analysis*, 116:139–154, 2017.

[46] Dechang Chen, Huan Wang, Donald E Henson, Li Sheng, Matthew T Hueman, and Arnold M Schwartz. Clustering cancer data by areas between survival curves. In *Connected Health: Applications, Systems and Engineering Technologies (CHASE), 2016 IEEE First International Conference on*, pages 61–66. IEEE, 2016.

[47] Elnaz Barshan, Ali Ghodsi, Zohreh Azimifar, and Mansoor Zolghadri Jahromi. Supervised principal component analysis: Visualization, classification and regression on subspaces and submanifolds. *Pattern Recognition*, 44(7):1357–1371, 2011.

[48] Maryam Farhadian, Hossein Mahjub, Jalal Poorolajal, Abbas Moghimbeigi, and Muharram Mansoorizadeh. Predicting 5-year survival status of patients with breast cancer based on supervised wavelet method. *Osong public health and research perspectives*, 5(6):324–332, 2014.

[49] Ewout W Steyerberg, Andrew J Vickers, Nancy R Cook, Thomas Gerds, Mithat Gonen, Nancy Obuchowski, Michael J Pencina, and Michael W Kattan. Assessing the performance of prediction models: a framework for some traditional and novel measures. *Epidemiology (Cambridge, Mass.)*, 21(1):128, 2010.

[50] Frank Harrell. *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*. Springer, 2015.

[51] Walter K Kremers. Concordance for survival time data: fixed and time-dependent covariates and possible ties in predictor and time. *Mayo Foundation*, 2007.

[52] Thomas A Gerds and Martin Schumacher. Consistent estimation of the expected brier score in general survival models with right-censored event times. *Biometrical Journal*, 48(6):1029–1040, 2006.

[53] Hirotugu Akaike. A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723, 1974.

[54] Clifford M. Hurvich and Chih-Ling Tsai. Regression and time series model selection in small samples. *Biometrika*, 76(2):297–307, 1989.

[55] Hamparsum Bozdogan. Model selection and akaike's information criterion (aic): The general theory and its analytical extensions. *Psychometrika*, 52(3):345–370, Sep 1987.

[56] Terry M Therneau and Thomas Lumley. Package 'survival'. *R Top Doc*, 128, 2015.

[57] Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of Classification*, 2(1):193–218, Dec 1985.

[58] Hesham Elhalawani, Abdallah SR Mohamed, Aubrey L White, James Zafereo, Andrew J Wong, Joel E Berends, Shady AboHashem, Bowman Williams, Jeremy M Aymard, Aasheesh Kanwar, et al. Matched computed tomography segmentation and demographic data for oropharyngeal cancer radiomics challenges. *Scientific data*, 4:170077, 2017.

[59] Stef van Buuren and Catharina Gerarda Maria Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of statistical software*, 45(3), 2011. Open Access.

[60] MD Anderson Cancer Center Head, Neck Quantitative Imaging Working Group, et al. Investigation of radiomic signatures for local recurrence using primary tumor texture analysis in oropharyngeal head and neck cancer patients. *Scientific reports*, 8, 2018.

[61] Marko Robnik-vSikonja and Igor Kononenko. Theoretical and empirical analysis of relieff and rrelieff. *Machine learning*, 53(1-2):23–69, 2003.

[62] Hemant Ishwaran, Udaya B Kogalur, and Maintainer Udaya B Kogalur. Package 'randomforestsrc'. 2018.

[63] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[64] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, pages 289–300, 1995.

[65] Greg Ridgeway, Maintainer Harry Southworth, and Suggests RUnit. Package 'gbm'. *Viitattu*, 10(2013):40, 2013.

[66] Xue Li, Osmar R Zaïane, and Zhanhuai Li. Advanced data mining and applications. In *Proceedings of Second International Conference, ADMA*, pages 14–16. Springer, 2006.