Spring 2018

Evaluating the effect of right-censored endpoint transformation for dimensionality reduction of radiomic features of oropharyngeal cancer patients

Luka Zdilar
University of Iowa

Recommended Citation

Zdilar, Luka. "Evaluating the effect of right-censored endpoint transformation for dimensionality reduction of radiomic features of oropharyngeal cancer patients." MS (Master of Science) thesis, University of Iowa, 2018.
https://doi.org/10.17077/etd.zp4ollz2

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

Evaluating The Effect of Right-Censored Endpoint Transformation for Dimensionality Reduction of Radiomic Features of Oropharyngeal Cancer Patients

by

Luka Zdilar

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

May 2018

Thesis Supervisor: Assistant Professor Guadalupe Canahuate

Graduate College
The University of Iowa
Iowa City, Iowa

CERTIFICATE OF APPROVAL

MASTER’S THESIS

This is to certify that the Master’s thesis of

Luka Zdilar

has been approved by the Examining Committee for
the thesis requirement for the Master of Science degree
in Electrical and Computer Engineering at the May 2018 graduation.

Thesis Committee: ______________________________
Guadalupe Canahuate, Thesis Supervisor

____
Er-Wei Bai

____
Hans J. Johnson

ABSTRACT

Radiomics is the process of extracting quantitative features from tomographic images (computed tomography [CT], magnetic resonance [MR], or positron emission tomography [PET] images). Thousands of features can be extracted via quantitative image analyses based on intensity, shape, size or volume, and texture. These radiomic features can then be used in combination with demographic, disease, and treatment indicators to increase precision in diagnosis, assessment of prognosis, and prediction of therapy response.

However, for models to be effective and the analysis to be statistically sound, it is necessary to reduce the dimensionality of the data through feature selection or feature extraction. Supervised dimensionality reduction methods identify the most relevant features given a label or outcome such as overall survival (OS) or relapse-free survival (RFS) after treatment. For survival data, outcomes are represented using two variables: time-to-event and a censor flag. Patients that have not yet experienced an event are censored and their time-to-event is their follow up time. This research evaluates the effect of transforming a right-censored outcome into binary, continuous, and censored aware representations for dimensionality reduction of radiomic features to predict overall survival (OS) and relapse-free survival (RFS) of oropharyngeal cancer patients. Both feature selection and feature extraction are considered in this work.

For feature selection, eight different methods were applied using a binary outcome indicating event occurrence prior to median follow-up time, a continuous outcome using the Martingale residuals from a proportional hazards model, and the raw right-censored time-to-event outcome. For feature extraction, a single covariate was extracted after clustering the patients according to radiomics data. Three different clustering techniques were applied using the same continuous outcome and raw right-censored outcome.

The radiomic signatures are then combined with clinical variables for risk prediction. Three metrics for accuracy and calibration were used to evaluate the performance of five predictive models and an ensemble of the models. Analyses were performed across 529 patients and over 3800 radiomic features. The data was preprocessed to remove redundant and low variance features prior to either selection or clustering. The results show that including a radiomic signature or radiomic cluster label predicts better than using only clinical data. Randomly generating signatures or generating signatures without considering an outcome results in poor calibration scores. Random forest feature selectors with the continuous and right-censored outcomes give the best predictive scores for OS and RFS in terms of feature selection while hierarchical clustering for feature extraction gives similarly predictive scores with compact representation of the radiomic feature space.

PUBLIC ABSTRACT

Radiomics is the process of extracting quantitative features to characterize cancer tumors from medical images. These radiomic features hold the promise to increase the precision of diagnosis, assessment of prognosis, and prediction of therapy response. Numerous machine learning algorithms exist that can be used for prediction and classification. Typically, predictive models are built over a training set where the outcomes are known and then applied over previously unseen cases. These models can be also adapted to predict survival of head and neck cancer patients and whether they will experience a relapse and when. The first challenge is to identify or extract a set of relevant features to train the models. The second is to handle the incompleteness of the data, as most patients may have not yet experienced a death or relapse event which limits the instances to base predictions on.

This work evaluates the effect of transforming these time-to-event outcomes for consolidating the high-dimensional data into a more concise, useful, and interpretable form (dimensionality reduction). Dimensionality reduction of radiomic features can later improve outcome prediction for oropharyngeal cancer patients. Two approaches are explored. The first is feature selection, which is selecting a subset of the most important features. The second is feature extraction, which transforms the features into a new representation. In both cases, the use of a radiomic signature alongside clinical data (e.g. age, sex, and smoking status) improves the results when compared to predicting with only clinical data.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1 Introduction

This paper explores the use of radiomics data in predicting survival rates of head and neck cancer patients. Head and neck cancers account for approximately 4% of all cancers in the United States [1]. Radiomics entails extraction of quantitative imaging features from computed tomography (CT), magnetic resonance imaging (MRI), or positron-emission tomography (PET) images. A large number of radiomic features can be extracted from these images to characterize tumor intensity, shape, and texture. Dimensionality reduction can greatly reduce the number of features which represent the high-dimensional radiomic space. Dimensionality reduction algorithms can be grouped into two forms, namely feature selection and feature extraction. Feature selection identifies tumor signature profiles that can be used for prognostic or predictive evaluation of patient outcomes [2], and have been putatively associated with clinical and survival outcomes [3, 4, 5, 6]. Feature extraction on the other hand transforms the radiomic feature space into a condensed yet explanatory representation, and can similarly be used for prediction of patient outcomes.

Several outcomes, such as overall survival (OS), local control (LC), freedom from distant metastasis (FDM), or combined outcomes such as relapse-free survival (RFS) are said to be right-censored because an individual may not have experienced the event before the end of their follow-up duration. At any given point in a study follow-up, some patients are yet without an event, but still potentially at risk for an event with further follow-up. Samples for which the outcome has not been observed at the last follow up are said to be censored. However, the majority of machine learning approaches and feature selection algorithms are built with either or both binary (1/0, yes/no) or continuous non-censored inputs. Although some methods have been developed to perform feature selection using the right-censored outcomes directly as time-to-event with a censor flag, censored time-to-event data frequently must be pre-processed depending on the machine learning (ML) and/or feature selection algorithm of choice. However, converting the algorithm to select for example, discrimination of 5-year control necessitates a decision regarding how to code patients with less than 5-years of follow-up, who are still at risk, but yet to have an event. Frequently, ad hoc approaches such as removing those subjects from the analysis or treating them as non-events is used. A third alternative is to use the Martingale residual (a continuous outcome) generated from a Cox proportional hazards model as the discriminant variable [3, 7, 8].

It is well-known that machine learning approaches non-cognizant to right-censoring lead to poorly calibrated risk prediction as they fail to accurately implement risk over the entire follow-up period (but discrimination is relatively unaffected) [9]. However, it is unclear what the effect of data pre-preprocessing approaches to handle right censoring have on algorithms for feature selection. Consequently, to interrogate the potential impact of right-censoring on machine learning assisted radiomics feature identification or extraction for longitudinal outcomes, we implemented the following specific aims, leveraging an extant tumor signature or transformation identified together with some relevant clinical features in oropharyngeal cancer (OPC) prognosis and comparing performance over several predictive models including:

1. Evaluate the performance of different dimensionality reduction algorithms using three distinct data "right-censored data preprocessing" approaches to represent the censored time-to-event survival data; binary, censor-incorporating (continuous), and censor-aware.
2. Assessment of the reduced feature set upon subsequent predictive model specification, using a library of established and novel risk prediction approaches (Cox proportional hazard, random forest, random survival forest, logistic regression, logistic-elastic net)

which have been adapted to right-censored outcomes using inverse probability of censoring weights [9].

3. Propose a novel dimensionality reduction technique that clusters the patients based on their similarity using the high-dimensional set of radiomic features and use the cluster label as a covariate for risk prediction.

Chapter 2 describes the dataset and the preprocessing of the dataset before radiomic dimensionality reduction. Chapters 3 and 4 describe the methodology and results of radiomic feature selection and feature extraction respectively. Chapter 5 concludes this thesis and includes directions for future work.

Related Work and Significance

Radiomics analysis and in particular dimensionality reduction of radiomic data has recently gained traction in a number of clinical studies [10]. Because of the high dimensionality of radiomic data, it is impractical to use directly in learning, and so a number of studies have focused on the reduction of the radiomic feature set for survival analysis. In [11], feature selection is performed using the right-censored outcomes directly as time-to-event with a censor flag with no additional processing of the outcome. While in studies like [12] censored time-to-event data is pre-processed and the outcome is transformed into a binary outcome using a predetermined time (e.g., whether or not the individual experienced local recurrence within 5 years) prior to feature selection. Feature clustering the radiomic data has also been explored in a number of studies [13, 14].

To the best of our knowledge, this is the first attempt to systematically evaluate the effect of outcome transformation for radiomics dimensionality reduction in model performance. The use of a continuous outcome derived from the Martingale residuals have been used in other imaging biomarkers studies [15] but has not been previously used in radiomic feature selection and extraction studies.

For feature extraction, we propose the use of the proximity matrix produce by random forest [16] to cluster the patients. This is different than the radiomic feature clustering previously proposed. In our case, it is the patients that are clustered using all the radiomic features and not the features. A single covariate representing the cluster label for each patient is then used to build the predictive models. Similar approaches have been applied in studies with gene expression data but not radiomics [17].

Chapter 2 Patient Data and Preprocessing

Data Source

Our institutional database was retrospectively reviewed for oropharyngeal cancer patients who were treated at MD Anderson Cancer Center between the periods of (2005-2013) following an IRB approval. Eligible patients who were diagnosed with oropharyngeal cancers that were pathologically confirmed either by a biopsy or a surgical excision and received their treatment (i.e. chemo-radiotherapy) on a curative intent were eligible. Demographical, clinical, toxicity and outcome data were collected for these patients.

For imaging data, contrast-enhanced computed tomography (CECT) at initial diagnosis, CT simulation and treatment planning (i.e. CT simulation images, planning structure sets, and dose) were retrieved to our commercially available contouring software (Velocity AI v3.0.1). The region of interests (ROIs) including the primary gross tumor volume (GTV) and the affected lymph node GTV was manually segmented by a radiation oncologist and was inspected by a second radiation oncologist. The generated ROIs and CT images were exported in the format of DICOM and DICOM-STRUCT to be used for radiomics feature extractions.

Radiomics Analysis

Each subject's head and neck contrast-enhanced CT image (CECT) was identified and individually checked. For each subject, the primary tumor volume was identified by two expert radiation oncologists to whom the relevant clinical data was not released. The gross tumor volumes (GTVs) were contoured based on the ICRU 62/83 definition of the gross tumor representing the gross demonstrable extent and location of the tumor [18]. A common ontology used to represent the primary (p) tumor volumes was GTVp. The manual segmentation of the

GTVp was done by commercial treatment planning software VelocityAI 3.0.1 software (powered by VelocityGrid). Also, the contours were done with the guidance of the findings from the physical examination, endoscopic examination, and other radiology such as magnetic resonance imaging (MRI) and positron emission tomography (PET). In case of metal artifacts that did not allow an accurate identification of the GTVp, these slices were omitted. Following that, CT images with GTVp generated were extracted in the format of digital imaging and communications and medicine (DICOMRT). Radiomics analysis was performed by the use of the freely available open source software "Imaging Biomarker Explorer" (IBEX), which was developed by the University of Texas MD Anderson Cancer Center and utilizes the Matlab platform (Mathworks Inc, Natick, VA). The CT images in the format of DICOM and the GTVp contours in the format DICOMRTSTRUCT were imported into IBEX. We extracted features that represent the intensity, shape, and texture. The categorization of these features was ranked as first, second, and higher texture features based on the applied method from pixel to pixel [19]. The intensity values (HU) and shape of the region of interest (ROI) are ranked as first order features, and they are extracted directly or by a histogram analysis before any mathematical transformation with no respect to the spatial configuration. The intensity features (entropy and variance) describe the gray level dispersion but it depends on the grey level spatial distribution precision. The second order features represent intratumoral heterogeneity with the integration of the spatial distribution. These second order features include gray level cooccurrence matrix (GLCM), gray level run length matrix (GLRLM), as well as neighbor intensity difference [19]. These features also involve the development of a parent matrix which is an equation of energy, entropy, dissimilarity and correlation features. The trilinear interpolation preprocessing filter was applied to resample the voxel size in the three dimensions to a constant value.

The x,y,z demission of the voxel size was set to 0.488 mm, 0.488mm and 1 mm respectively. The calculation intensity-based features were preceded by applying the Laplacian of Gaussian (LoG). The standard deviation (sigma) of the LoG filtered ranged from 0.5 to 2.5 voxels for a total of 5 iterations [20]. The Butterworth smoothing preprocessing filter was applied to the intensity and texture features to calculate the impact of smoothing and noise removal on the radiomics features. The ROIs were fitted to 512 512 pixels when applying the Butterworth filters. The uniformity of voxel size was done by applying the 2-dimensional Butterworth filters with the 3d voxel size before the smoothing process. More of these statistical texture features, along with their relevant equations were illustrated by Davnall et al. [21]. More details describing the data can be found in [22].

Data

Table 2-1 shows the demographics of the 529 patients for the clinical features and the outcomes for OS and RFS. The cohort was predominately male (87%) and the median age was 58 with ages ranging from 21 to 88. Most cancers (87%) were stage 4 according to AJCC staging. Over half of the cohort (58%) was HPV positive. 20% of patients died during follow-up and 18% experienced a relapse indicating that most patients are censored for the OS and RFS outcomes. Over 3800 radiomic features accompanied the clinical data for the patients.

Table 2-1. Demographics and disease characteristics for the OPC dataset.

Total # Patients	529
Gender	
Male (%)	462 (87%)
Female (%)	67 (13%)
Age At Diagnosis (years)	
Median (Range) (25th-75th Centiles)	58.1 (21 – 88) (52 – 65)
T Category	
T1/T2 (%)	329 (62%)
T3/T4 (%)	200 (38%)
N Category	
< N2b (%)	120 (23%)
≥ N2b (%)	409 (77%)
AJCC Stage (7th Edition)	
I (%)	2 (<1%)
II (%)	8 (2%)
III (%)	59 (11%)
IV(%)	459 (87%)
Smoking Packs Per Year	
Median (Range) (25th-75th Centiles)	5 (0 –120) (0 – 30)
Smoking Status	
Former (%)	185 (35%)
Current (%)	111 (21%)
Never (%)	233 (44%)
Subsite	
Tonsil (%)	199 (38%)
Base of Tongue (%)	285 (54%)
Other (%)	45 (8%)
HPV Status	
Positive (%)	307 (58%)
Negative (%)	49 (9%)
Unknown (%)	173 (33%)
Vital Status	
Alive (%)	423 (80%)
Deceased (%)	106 (20%)
Median in months (Range) (25th-75th Centiles)	70.5 (1.10 - 148.37) (47.37 - 99.77)
Relapse Free Survival	
Yes (%)	435 (82%)
No (%)	94 (18%)
Median in months (Range) (25th-75th Centiles)	64 (1.10 - 144.37) (40.57 - 97.80)
Local Control	
Yes (%)	483 (91%)

Table 2-1. Continued

No (%)	46 (9%)
Median in months (Range) (25th-75th Centiles)	67.47 (1.10 - 148.37) (44.03 - 98.37)

Data Processing

Features which have a high Spearman rank correlation (≥99%) with at least one other feature were removed to minimize redundancy. Features with little variability were also removed as they were not considered informative for prediction. Features which were highly skewed were log transformed. Features with negative values were shifted up by 1 + the most negative value in the case where the median value of the feature was positive. If the median value of the feature was negative then the feature was shifted downward in the same way and then negated before applying a log transform and negating once again.

Chapter 3 Radiomic Feature Selection

This chapter considers feature selection as a method of reducing the dimensionality and selecting a radiomic signature. Feature selection [23] is a process where a subset of features are selected to train and predict with machine learning models rather than the entire set of features. For a set of n features, there are 2^n possible subsets of features including the set of all features. It is not feasible to try all combinations of features since there are over one million possible combinations with just 20 features. Instead, feature selection algorithms systematically choose features and fall within one of three categories: filter, wrapper, or embedded. Filter based feature selectors remove features by considering their inherent importance like correlation with the outcome or variance of the feature itself. As an example, a filter-based technique was already utilized during the preprocessing of the dataset when features with extremely low variance were pruned. Wrapper based feature selectors consider the performance with a learning model in choosing relevant features. For example, with sequential forward selection, the subset of features is built one-by-one, and at each step the feature that is chosen is the one which best predicts the outcome. The last class of feature selector is the embedded selector. With an embedded selector features are chosen as part of the learning model itself. Decision trees are an example of this and are explained later. This chapter explores many feature selection algorithms with different endpoint types: binary, continuous, and right-censored, and evaluates the predictive power of the feature selection algorithms themselves and their efficacy with the different endpoint types.

Methods

To further select the radiomic features after the initial pruning during preprocessing, we considered eight feature selection and feature extraction algorithms. Each feature selection algorithm assumes that the outcome of interest is either (1) binary, (2) continuous, or (3) time-to-event with censoring indicator. Pre-processing the data for these three different outcomes permits the use of many different feature selection algorithms beyond just those for right-censored data. We consider three different ways of pre-processing the time-to-event outcome to be used in the feature selection algorithms described below.

1. **Binary outcome.** The outcome is dichotomized based on if the event was experienced prior to the median observed follow-up time. Censored patients whose follow-up time is less than the median value are removed from the feature selection process. Thus the binary outcome feature selection is informed on fewer patients.

2. **Censor-incorporating outcome.** The Martingale residual is computed from a Cox proportional hazard model. The Martingale residual can be thought of as the variability in the time-to-event outcome which is not explained by the clinical covariates included in the model. The multivariable Cox model included the following variables: sex, age, tumor subsite, T stage, N stage, American Joint Committee on Cancer (AJCC) 7th edition stage, HPV status, and smoking status. The Martingale residual is a continuous outcome.

3. **Censored-aware outcome.** The time-to-event outcome and censoring information is used directly with no additional transformation.

Feature Selector/ Extractor	Abbrev.	Supervised	Outcome Type	Classifier
Minimum Redundancy Maximum Relevance	MRMR	Y	B	N
Wilcoxon Rank Sum Test	Wilcoxon	Y	B	N
Random Forest	RF	Y	B	Y
RReliefF	RReliefF	Y	R	N
Random Regression Forest	RRF	Y	R	Y
Incremental Association Markov Blanket	IAMB	Y	C	N
Random Survival Forest	RSF	Y	C	Y
Principal Component Analysis	PCA	N	-	N

Table 3-1. Feature selection methods summary. Outcome type: Binary (B), Continuous (R), Censored-aware (C).

Each of the three outcome types is used with at least two different feature selection algorithms from the ML literature. A total of seven feature selection methods were applied to the dataset all of which are supervised. We also considered one unsupervised feature extraction method, principal component analysis, due to its popularity in high-dimensional data analysis. Feature extraction differs from feature selection in that with feature extraction the feature space is transformed to a smaller dimensional representation, and so none of the original features exist as they were after the reduction. For completeness, we also compare the performance of the feature selection methods to randomly selecting 10 radiomic features and using clinical features only. The feature selectors include Minimum Redundancy Maximum Relevance (MRMR), Wilcoxon rank sum test (Wilcoxon), random forest (RF), RReliefF, random regression forest (RRF), Incremental Association Markov Blanket (IAMB), random survival forest (RSF), and principal component analysis (PCA). Table 3-1 summarizes the algorithms and the type of outcome variable they require (binary (B), continuous (R), time-to-event (C)). Some of the methods can also be used as predictive models (noted in the table). We describe each feature selection method and any relevant parameters and implementation details below.

The **Minimum Redundancy and Maximum Relevance (MRMR)** feature selector is frequently used in gene expression experiments [24]. It seeks to find a subset of features which are individually highly correlated with the outcome (relevance), yet distinct from any other selected features (redundancy). Redundancy, W, is minimized and is defined by the following equation:

$$W = \frac{1}{|S|^2} \sum_{i,j \in S} I(i,j)$$

Relevance, *V*, on the other hand is maximized and is defined by the equation:

$$V = \frac{1}{|S|} \sum_{i \in S} I(i,h)$$

In both equations, *S* refers to the set of all features considered. *I*(*i, j*) and *I*(*i, h*) are both measures of correlation or association between covariates or a covariate and outcome. Maximum relevancy and minimum redundancy can then be achieved by obtaining the maximum difference between *V* and *W* or the maximum ratio of *V* to *W*. The mRMRe R package [25] is used for selecting the features. We specify 10 features as the number of features to select as this is near the average number of features selected by the other methods.

The **Wilcoxon Rank-Sum Test (Wilcoxon)** [26], also known as the Mann-Whitney test, is a statistical test which selects features based on whether or not two distributions differ by some shift and does not assume a normal distribution of the data. We run Wilcoxon with 10 different splits of the dataset. We select the features by splitting the dataset into 50-folds via Monte-Carlo cross-validation and running the feature selector over these splits where the test set is a tenth of the number of rows. The top features are those which appear the greatest number of times in the top 20 features of each fold. The cutoff for the features is determined based on where the largest decline in occurrences is. For example, if feature A appears 5 times, feature B appears 4 times, feature C appears 1 time, and all other features appear 0 times, then only features A and B will be selected as the largest jump in number of occurrences happens between B and C. The provided the R package, WilcoxCV [27], is used for the Mann-Whitney test with cross-validation.

A **random forest (RF)** [28] is an ensemble-based method consisting of decision trees and is typically used for classification. A single decision is formed by splitting a single feature into multiple nodes where each node is some value or set of values for the feature. In a random forest, instances for each tree are bootstrap sampled from the dataset, and the splitting feature for a node of a tree is chosen from a random subset of features. Two other Random Forest feature selectors, **random regression forests (RRF)** [28] and **random survival forests (RSF)** [29], are used as well. They are also based on an ensemble of trees, however, they predict different outcome types; random survival forests predict right-censored outcomes with survival trees and random regression forests predict continuous outcomes with regression trees. A combination of variable hunting and variable importance (VIMP) [30] is used for feature selection with all of the random forests. The Random Forests are over five Monte Carlo iterations. All the random forests are implemented with the R package randomForestSRC from [31].

RReliefF is a feature selector and an extension of the Relief and ReliefF algorithms. 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 [32]. 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. Unlike Relief and ReliefF which require a class based outcome, RReliefF can calculate feature importance based on a continuous outcome. This is achieved 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:

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

Choosing the cutoff point for which features to select is done in the same way as the Wilcoxon feature selector, except instead of basing the cutoff point on number of occurrences, it is established by finding the largest gap in the feature importance value returned from the algorithm for each feature. The RReliefF algorithm is implemented with the R package CORElearn from [32].

The **Incremental Association Markov Blanket (IAMB)** [33] feature selector finds a subset of features which excludes those independent of the target outcome. IAMB works in two phases: a growth phase and shrink phase. The growth phase adds independent features based on mutual information and continues until no new features are added. The shrink phase eliminates false positives by measuring conditional independence between the outcome and each feature chosen in the growth phase. We utilize the R package MXM [34] which provides a variant of

IAMB suitable for right-censored outcomes.

Principal component analysis (PCA) is the only unsupervised method as well as the only feature extraction method used in this chapter. PCA transforms the set of features into a set of components which are uncorrelated and thus can reduce dimensionality [35]. We don't desire every component as most do not give much additional information. Instead we retain a number of components which explains at least 95% of the variance in the data, and this can be a very small number of components in comparison to the actual number of features. Since with this dimensionality reduction technique, the feature space is transformed, it's not as clear which features are indicative of the outcome, and so interpretation of feature importance is not as straightforward when using PCA compared to the other methods which return a subset of the original features. No features are log transformed prior to extraction like in the other methods, however, all features are scaled and centered.

The selected radiomics features identified using the feature selection algorithms together with other relevant clinical features are used for outcome prediction. For this study we considered estimating 5-year overall survival (OS) and relapse-free survival (RFS). All the patients included in this study have complete radiomic data for the primary tumor but may have other clinical data missing. Missing values are imputed using Multivariate Imputation by Chained Equations (MICE) [36] prior to evaluation.

The predictive models and their corresponding R-packages include logistic regression from R-core, Cox proportional hazards from [37], random forest and random survival forest from [31], logistic elastic-net (LEN) from [38], and an ensemble (i.e., combination) of these five models. Some of these prediction algorithms do not directly handle right censored survival data. We used inverse probability of censoring weighting (IPCW) to extend machine learning methods for survival analysis [9].

An overview of performance metrics for model comparison in survival analysis is provided in [39]. Three different metrics are used to evaluate performance: Harell's C-index, ROC, and Nam-D'Agostino calibration test statistic. If the models are well-calibrated, the calibration test statistic follows a chi-square distribution with 8 degrees of freedom. Test statistics larger than 15.5 would indicate that the models are significantly miscalibrated at the 0.05 significance level. The ROC and Harell C-index are measures of the predictive power of the learning model where higher values indicate better predictive power. A separate test set is not included for evaluation. Instead, a five-fold cross-validation is used for evaluation. Figure 3-1 shows the overall processing pipeline for the feature selection experiment.

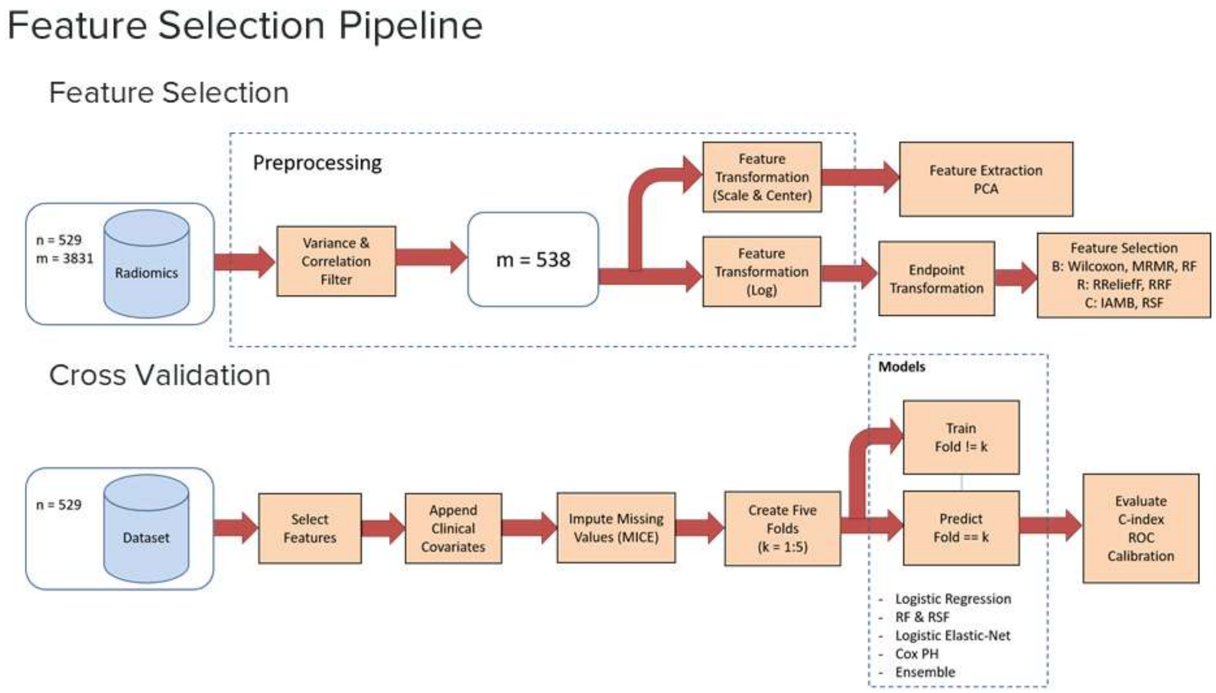


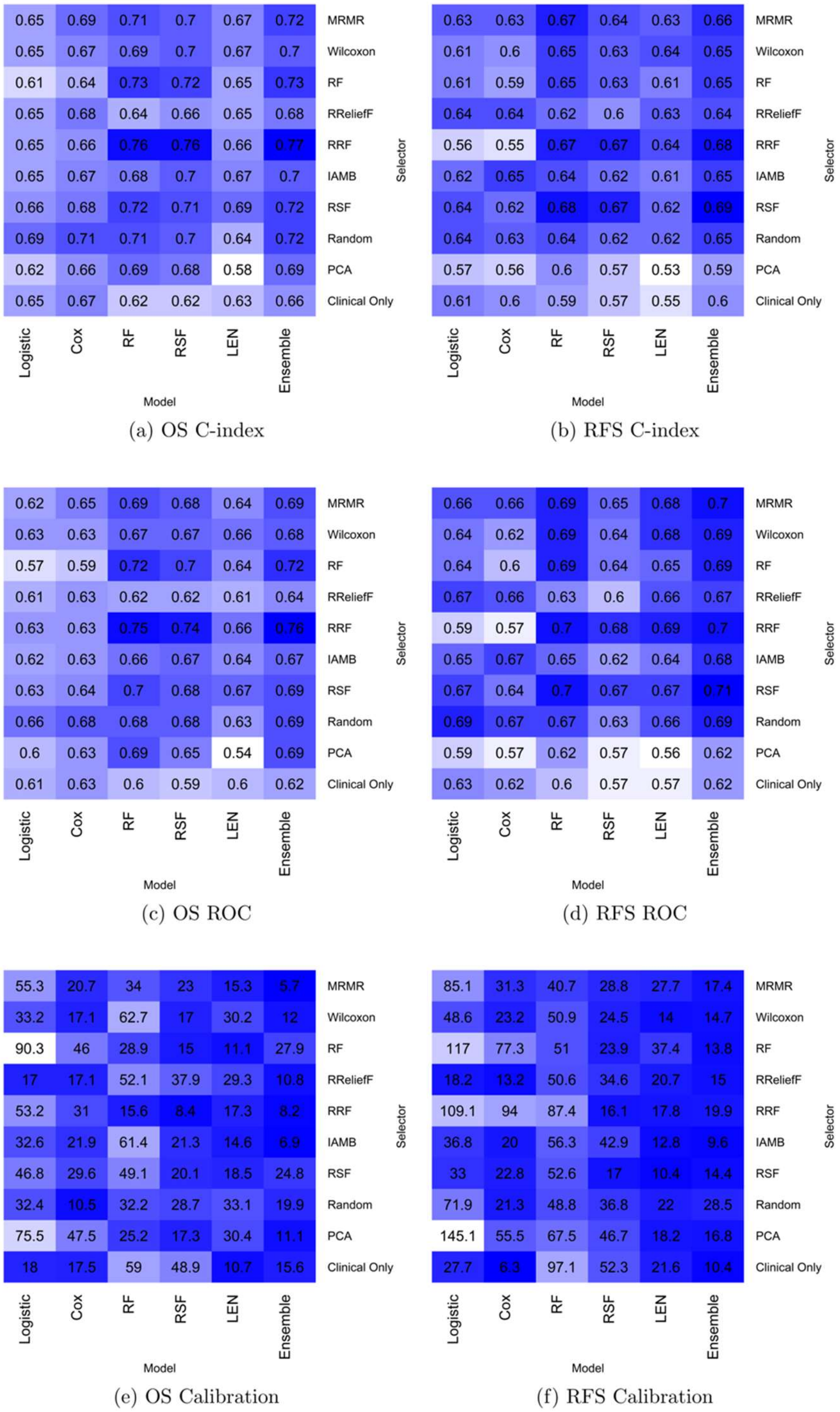
Figure 3-1. The processing pipeline for feature selection and evaluation. The top portion of the figure details the steps for feature selection and extraction. The bottom portion of the figure illustrates the evaluation procedure with cross validation.

Results

Models Performance

Figure 3-2 shows heatmaps displaying the scores for each feature selector and model pair in predicting OS and RFS respectively. Comparing the different risk prediction methods, the ensemble model resulted in the best scores for many feature selectors across all metrics. Often ROC and C-index were greater than 0.7 and calibration test statistic was less than 15.5. For all selectors with the ensemble model, the average C-index, ROC, and calibration scores are 0.709, 0.685, and 14.29 for OS, respectively. The averages for RFS are 0.646, 0.677, and 16.05, respectively. Logistic regression and RF models tended to result in poor calibration test statistics (typically > 15.5) independent of the feature selection method. The RSF predictions are comparable in ROC and C-index scores to RF, with both following the ensemble, but a few of the calibration scores were poor for some selectors, although more frequently for RFS. Elastic net and Cox on the other hand, had more acceptable values for calibration in general (with Elastic net being better more consistently), however, its ROC and C-index scores were lower and comparable to the logistic model.

Considering the different radiomic feature selector algorithms, compared to only clinical data, all supervised methods selected a subset of features which resulted in higher ROC and C-index values for both PFS and OS, many times with an increase of more than 0.1 for those metrics. Clinical only ROC and C-index values always fell below 0.68 for OS and 0.64 for RFS. From here on, we discuss the results of the different feature selection algorithms in the context of the ensemble risk prediction model as this was consistently the best performing model. PCA and random selection achieved good results for OS but not so for RFS. This indicates that radiomic features are more informative for RFS than OS, as can be expected since cause of death is often non-disease related. Random selection performed better than PCA and occasionally achieved fairly high ROC and C-index scores, however it rarely achieved acceptable calibration scores, especially for RFS. The feature selection procedures which use a binary outcome, MRMR, Wilcoxon, and RF all had similar scores in C-index (0.65-0.66) and ROC (0.69-0.70) for the RFS outcome. However, RF was the best performing among the features selectors using a binary outcome when considering OS with C-index (0.7-0.73) and ROC (0.68-0.72). In general, the random forest (RF) selectors (RF, RRF, or RSF) achieved the best scores all-around. Occasionally, a RF selector would tie with another selector or be beat by a small margin (≤0.01 for ROC and C-index), and sometimes calibration scores were high as was for RF and RSF in predicting OS. Despite that, an RF-based selector always achieved the highest ROC and C-index scores with a reasonable calibration for both outcomes. RRF performed the best for OS (ROC: .76, C-index: .77, Calibration: 8.2) and RSF performed the best for RFS (ROC: .71, C-index: .69, Calibration: 14.4).



Selector	Logistic	Cox	RF	RSF	LEN	Ensemble
MRMR	0.65	0.69	0.71	0.7	0.67	0.72
Wilcoxon	0.65	0.67	0.69	0.7	0.67	0.7
RF	0.61	0.64	0.73	0.72	0.65	0.73
RReliefF	0.65	0.68	0.64	0.66	0.65	0.68
RRF	0.65	0.66	0.76	0.76	0.66	0.77
IAMB	0.65	0.67	0.68	0.7	0.67	0.7
RSF	0.66	0.68	0.72	0.71	0.69	0.72
Random	0.69	0.71	0.71	0.7	0.64	0.72
PCA	0.62	0.66	0.69	0.68	0.58	0.69
Clinical Only	0.65	0.67	0.62	0.62	0.63	0.66

(a) OS C-index

Selector	Logistic	Cox	RF	RSF	LEN	Ensemble
MRMR	0.63	0.63	0.67	0.64	0.63	0.66
Wilcoxon	0.61	0.6	0.65	0.63	0.64	0.65
RF	0.61	0.59	0.65	0.63	0.61	0.65
RReliefF	0.64	0.64	0.62	0.6	0.63	0.64
RRF	0.56	0.55	0.67	0.67	0.64	0.68
IAMB	0.62	0.65	0.64	0.62	0.61	0.65
RSF	0.64	0.62	0.68	0.67	0.62	0.69
Random	0.64	0.63	0.64	0.62	0.62	0.65
PCA	0.57	0.56	0.6	0.57	0.53	0.59
Clinical Only	0.61	0.6	0.59	0.57	0.55	0.6

(b) RFS C-index

Selector	Logistic	Cox	RF	RSF	LEN	Ensemble
MRMR	0.62	0.65	0.69	0.68	0.64	0.69
Wilcoxon	0.63	0.63	0.67	0.67	0.66	0.68
RF	0.57	0.59	0.72	0.7	0.64	0.72
RReliefF	0.61	0.63	0.62	0.62	0.61	0.64
RRF	0.63	0.63	0.75	0.74	0.66	0.76
IAMB	0.62	0.63	0.66	0.67	0.64	0.67
RSF	0.63	0.64	0.7	0.68	0.67	0.69
Random	0.66	0.68	0.68	0.68	0.63	0.69
PCA	0.6	0.63	0.69	0.65	0.54	0.69
Clinical Only	0.61	0.63	0.6	0.59	0.6	0.62

(c) OS ROC

Selector	Logistic	Cox	RF	RSF	LEN	Ensemble
MRMR	0.66	0.66	0.69	0.65	0.68	0.7
Wilcoxon	0.64	0.62	0.69	0.64	0.68	0.69
RF	0.64	0.6	0.69	0.64	0.65	0.69
RReliefF	0.67	0.66	0.63	0.6	0.66	0.67
RRF	0.59	0.57	0.7	0.68	0.69	0.7
IAMB	0.65	0.67	0.65	0.62	0.64	0.68
RSF	0.67	0.64	0.7	0.67	0.67	0.71
Random	0.69	0.67	0.67	0.63	0.66	0.69
PCA	0.59	0.57	0.62	0.57	0.56	0.62
Clinical Only	0.63	0.62	0.6	0.57	0.57	0.62

(d) RFS ROC

Selector	Logistic	Cox	RF	RSF	LEN	Ensemble
MRMR	55.3	20.7	34	23	15.3	5.7
Wilcoxon	33.2	17.1	62.7	17	30.2	12
RF	90.3	46	28.9	15	11.1	27.9
RReliefF	17	17.1	52.1	37.9	29.3	10.8
RRF	53.2	31	15.6	8.4	17.3	8.2
IAMB	32.6	21.9	61.4	21.3	14.6	6.9
RSF	46.8	29.6	49.1	20.1	18.5	24.8
Random	32.4	10.5	32.2	28.7	33.1	19.9
PCA	75.5	47.5	25.2	17.3	30.4	11.1
Clinical Only	18	17.5	59	48.9	10.7	15.6

(e) OS Calibration

Selector	Logistic	Cox	RF	RSF	LEN	Ensemble
MRMR	85.1	31.3	40.7	28.8	27.7	17.4
Wilcoxon	48.6	23.2	50.9	24.5	14	14.7
RF	117	77.3	51	23.9	37.4	13.8
RReliefF	18.2	13.2	50.6	34.6	20.7	15
RRF	109.1	94	87.4	16.1	17.8	19.9
IAMB	36.8	20	56.3	42.9	12.8	9.6
RSF	33	22.8	52.6	17	10.4	14.4
Random	71.9	21.3	48.8	36.8	22	28.5
PCA	145.1	55.5	67.5	46.7	18.2	16.8
Clinical Only	27.7	6.3	97.1	52.3	21.6	10.4

(f) RFS Calibration

Figure 3-2. Heatmaps for each of the different feature selection and learning models showing C-Index for (a) OS and (b) RFS, ROC for (c) OS and (d) RFS, and Calibration for (e) OS and (f) RFS. Darker colors indicate a better score for all metrics.

Effect of Censored Outcome Transformation

Table 3-2 shows the number of features selected by each method for the OS and RFS outcomes. The number of features selected ranged from 1-24 (mean 10.1, median 9.5), with RF methods selecting the largest number of features. There is not significant overlap between the features selected by the different methods. The number of features selected for both outcomes is similar but larger for RFS (OS: 64 features, RFS: 78 features). RFS has a considerably larger amount of overlap in selected features between the methods with three times as much overlap as OS. For OS, four features were selected by at least two methods (with F32.NeighborIntensityDifference25Complexity being the only one selected by all three binary methods). The features selected by the continuous methods had no overlap with the binary selected ones. For RFS 12 features were selected by at least two methods.

Method	# features selected		Features selected by two or more methods
	OS	RFS	
MRMR (B)	10	10	OS: F2.GrayLevelCooccurenceMatrix25270.4ClusterProminence (R+C) F32.NeighborIntensityDifference25Complexity (3B) F48.GrayLevelCooccurenceMatrix25225.6ClusterProminence (B+C) F7.IntensityDirectEnergy (B+C)
Wilcoxon (B)	5	7	
RF (B)	20	22	
RReliefF (R)	1	4	RFS: F2.GrayLevelCooccurenceMatrix25180.3InverseDiffMomentNorm (B+C) F2.GrayLevelCooccurenceMatrix25180.5ClusterShade (B+R) F2.GrayLevelCooccurenceMatrix25225.2ClusterProminence (2B) F2.GrayLevelCooccurenceMatrix25270.5ClusterShade (B+R) F2.GrayLevelCooccurenceMatrix25315.6ClusterProminence (2B+R) F20.NeighborIntensityDifference25Complexity (B+R) F29.IntensityDirectGlobalMax (R+C) F29.IntensityDirectLocalRangeMax (B+R) F48.GrayLevelCooccurenceMatrix25180.7ClusterShade (B+R) F48.GrayLevelCooccurenceMatrix25225.7ClusterShade (2B) F48.GrayLevelCooccurenceMatrix25270.1ClusterProminence (2B) F50.NeighborIntensityDifference25TextureStrength (B+R)
RRF (R)	16	24	
IAMB (C)	2	2	
RSF (C)	10	9	

Table 3-2. Number of features selected by the different methods for OS and RFS, and features selected by at least two different methods. Outcome type: B=Binary, R=Continuous, and C=Censored-aware.

Figure 3-3 shows the ROC and calibration metrics for the ensemble model using random forest as the feature selection algorithm for binary outcome (RF), continuous (RRF), and censored-aware (RSF) for both OS and RFS. For comparison, we also include random selection of 10 features (RAND), PCA, and clinical only results. As can be seen, for OS the Random Regression Forest (RRF) exhibits both the best calibration and ROC. For RFS, Random Survival Forest (RSF) shows the best ROC with a good calibration.

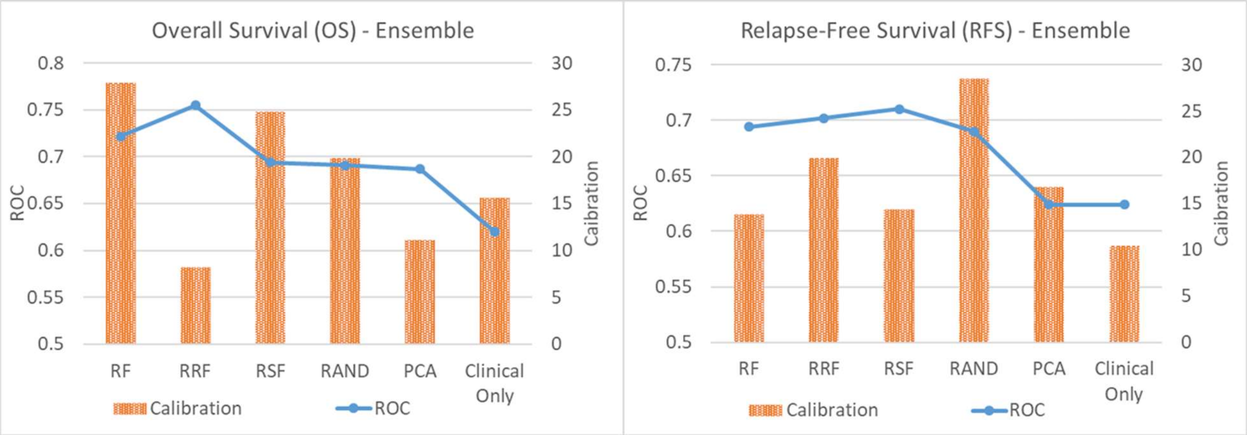


Figure 3-3. ROC and Calibration for the ensemble model when different feature selectors are used for OS (left) and RFS (right).

Discussion

The ensemble model consistently makes the best predictions across all selectors considering all three metrics for both outcomes. Accompanying clinical features with radiomic features certainly improves predictions, however unsupervised feature selection results in either miniscule improvement, bad calibration, or both. RF-based selectors in general select a larger number of features and tend to produce the best accuracy results while maintaining acceptable calibration levels. In particular, the RF selectors for the censored-aware outcome and censor-incorporating outcome (RSF and RRF respectively) achieve the highest predictive power. The different outcomes, OS and RFS, don’t significantly affect the number of features selected in total by all of the methods, however, the RFS outcome results in more overlap in features selected between the methods. Although, neither outcome results in a large overlap in total.

We observed similar results to other previous studies. A few of the selected features, i.e. F29.IntensityDirectGlobalMax, have also been selected in [3], which indicates that these features in particular have predictive value and may be enhanced by the inclusion of other non-redundant radiomic features. As discussed in [12] where only the binary outcome was considered, MRMR and Wilcoxon performed well with MRMR performing slightly better depending on the model. As presented in [11] where only the right-censored outcome was considered, RF feature selectors perform well in predicting depending on the model used. In general, RF models are considered state-of-the-art in machine learning literature, and their efficacy is also apparent in our results.

We conclude that supervised methods should be preferred over unsupervised ones such as PCA as the metric scores are consistently better and resulting features can be interpreted more easily. Among all feature selectors and considering all prediction models, the RF, RRF, and RSF feature selectors give the best predictions for OS and RFS. In particular, of the three, RRF and RSF select the most predictive radiomic signatures. Additionally, RF's implication of a binary outcome imposes some limitations in survival analysis especially when the number of censored samples is high. We recommend the use of RRF or RSF instead of RF in these cases.

A limitation in the experiment is performing dimensionality reduction on the same set of patients for which the learning models are applied due to the small number of patients with radiomic data, large number of radiomic features, and small number of patients with uncensored outcome. However, since we are comparing the performance between the feature selectors, and each feature selector is informed by the same set of patients, we do not expect that any feature selector gets an advantage over others. A limitation in the feature selection process is the variance in number of selected features and namely that a few methods select very few features.

Those feature selectors which selected fewer features like IAMB and RReliefF which both selected < 5 features, performed consistently worse than the others. The random forest selectors (RF, RRF, and RSF) consistently select the higher number of features and tended to produce better scores.

Chapter 4 Radiomic Feature Extraction

This chapter explores the transformation of radiomics data into a single covariate produced by clustering patients. Clustering [40] is a machine learning task which seeks to group instances together even if there is no clear label to assign. In this case, however, we seek to use clustering to obtain a single covariate from many features based on their relations to relapse-free survival (RFS). Other studies [13, 14] have explored clustering based on radiomic data, and the work here differs in that this clustering is performed according to proximity derived from a random forest model [41] and for the different end-point transformations as described in chapter 3.

Methods

Radiomic features are preprocessed and pruned exactly as was done prior to feature selection in chapter 3. In this study, however, the patient dataset is split into a separate training and test set stratified by RFS event occurrence. This allows for a fair comparison between feature selection and feature extraction and how much they improve prediction over only clinical data. Splitting the dataset into training and test sets also more closely emulates the procedure required for establishing prognoses of new patients. Table 4-1 shows the clinical demographics for the training and test splits. 80% of the samples are in the training set, and 20% of samples are in the test set. The clinical covariates are mostly proportional between the sets.

Total # Patients		529
	Train (424, 80%)	**Test (105, 20%)**
Gender		
Male (%)	369 (87%)	93 (88%)
Female (%)	55 (13%)	12 (12%)
Age At Diagnosis (years)		
Median (Range) (25th-75th Centiles)	58.20972 (21 – 88) (53 – 65)	57 (29 – 85) (52 – 65)
T Category		
T1/T2 (%)	259 (61%)	70 (67%)
T3/T4 (%)	165 (39%)	35 (33%)
N Category		
< N2b (%)	95 (22%)	25 (24%)
≥ N2b (%)	329 (78%)	80 (76%)
AJCC Stage (7th Edition)		
I (%)	2 (< 1%)	0 (0%)
II (%)	6 (1%)	2 (2%)
III (%)	47 (11%)	12 (11%)
IV(%)	368 (87%)	91 (87%)
Smoking Packs Per Year		
Median (Range) (25th-75th Centiles)	5 (0 –120) (0 – 30)	3 (0 – 96) (0 – 33)
Smoking Status		
Former (%)	154 (36%)	31 (29%)
Current (%)	84 (20%)	27 (26%)
Never (%)	186 (44%)	47 (45%)
Subsite		
Tonsil (%)	159 (37%)	40 (38%)
Base of Tongue (%)	228 (54%)	57 (54%)
Other (%)	37 (9%)	8 (8%)
HPV Status		
Positive (%)	244 (58%)	63 (60%)
Negative (%)	38 (9%)	11 (10%)
Unknown (%)	142 (33%)	31 (30%)
Vital Status		
Alive (%)	336 (79%)	87 (83%)
Deceased (%)	88 (21%)	18 (17%)
Median in months (Range) (25th-75th Centiles)	69 (2.4 - 148) (47 - 100)	75 (1 - 139) (47 - 100)
Relapse Free Survival		
Yes (%)	344 (81%)	91 (87%)
No (%)	80 (19%)	14 (13%)
Median in months (Range) (25th-75th Centiles)	63 (1 - 144) (40 - 98)	69 (1 – 139) (45 – 95)

Table 4-1. The clinical demographics for patients in each of the test and train splits.

A random forest is generated and a proximity matrix is built based on how frequently a pair of a patients fall within the same nodes of the trees making up the forest. The proximity matrix indicates the proportion of times two patients fall into the same terminal node. Converting this similarity matrix into a dissimilarity measure is straightforward. The resulting dissimilarity matrix can then be processed by clustering methods to assign cluster labels to each patient. Using random forest generated proximity matrices for clustering is known as unsupervised learning by random forests [41]. Following clustering, test set patients are added to a cluster based on their proximities to patients in the clusters. The exact method for assigning test patients to clusters depends on the clustering algorithm. The cluster label is used as a covariate and evaluation is done using the training and test split as opposed to the cross-validation approach in chapter 3. The pipeline in Figure 4-1 illustrates the entire feature extraction and evaluation process.

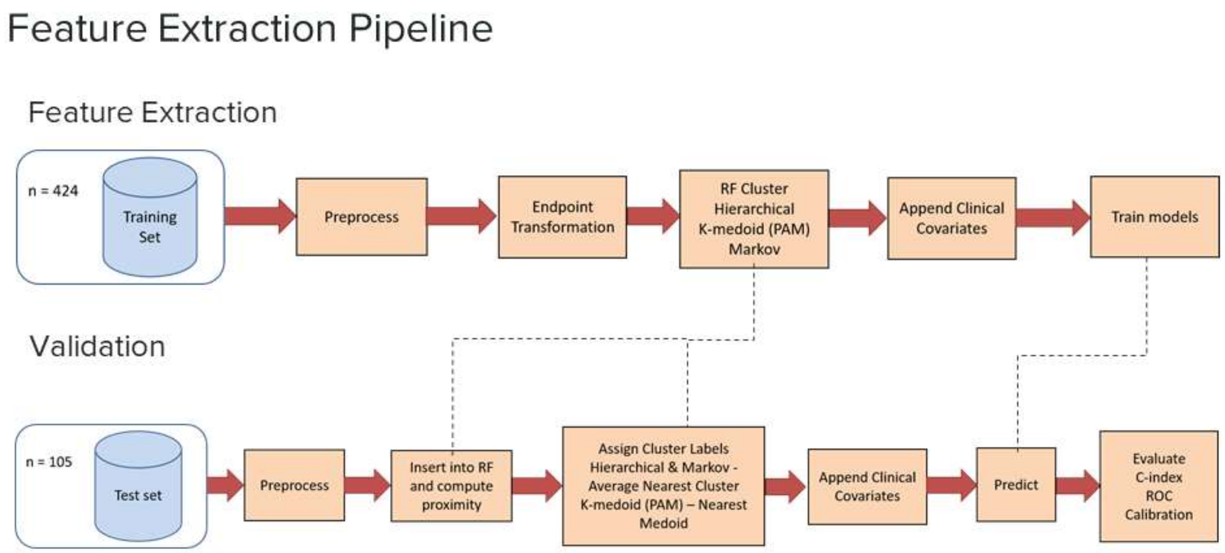

Feature Extraction Pipeline
Feature Extraction
n = 424
Training Set
Preprocess
Endpoint Transformation
RF Cluster Hierarchical K-medoid (PAM) Markov
Append Clinical Covariates
Train models
Validation
n = 105
Test set
Preprocess
Insert into RF and compute proximity
Assign Cluster Labels Hierarchical & Markov - Average Nearest Cluster K-medoid (PAM) – Nearest Medoid
Append Clinical Covariates
Predict
Evaluate C-index ROC Calibration

Figure 4-1. The processing pipeline for feature extraction and evaluation. The top portion of the figure details the steps for feature extraction. The bottom portion of the figure illustrates the evaluation procedure with the test set.

We explore clustering using two outcome transformations, the continuous outcome transformation based on the Martingale residuals and the raw right-censored outcome. The binary outcome transformation is not included in this portion of the study to keep the number of patients sufficient and consistent across methods. An ensemble method that combines the proximity matrices produced by the random forests using the continuous transformation and raw right-censored outcomes is also considered. We refer to it as the combined method from here on. Rather than combining the clusters after they are created, we opted to take the maximum of the proximity between two patients in the proximity matrix for the raw right-censored and the continuous transformed outcomes. We then use this combined matrix to generate the ensemble clustering. The intuition behind taking the maximum of the proximities is that we consider the transformations as complementary and as long as one of the methods considers the patients similar then they should be considered similar. The alternative is taking the minimum, which implies that the patients need to be similar under both transformations. A method for combining clusters after their creation based on Jaccard similarity or other cluster similarity measures as in [42] was considered but not implemented as it did not guarantee a hard cluster assignment, i.e. a patient may be assigned to multiple clusters, and this would not allow for a single covariate representation. A random regression forest is used to generate the proximity between patients for the continuous outcome transformation, and a survival forest is used for the raw right-censored outcome. The random forests are built with a node size of 5, meaning that on average each terminal node contains 5 patients. [43] uses a minimal node size of 5 for feature selection with decision trees for prediction of parkinson’s disease with radiomic data. Proximities for the training set are based only on in-bag samples; the patients which were selected for that bootstrap sample. The number of trees per forest is 1000. Aside from the splitting rule, the random forests’ parameters are the same between the outcomes.

Three clustering algorithms are assessed. Hierarchical clustering, k-medoids clustering, and Markov clustering. After the patients are clustered, a new covariate is created where each patient's value for that covariate is the cluster they were assigned to (e.g. if a patient was clustered in the 3rd cluster, his/her value for the covariate is 3). We evaluate predictive power as was done with radiomic signature selection using the same predictive models, however, we only report the ensemble model's predictions as it was the best performing. For completeness, we compare the clustering methods to random cluster assignment as well as to using only clinical data. We also provide a comparison to some of the feature selection algorithms. We now describe the three clustering methods.

Hierarchical clustering [44] is a greedy approach where clusters are built either by starting with one large cluster and splitting it apart (divisive), or starting with a cluster for each individual point and then merging them at each step (agglomerative). We used the agglomerative approach. In this approach first the two most similar patients are clustered together. Following that, either two other patients are clustered together, or a patient is merged into the cluster from the first step. Figure 4-2 shows an example dendrogram resulting from hierarchical clustering with an established cut resulting in two clusters. First, observations 5 and 3 are clustered, then observations 2 and 4 are clustered, and finally the cluster observation 1 is clustered with the cluster consisting of observations 5 and 3. Thus, the two clusters are {2, 4} and {1, 5, 3}. A number of different ways exist to measure distance between existing clusters to determine which clusters to merge at each step, and in this study, we used average linkage [45]. Average linkage is calculated with the following equation:

$$AL = \frac{1}{|A||B|} \sum_{a \in A} \sum_{b \in B} d(a, b)$$

A and B are the sets of points defining two clusters and $d(a, b)$ is the distance between points a and b. For hierarchical clustering, new patients are assigned to the cluster for which the average distance between the new patient and the patients within the cluster is lower than for any other cluster. The hierarchical clustering method used is included in the standard R package.

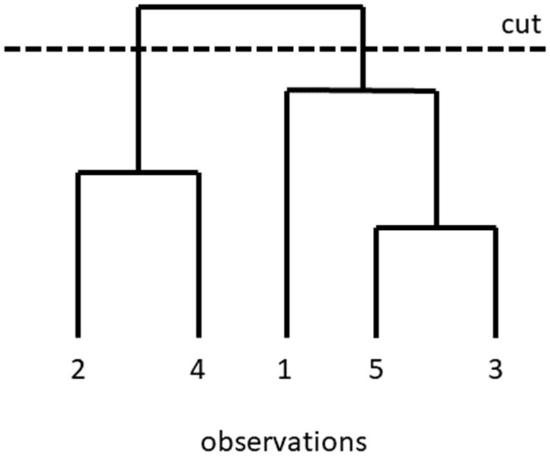


Figure 4-2. A dendrogram with a cut establishing two clusters.

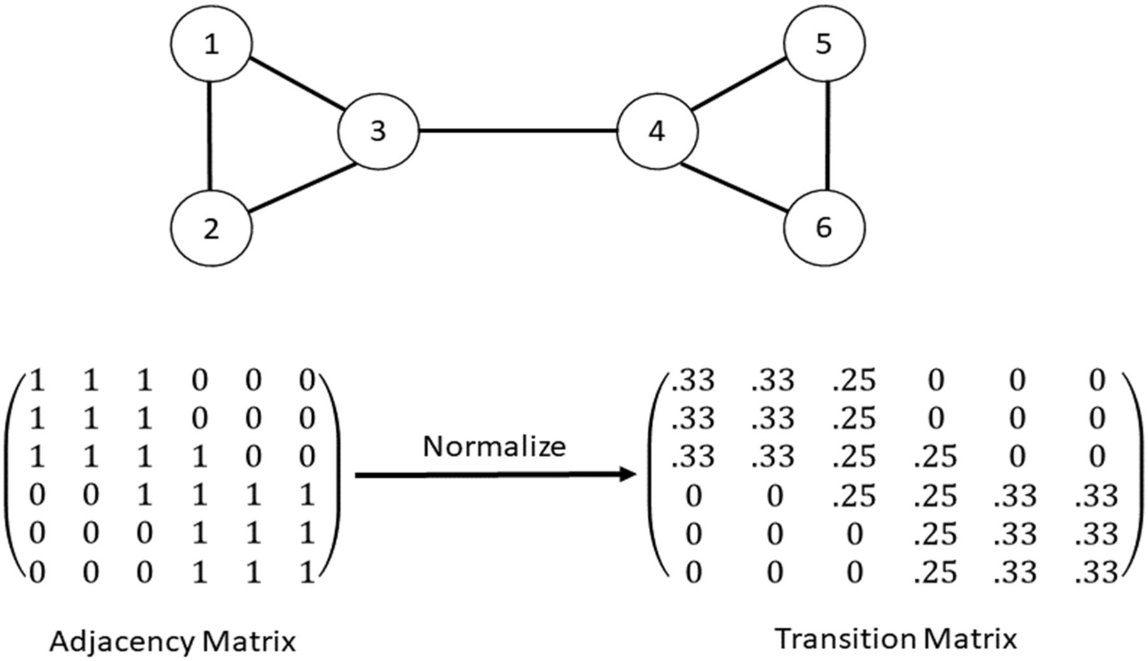
K-medoids clustering [46] is a relative of k-means clustering which does not require the position of the points being clustered. Instead, k-medoids can construct clusters just from the dissimilarity between the points. This is possible since a cluster center is always a point within the cluster, and so distances for each point from the cluster center can be obtained from the dissimilarity matrix itself without the need for recalculating distances. K-medoids has the additional benefit of being more robust to outliers than k-means. K-medoids clustering is performed with the partitioning around medoids algorithm [47]. This is a greedy approach and works as follows:

```
1. Select k points as cluster centers and assign each point to the closest center
2. While the sum of the distances from each point to its cluster center decreases
    a. For each center c, for each non-center n
        i. Swap c and n, and recalculate the sum of distances
        ii. If the sum of distances has increased reverse the swap, otherwise maintain it
```

Test set patients are assigned to the cluster whose median is closest in proximity. The R package cluster [48] provides the implementation of PAM used for this experiment.

Markov clustering [49] is a clustering method intended for clustering the nodes of a graph. In Markov clustering a random walk is performed across the nodes to form a Markov chain. A transition probability matrix is formed from an adjacency matrix by normalizing columns of the matrix. Figure 4-3 illustrates a simple example with six nodes 1 through 6 where the connectivity suggests {1, 2, 3} and {4, 5, 6} are two separate clusters based on visual inspection. A highly connected cluster of nodes is likely to have a random walk which stays within that cluster rather than venture to another cluster of nodes because the transition probabilities dictates that a walk will likely happen from one node in the cluster to another in that same cluster.

Markov clustering requires an adjacency matrix or some other graph representation which denotes which nodes are adjacent to each other. A similarity matrix can be thought of as a fully connected graph of nodes with edge weights equal to the similarities. Since Markov clustering requires adjacencies with no weights, we can instead establish some proximity value as the cutoff value for whether an edge exist between two nodes. In this experiment we establish a cutoff so that the upper 25% of adjacencies are maintained from the proximity matrix and are set to 1. The remaining proximity values are set to 0. An alternative to this approach would be to use the edge weights directly to determine the transitional probabilities. Unlike the other clustering methods, Markov clustering itself determines the number of clusters based on the connectivity of the graph. Assigning test set patients to clusters is the same as for hierarchical clustering. The R package MCL [49] provides the implementation of Markov clustering used.



$$\begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 \end{pmatrix} \xrightarrow{\text{Normalize}} \begin{pmatrix} .33 & .33 & .25 & 0 & 0 & 0 \\ .33 & .33 & .25 & 0 & 0 & 0 \\ .33 & .33 & .25 & .25 & 0 & 0 \\ 0 & 0 & .25 & .25 & .33 & .33 \\ 0 & 0 & 0 & .25 & .33 & .33 \\ 0 & 0 & 0 & .25 & .33 & .33 \end{pmatrix}$$

Figure 4-3. A simple graph and its associated adjacency matrix (self-loops implied). A transition matrix for Markov clustering is formed by normalizing the adjacency matrix by columns.

Results

Models Performance

Figure 4-4 shows the ROC, C-index, and calibration scores for the RFS outcome represented as censored-aware (S), continuous (R), and combined (C) with hierarchical, k-medoid (PAM), and random clustering. Results are shown for 2 and 3 clusters as both PAM and hierarchical clustering require cluster sizes, and these are the lowest possible cluster amounts. For hierarchical clustering, all three outcomes produce ROCs that are consistently above 0.65 for both 2 and 3 clusters. C-indices follow behind ROC for almost all clustering algorithms, however they are usually above 0.6 for hierarchical for different outcome transformations and cluster count. Hierarchical clustering calibrations are typically low except for the combined method with 2 clusters, however, the value is still within an acceptable range. PAM typically achieved ROC and C-index scores in the 0.5 – 0.6 range. With 3 clusters C-index scores were higher than ROC scores for the survival and the combined outcome transformation. In one instance, with the continuous outcome transformation and 2 clusters, and the PAM achieved very high C-index and a ROC over 0.8. PAM performed well with the continuous outcome transformation and 3 clusters, however not nearly as well as with only 2 clusters. Results are only shown for the Markov cluster for the continuous (2 clusters) and combined (3 clusters) outcome transformations. The censored-aware outcome in many cases was not much better than clinical only, so it is not included. With the continuous transformation, Markov clustering achieved scores better than clinical at an acceptable calibration, but not much better than randomly assigned clusters.

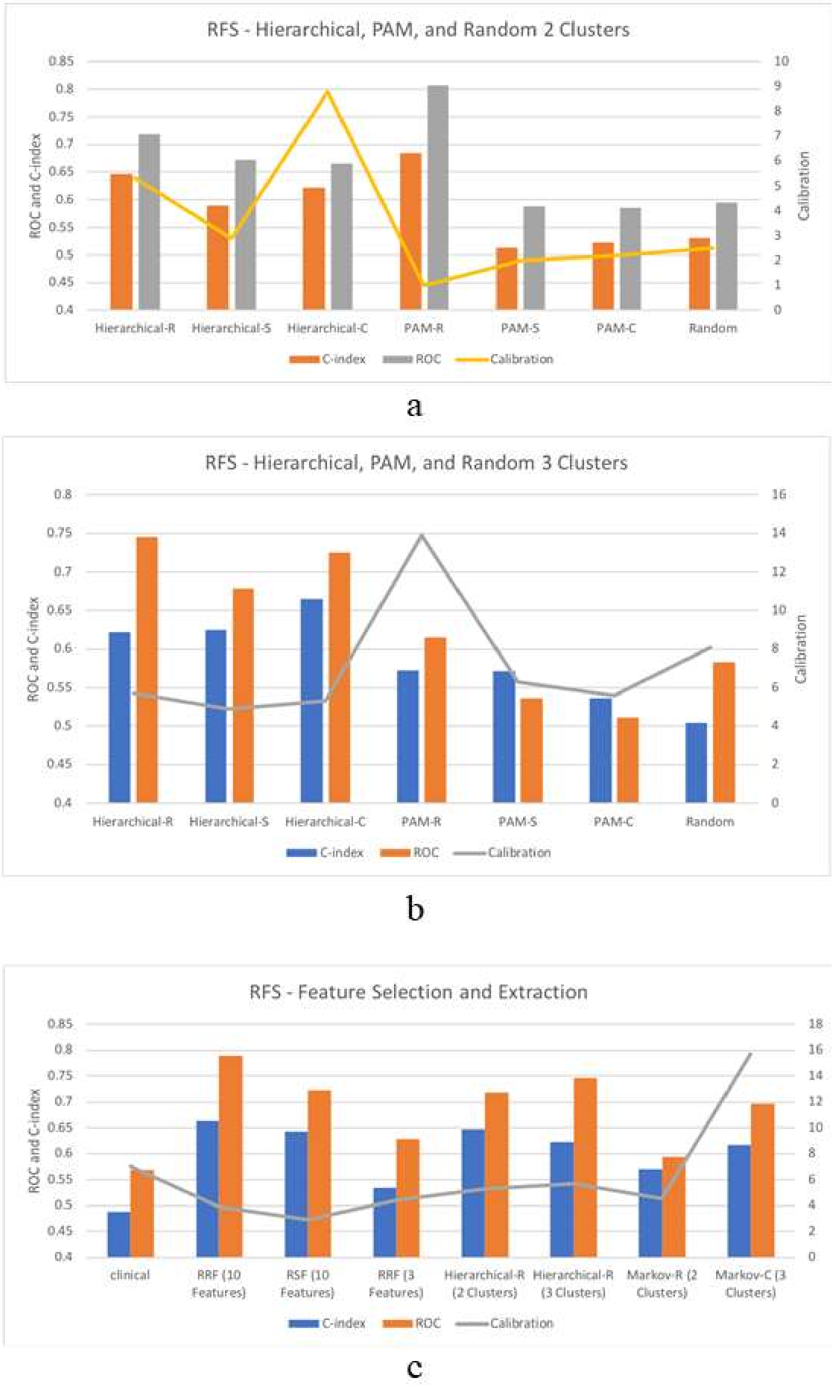


Figure 4-4. The ROC, C-index and calibration scores for RFS prediction. (a) and (b) show metrics for hierarchical, k-medoid (PAM), and random clustering with 2 and 3 clusters respectively. (c) shows metrics for feature selection with RRF and RSF and feature extraction with hierarchical clustering for 3 clusters. It also shows prediction with clinical features alone. The RF clusterings are R: continuous, S: censored-aware, and C: combined.

Compared to random clustering and clinical only, hierarchical clustering improves prediction of RFS substantially for all three outcome transformations and for both cluster counts. PAM likewise shows an improvement over just clinical or random cluster assignment, but only for the continuous outcome transformation. The random forest feature selectors, RRF and RSF, which selected 10 features achieved high ROC and C-index scores at good calibration. The scores for hierarchical clustering, however, followed closely behind. In the case where only the top three features of RRF feature selection were used, ROC and C-index were substantially worse and hierarchical clustering outperformed them with the continuous outcome. The combined method typically scored somewhere between the censored-aware method and the continuous method. In one instance where C-index was similar between the continuous and censored-aware representations, the combined method achieved the highest C-index score.

Effect of Censored Outcome Transformation

Figure 4-5 shows the RFS Kaplan-Meier survival curves for the clusters with each of the proximity matrix sources and for T category with the test set only. A Kaplan-Meier survival curve shows the probability of survival over time in intervals [50]. Drops indicate an event, and vertical ticks indicate a censored patient. The T category is part of the TNM classification which is used for AJCC staging and it describes the size of the primary tumor [51] which is potentially informative of RFS. The T category is represented as two groups: T3 and T4 (largest tumors) and anything below that (smaller or immeasurable tumors). For both groups of T category, there is a visible separation between the two groups. For all the clustering methods there was also visible separation between the clusters. The log-rank test [52] was used to determine whether there was truly a significant difference between the two curves. The resulting p-values are 0.133 for T category, 0.032 for the continuous method, 0.092 for the censored-aware method, and 0.091 for the combined method. At the 95% confidence interval, this indicates that there is a statistically significant difference between the survival curves for the clusters obtained using the continuous method but not for T category or the other clustering methods. The censored-aware and combined methods both have a low number of points in their second clusters which may contribute to lack of statistical significance. The T category groups have similar survival rates up to about 10 months, at which point the two groups diverge. The survival rate of the low survival cluster groups with the continuous method is lower (< 0.8 at < 50 months) than the survival rate of the more severe T category group (< 0.8 at > 50 months). The groups with the higher survival rate are more similar between the continuous and T category curves, however, the T category group sees a quicker drop in survival rate at the beginning (< 0.95 at < 25 months) compared to the cluster group (< 0.95 at > 25 months).

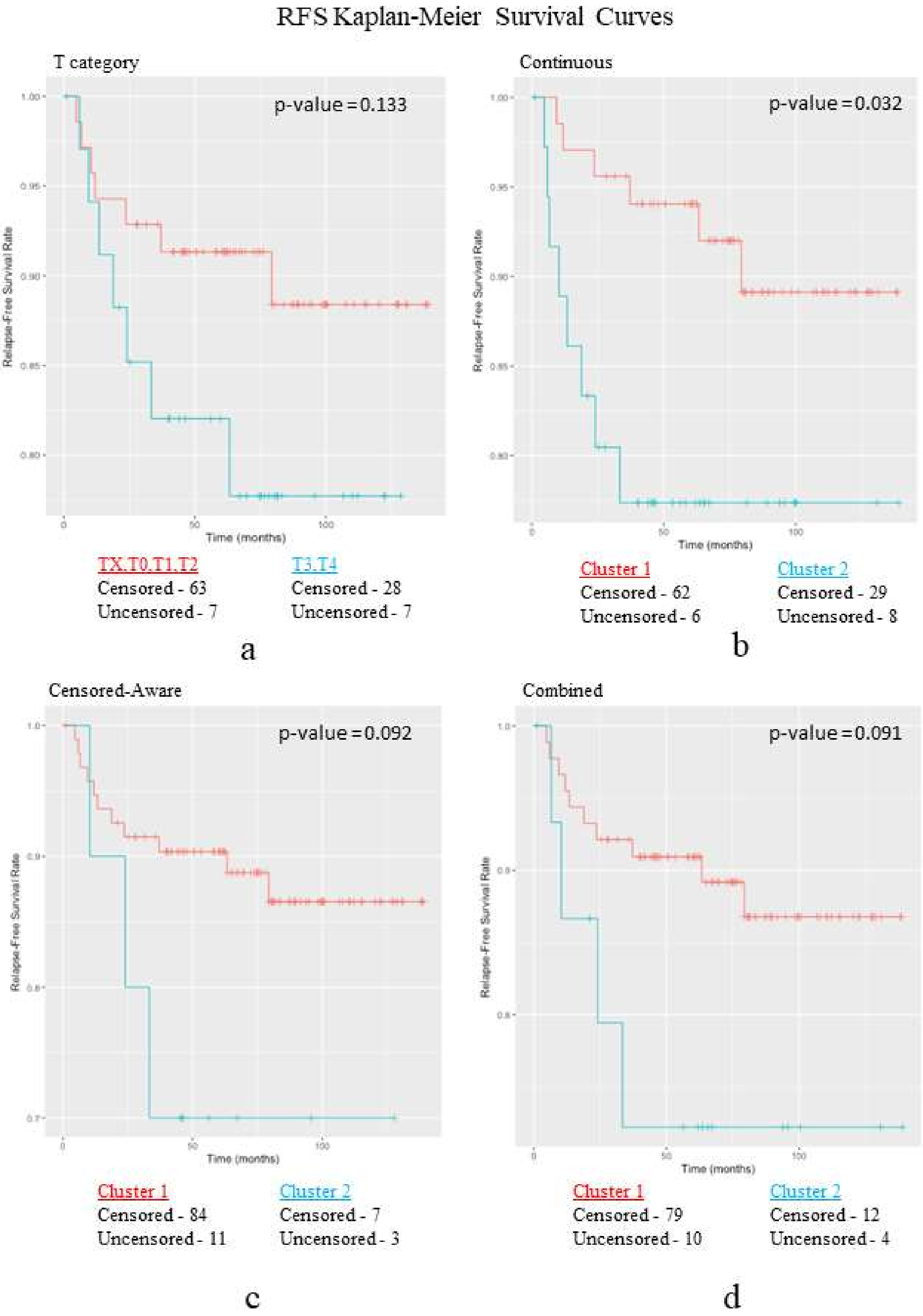


Figure 4-5. The Kaplan-Meier curves for (a) T category, (b) continuous method, (c) censored-aware method, and (d) combined method. All are based on the test set only. The p-value is shown for each of the plots as well as the number of censored and uncensored in each group.

Table 4-2. The demographics of patients in the different clusters with two clusters for T category and the OS and RFS outcomes. These are the cluster distributions for the test set only. S is the censored-aware proximity method, R is the continuous proximity method, and C is the combined proximity method.

Feature	Total	%	S				R				C			
			C1	%	C2	%	C1	%	C2	%	C1	%	C2	%
			95	90%	10	10%	68	65%	37	35%	89	85%	16	15%
T Category														
T1/T2	70	67%	66	69%	4	40%	49	72%	21	57%	61	69%	9	56%
T3/T4	35	33%	29	31%	6	60%	19	28%	16	43%	28	31%	7	44%
Vital Status														
Alive (%)	87	83%	80	84%	7	70%	63	93%	24	65%	76	85%	11	69%
Deceased (%)	18	17%	15	16%	3	30%	5	7%	13	35%	13	15%	5	31%
Median (months)	75	-	75	-	51	-	80	-	56	-	75	-	65	-
Relapse Free Survival														
Yes (%)	91	87%	84	88%	7	70%	62	91%	29	78%	79	89%	12	75%
No (%)	14	13%	11	12%	3	30%	6	9%	8	22%	10	11%	4	25%
Median (months)	69	-	73	-	46	-	76	-	47	-	74	-	63	-

Table 4-2 shows the distribution of patient demographics across clusters for T category and the OS and RFS outcomes. The clusters generated by the censored-aware method are the most lopsided with 95 patients in one cluster and 10 patients in the other cluster. The continuous transformation generated clusters are still unequal, however, they are closer in size than the censored-aware (68 and 37 per cluster). The combined cluster is in between the two but closer in distribution to the survival one (89 and 16 per cluster). Most covariates show no change in distribution in the clusters like gender which is proportionally distributed among clusters so they are not shown in the table. The T category distribution changes for the second clusters, and in particular for the censored-aware method. This indicates that radiomics are associated with the primary tumor size like T category is, but are not completely redundant with T category as the distributions do not change too drastically. For relapse-free survival there is a distinction in the distribution between clusters for all methods in both of the clusters. Additionally, the distribution of vital status across clusters 1 and 2 for all methods is proportionally different indicating a correlation between RFS and OS. The median follow-up times are lower for the second cluster in censored-aware (OS: 51 months, RFS: 46 months) than continuous (OS: 56 months, RFS: 47 months). The combined outcome transformation's second cluster saw the largest median time for OS (65 months) and for RFS (63 months).

Discussion

Overall, including a covariate derived from clustering patients on their radiomic data is indicative of the RFS outcome. The hierarchical clustering method is robust and generates informative clusters across the different outcomes but predicts especially well with the continuous method. PAM improves predictability with the continuous method as well but does not perform as well with the other outcomes. The generated clusters show a divergence in survival time and survival rates as well as other associated clinical features like T category.

Feature extraction with RF clustering has the benefit of only resulting in a single covariate. Compared to radiomic signatures output by feature selection algorithms, the clustering covariate can be much more concise. Having a smaller subset of features to represent the radiomics is especially useful when the number of clinical covariates increases or the number of samples decreases, and the dimensionality relative to the sample size becomes too large (i.e. curse of dimensionality) [53]. The results show that at a lower number of selected features, the single covariate cluster assignment can be more informative. An additional benefit of this method is for preprocessing before training models as well as for predicting for new patients. Because random forest algorithms are not affected by monotonic transformations, there is no need to apply a log transformation or any other type of monotonic transformation to any of the radiomic features either during the feature extraction phase or the training or predicting phase. This avoids issues like needing to recompute transformations when new patients are added. When a very low-dimensional explanation of radiomic data is required, we recommend the use of feature extraction via random forest clustering, and furthermore, we recommend hierarchical clustering and transforming the right-censored outcome to the continuous outcome based on Martingale residuals.

A limitation of this experiment is that it doesn't evaluate or consider the different parameters in the random forest construction. Parameters like node size are likely to have a substantial effect on clustering assignments and subsequently on predictive capabilities. Similarly, only a small number of clusters was tested. By increasing the number of clusters beyond three, it may be possible to capture more of the variability of radiomic features between patients. As an extension of this experiment, node size, number of clusters, and other relevant random forest clustering parameters may be varied over a range of values to observe their effect on clusters and subsequent prediction of patient outcome and perhaps to find the best performing values for each. Some radiomic clustering studies have used techniques to determine an optimal number of clusters; [54] uses PCA and cluster validation while [13] uses consensus clustering. Another limitation of this experiment is that the test set is quite small with only 14 RFS events, only 12 of which occurred before the 5-year follow-up period used for evaluation. The preparation of a separate test set should resolve this and consequently allow for more samples to train the models.

In conclusion, feature extraction via random forest clustering provides a concise yet informative representation of the wide set of radiomic features. Depending on the clustering method and endpoint transformation used to generate similarities, its performance nears that of feature selection, but with a more minimal representation. For this reason, radiomic feature extraction is a viable method when a more compact representation of the feature space is desired or necessary.

Chapter 5 Conclusion

Dimensionality reduction consolidates a large number of radiomic features which are difficult to interpret and apply to learning models as they are. Furthermore, the inclusion of features resulting from dimensionality reduction gives good results in terms of prognosis of OS and RFS. Both feature selection and feature extraction with random forest clustering of patients give substantial improvement over using clinical data alone or using randomly selected features or randomly assigned covariates.

The largest limitation of this study for both forms of dimensionality reduction include the lack of a sufficiently large separate test set of patients. Although both feature selection and extraction were evaluated with a test set, the number of patients in it is not large. Also, to maintain a proportional number of events in each of the test and training set, the number of events is very low in the test set.

In terms of future work, patient data for an independent validation set is being collected which will be used in future analyses to remedy the limitation imposed by the high number of censored patients for both RFS and OS outcomes. Additionally, as more patient data is collected and the amount of censoring changes, we will be able to evaluate outcome transformation across various amounts of censoring. In the initial pruning of radiomic features, we kept features less than 99% correlated with another feature. 99% was chosen as a conservative threshold, however the number of selectable features remains large. With removal of features correlated 90% or above, over 90% of the raw features can be removed. However, this could result in lower performance if relevant features are pruned. Combining the pruned features into a single feature can minimize this information loss. Future work could involve varying the pruning threshold or implementing another approach to reduce the initial redundant set of radiomic features.

In conclusion, dimensionality reduction is efficacious in improving prediction of OS and RFS. For feature selection, the random forest group exhibit the highest predictive scores, with RRF (continuous outcome) and RSF (censored-aware outcome) being the most effective. This could be partly due to the fact that there is a substantial number of censored samples in the dataset. For a similar number of censored patients, we advocate using the censor-incorporating and censored-aware outcomes with RRF and RSF respectively when utilizing feature selection. For feature extraction, clustering with a continuous outcome transformation from the martingale residual via the hierarchical clustering algorithm is the most consistently calibrated and best performing approach for RFS prediction. Additionally, clustering patients with the Martingale outcome transformation compares well to feature selection. Because feature selection results in larger space of features, the choice between feature selection and feature extraction should be based on the number of patients available for training and the number of non-radiomic features included.

References

[1] R. L. Siegel, K. D. Miller and A. Jemal, Cancer statistics, 2017 CA: a cancer. J Clin 67: 7--30, 2017.

[2] K. M. Panth, R. T. H. Leijenaar, S. Carvalho, N. G. Lieuwes, A. Yaromina, L. Dubois and P. Lambin, "Is there a causal relationship between genetic changes and radiomics-based image features? An in vivo preclinical experiment with doxycycline inducible GADD34 tumor cells," *Radiotherapy and Oncology,* vol. 116, pp. 462-466, 2015.

[3] M. D. A. C. C. Head, N. Q. I. W. Group and others, "Investigation of radiomic signatures for local recurrence using primary tumor texture analysis in oropharyngeal head and neck cancer patients," *Scientific reports,* vol. 8, 2018.

[4] H. J. W. L. Aerts, E. R. Velazquez, R. T. Leijenaar, C. Parmar, P. Grossmann, S. Cavalho, J. Bussink, R. Monshouwer, B. Haibe-Kains, D. Rietveld and others, "Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach," *Nat. Commun,* vol. 5, pp. 1-8, 2014.

[5] Y. Huang, Z. Liu, L. He, X. Chen, D. Pan, Z. Ma, C. Liang, J. Tian and C. Liang, "Radiomics Signature: A Potential Biomarker for the Prediction of Disease-Free Survival in Early-Stage (I or II) Non—Small Cell Lung Cancer," *Radiology,* vol. 281, pp. 947-957, 2016.

[6] A. J. Wong, A. Kanwar, A. S. Mohamed and C. D. Fuller, "Radiomics in head and neck cancer: from exploration to application," *Translational Cancer Research,* vol. 5, pp. 371-382, 2016.

[7] P. Royston and D. G. Altman, "External validation of a Cox prognostic model: principles and methods," *BMC medical research methodology,* vol. 13, p. 33, 2013.

[8] T. M. Therneau, P. M. Grambsch and T. R. Fleming, "Martingale-based residuals for survival models," *Biometrika,* vol. 77, pp. 147-160, 1990.

[9] D. M. Vock, J. Wolfson, S. Bandyopadhyay, G. Adomavicius, P. E. Johnson, G. Vazquez-Benitez and P. 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,* vol. 61, pp. 119-131, 2016.

[10] R. J. Gillies, P. E. Kinahan and H. Hricak, "Radiomics: images are more than pictures, they are data," *Radiology,* vol. 278, pp. 563-577, 2015.

[11] S. Leger, A. Zwanenburg, K. Pilz, F. Lohaus, A. Linge, K. Zöphel, J. Kotzerke, A. Schreiber, I. Tinhofer, V. Budach and others, "A comparative study of machine learning

methods for time-to-event survival data for radiomics risk modelling," *Scientific Reports,* vol. 7, p. 13206, 2017.

[12] C. Parmar, P. Grossmann, D. Rietveld, M. M. Rietbergen, P. Lambin and H. J. W. L. Aerts, "Radiomic machine-learning classifiers for prognostic biomarkers of head and neck cancer," *Frontiers in oncology,* vol. 5, 2015.

[13] C. Parmar, R. T. H. Leijenaar, P. Grossmann, E. R. Velazquez, J. Bussink, D. Rietveld, M. M. Rietbergen, B. Haibe-Kains, P. Lambin and H. J. W. L. Aerts, "Radiomic feature clusters and prognostic signatures specific for lung and head & neck cancer," *Scientific reports,* vol. 5, p. 11044, 2015.

[14] C. Tang, B. Hobbs, A. Amer, X. Li, C. Behrens, J. R. Canales, E. P. Cuentas, P. Villalobos, D. Fried, J. Y. Chang and others, "Development of an Immune-Pathology Informed Radiomics Model for Non-Small Cell Lung Cancer," *Scientific reports,* vol. 8, p. 1922, 2018.

[15] H. Wang, M. B. Schabath, Y. Liu, A. E. Berglund, G. C. Bloom, J. Kim, O. Stringfield, E. A. Eikman, D. L. Klippenstein, J. J. Heine and others, "Semiquantitative computed tomography characteristics for lung adenocarcinoma and their association with lung cancer survival," *Clinical lung cancer,* vol. 16, pp. e141--e163, 2015.

[16] A. Liaw, M. Wiener and others, "Classification and regression by randomForest," *R news,* vol. 2, pp. 18-22, 2002.

[17] X. Chen and H. Ishwaran, "Random forests for genomic data analysis," *Genomics,* vol. 99, pp. 323-329, 2012.

[18] 4, "Definition of Volumes," *Journal of the ICRU,* vol. 10, pp. 41-53, 2010.

[19] R. M. Haralick, "Statistical and structural approaches to texture," *Proceedings of the IEEE,* vol. 67, pp. 786-804, 1979.

[20] B. Ganeshan, K. Skogen, I. Pressney, D. Coutroubis and K. Miles, "Tumour heterogeneity in oesophageal cancer assessed by CT texture analysis: preliminary evidence of an association with tumour metabolism, stage, and survival," *Clinical radiology,* vol. 67, pp. 157-164, 2012.

[21] F. Davnall, C. S. P. Yip, G. Ljungqvist, M. Selmi, F. Ng, B. Sanghera, B. Ganeshan, K. A. Miles, G. J. Cook and V. Goh, "Assessment of tumor heterogeneity: an emerging imaging tool for clinical practice?," *Insights into imaging,* vol. 3, pp. 573-589, 2012.

[22] H. Elhalawani, A. S. R. Mohamed, A. L. White, J. Zafereo, A. J. Wong, J. E. Berends, S. AboHashem, B. Williams, J. M. Aymard, A. Kanwar and others, "Matched computed

tomography segmentation and demographic data for oropharyngeal cancer radiomics challenges," *Scientific data,* vol. 4, p. 170077, 2017.

[23] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of machine learning research,* vol. 3, pp. 1157-1182, 2003.

[24] C. Ding and H. Peng, "Minimum redundancy feature selection from microarray gene expression data," *Journal of bioinformatics and computational biology,* vol. 3, pp. 185-205, 2005.

[25] N. De Jay, S. Papillon-Cavanagh, C. Olsen, G. Bontempi and B. Haibe-Kains, "mRMRe: an R package for parallelized mRMR ensemble feature selection," *Submitted,* p. ., 2012.

[26] C. Liao, S. Li and Z. Luo, "Gene selection using wilcoxon rank sum test and support vector machine for cancer classification," in *International Conference on Computational and Information Science*, 2006.

[27] A.-L. Boulesteix, "WilcoxCV: an R package for fast variable selection in cross-validation," *Bioinformatics,* vol. 23, pp. 1702-1704, 2007.

[28] L. Breiman, "Random forests," *Machine learning,* vol. 45, pp. 5-32, 2001.

[29] H. Ishwaran, U. B. Kogalur, E. H. Blackstone and M. S. Lauer, "Random survival forests," *Ann. Appl. Statist.,* vol. 2, pp. 841-860, 2008.

[30] H. Ishwaran, U. B. Kogalur, E. Z. Gorodeski, A. J. Minn and M. S. Lauer, "High-dimensional variable selection for survival data," *Journal of the American Statistical Association,* vol. 105, pp. 205-217, 2010.

[31] H. Ishwaran and U. B. Kogalur, "Random Forests for Survival, Regression, and Classification (RF-SRC)," manual, 2017.

[32] M. Robnik-Šikonja and I. Kononenko, "Theoretical and empirical analysis of ReliefF and RReliefF," *Machine learning,* vol. 53, pp. 23-69, 2003.

[33] I. Tsamardinos, C. F. Aliferis, A. R. Statnikov and E. Statnikov, "Algorithms for Large Scale Markov Blanket Discovery.," in *FLAIRS conference*, 2003.

[34] V. Lagani, G. Athineou, A. Farcomeni, M. Tsagris and I. Tsamardinos, "Feature Selection with the R Package MXM: Discovering Statistically Equivalent Feature Subsets," *Journal of Statistical Software,* vol. 80, pp. 1-25, 2017.

[35] S. Wold, K. Esbensen and P. Geladi, "Principal component analysis," *Chemometrics and intelligent laboratory systems,* vol. 2, pp. 37-52, 1987.

[36] S. Buuren and K. Groothuis-Oudshoorn, "mice: Multivariate Imputation by Chained Equations in R," *Journal of Statistical Software, Articles,* vol. 45, pp. 1-67, 2011.

[37] T. M. Therneau, "A Package for Survival Analysis in S," 2015.

[38] J. Friedman, T. Hastie and R. Tibshirani, "Regularization Paths for Generalized Linear Models via Coordinate Descent," *Journal of Statistical Software, Articles,* vol. 33, pp. 1-22, 2010.

[39] E. W. Steyerberg, A. J. Vickers, N. R. Cook, T. Gerds, M. Gonen, N. Obuchowski, M. J. Pencina and M. W. Kattan, "Assessing the performance of prediction models: a framework for some traditional and novel measures," *Epidemiology (Cambridge, Mass.),* vol. 21, p. 128, 2010.

[40] A. K. Jain and R. C. Dubes, "Algorithms for clustering data," 1988.

[41] T. Shi and S. Horvath, "Unsupervised learning with random forest predictors," *Journal of Computational and Graphical Statistics,* vol. 15, pp. 118-138, 2006.

[42] G. J. Torres, R. B. Basnet, A. H. Sung, S. Mukkamala and B. M. Ribeiro, "A similarity measure for clustering and its applications," *Int J Electr Comput Syst Eng,* vol. 3, pp. 164-170, 2009.

[43] A. Rahmim, P. Huang, N. Shenkov, S. Fotouhi, E. Davoodi-Bojd, L. Lu, Z. Mari, H. Soltanian-Zadeh and V. Sossi, "Improved prediction of outcome in Parkinson's disease using radiomics analysis of longitudinal DAT SPECT images," *NeuroImage: Clinical,* vol. 16, pp. 539--544, 2017.

[44] J. Franklin, "The elements of statistical learning: data mining, inference and prediction," *The Mathematical Intelligencer,* vol. 27, pp. 83-85, 2005.

[45] S. C. Johnson, "Hierarchical clustering schemes," *Psychometrika,* vol. 32, pp. 241-254, 1967.

[46] J. MacQueen and others, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, 1967.

[47] L. Kaufman and P. J. Rousseeuw, Finding groups in data: an introduction to cluster analysis, vol. 344, John Wiley & Sons, 2009.

[48] M. Maechler, P. Rousseeuw, A. Struyf, M. Hubert and K. Hornik, "cluster: Cluster Analysis Basics and Extensions," 2018.

[49] S. M. Van Dongen, "Graph clustering by flow simulation," 2001.

[50] M. K. Goel, P. Khanna and J. Kishore, "Understanding survival analysis: Kaplan-Meier estimate," *International journal of Ayurveda research,* vol. 1, p. 274, 2010.

[51] "Cancer Staging - National Cancer Institute," [Online]. Available: https://www.cancer.gov/about-cancer/diagnosis-staging/staging.

[52] V. Bewick, L. Cheek and J. Ball, "Statistics review 12: survival analysis," *Critical care,* vol. 8, p. 389, 2004.

[53] J. H. Friedman, "On bias, variance, 0/1—loss, and the curse-of-dimensionality," *Data mining and knowledge discovery,* vol. 1, pp. 55-77, 1997.

[54] M. Borri, M. A. Schmidt, C. Powell, D.-M. Koh, A. M. Riddell, M. Partridge, S. A. Bhide, C. M. Nutting, K. J. Harrington, K. L. Newbold and others, "Characterizing heterogeneity within head and neck lesions using cluster analysis of multi-parametric MRI data," *PloS one,* vol. 10, p. e0138545, 2015.