Published in final edited form as:

JCO Clin Cancer Inform. 2018 December ; 2: 1–19. doi:10.1200/CCI.18.00052.

Evaluating the Effect of Right-Censored End Point Transformation for Radiomic Feature Selection of Data From Patients With Oropharyngeal Cancer

Luka Zdilar, **David M. Vock**, **G. Elisabeta Marai**, **Clifton D. Fuller**, **Abdallah S.R. Mohamed**, **Hesham Elhalawani**, **Baher Ahmed Elgohari**, **Carly Tiras**, **Austin Miller**, and **Guadalupe Canahuate**

Luka Zdilar and Guadalupe Canahuate, University of Iowa, Iowa City, IA; David M. Vock, University of Minnesota, Minneapolis, MN; G. Elisabeta Marai, University of Illinois at Chicago, Chicago, IL; Clifton D. Fuller, Abdallah S.R. Mohamed, Hesham Elhalawani, Baher Ahmed Elgohari, and Carly Tiras, The University of Texas MD Anderson Cancer Center; and Austin Miller, McGovern Medical School, Houston, TX.

Corresponding author: Guadalupe Canahuate, MS, PhD, Department of Electrical and Computer Engineering, 4016 Seamans Center for the Engineering Arts and Sciences, University of Iowa, Iowa City, Iowa, 52242; guadalupe-canahuate@uiowa.edu.

AUTHOR CONTRIBUTIONS

Conception and design: Luka Zdilar, David M. Vock, G. Elisabeta Marai, Clifton D. Fuller, Abdallah S.R. Mohamed, Hesham Elhalawani, Baher Ahmed Elgohari, Guadalupe Canahuate

Financial support: David M. Vock, Guadalupe Canahuate

Administrative support: Clifton D. Fuller

Collection and assembly of data: Luka Zdilar, G. Elisabeta Marai, Clifton D. Fuller, Abdallah S.R. Mohamed, Hesham Elhalawani, Baher Ahmed Elgohari, Carly Tiras, Austin Miller, Guadalupe Canahuate

Data analysis and interpretation: Luka Zdilar, David M. Vock, G. Elisabeta Marai, Clifton D. Fuller, Abdallah S.R. Mohamed, Hesham Elhalawani, Baher Ahmed Elgohari, Guadalupe Canahuate

Manuscript writing: All authors

Final approval of manuscript: All authors

Accountable for all aspects of the work: All authors

Luka Zdilar
No relationship to disclose

David M. Vock
No relationship to disclose

G. Elisabeta Marai
Patents, Royalties, Other Intellectual Property: Patent licensed to C-Motion

Clifton D. Fuller
Honoraria: Elekta
Research Funding: Elekta (Inst), RaySearch Laboratories (Inst)
Patents, Royalties, Other Intellectual Property: Patent application pending on unrelated medical device (Inst)
Travel, Accommodations, Expenses: Elekta

Abdallah S.R. Mohamed
No relationship to disclose

Hesham Elhalawani
No relationship to disclose

Baher Ahmed Elgohari
No relationship to disclose

Carly Tiras
No relationship to disclose

Austin Miller
No relationship to disclose

Guadalupe Canahuate
No relationship to disclose

AUTHORS' DISCLOSURES OF POTENTIAL CONFLICTS OF INTEREST

The following represents disclosure information provided by authors of this manuscript. All relationships are considered compensated. Relationships are self-held unless noted. I = Immediate Family Member, Inst = My Institution. Relationships may not relate to the subject matter of this manuscript. For more information about ASCO's conflict of interest policy, please refer to www.asco.org/rwc or ascopubs.org/jco/site/ifc.

Abstract

Purpose—To evaluate the effect of transforming a right-censored outcome into binary, continuous, and censored-aware representations on radiomics feature selection and subsequent prediction of overall survival (OS) and relapse-free survival (RFS) of patients with oropharyngeal cancer.

Methods and Materials—Different feature selection techniques were applied using a binary outcome indicating event occurrence before 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. Radiomic signatures combined with clinical variables were used for risk prediction. Three metrics for accuracy and calibration were used to evaluate eight feature selectors and six predictive models.

Results—Feature selection across 529 patients on more than 3,800 radiomic features resulted in increases ranging from 0.01 to 0.11 in C-index and area under the curve (AUC) scores compared with clinical features alone. The ensemble model yielded the best scores for AUC and C-index (often > 0.7) and calibration (Nam-D'Agostino test statistic often < 15.5 with 8 *df*). The random forest feature selectors achieved the best performance considering all metrics. Random regression forest performed the best in OS prediction with the ensemble model (AUC, 0.75; C-index, 0.76; calibration, 8.7). Random survival forest performed the best in RFS prediction with the ensemble model (AUC, 0.71; C-index, 0.68; calibration, 19.1).

Conclusion—Including a radiomic signature results in better prediction than using only clinical data. Signatures generated randomly or without considering the outcome result in poor calibration scores. The random forest feature selectors for each of the three transformations typically selected the greatest number of features and produced the best predictions at acceptable calibration levels. In particular, random regression forest and random survival forest performed best for OS and RFS, respectively.

INTRODUCTION

Radiomics entails extraction of quantitative imaging features from computed tomography (CT), magnetic resonance imaging, or positron emission tomography images. A large number of radiomic features can be extracted from these images to characterize tumor intensity, shape, and texture. Feature selection identifies tumor signature profiles that can be used for prognostic or predictive evaluation of patient outcomes[1] and that have been putatively associated with clinical and survival outcomes.[2–5]

Several end points, such as overall survival (OS), local control, freedom from distant metastasis, 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 during study follow-up, a patient may not have yet experienced an event but is still 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 for either binary (eg, 1 or 0, yes or no) or continuous noncensored outcomes. Although some methods have been developed to perform feature selection using the right-censored outcomes directly (ie, follow-up time and censor flag),[6] censored time-to-event data frequently must be preprocessed depending on the machine learning and/or feature selection algorithm of choice. For example, to apply many existing machine learning algorithms, the outcome is often transformed into a binary outcome[7] using a predetermined follow-up time (eg, whether the individual experienced local recurrence within 5 years). However, converting the algorithm to select for 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 have not yet had an event. Frequently, ad hoc approaches, such as removing those patients from the analysis or treating them as nonevents, are used. A third alternative is to use the Martingale residual (a continuous outcome) generated from a Cox proportional hazards model as the outcome variable for feature selection algorithms requiring a continuous outcome.[2,8,9]

It is well known that machine learning approaches noncognizant to right-censoring lead to poorly calibrated risk prediction because they fail to accurately account for unequal follow-up times (but discrimination is relatively unaffected).[10] However, it is unclear what effect data preprocessing approaches to handle right-censoring have on algorithms for feature selection. To interrogate the potential impact of right-censoring on radiomics feature identification for longitudinal outcomes, we implemented the following specific aims, comparing performance over several predictive models for oropharyngeal cancer (OPC). First, we evaluated the performance of different feature selection algorithms using three distinct right-censored data preprocessing approaches (binary, censor incorporating [continuous], and censor aware) to represent the censored time-to-event survival data. Second, we assessed the identified radiomic features upon subsequent predictive model specification, using a library of established and novel risk prediction approaches (Cox proportional hazards, random forest [RF], random survival forest [RSF], logistic regression, logistic elastic net), which have been adapted to right-censored outcomes using inverse probability of censoring weights.

To the best of our knowledge, this is the first attempt to systematically evaluate and generate hypotheses on the effects of outcome transformation for radiomics feature selection in model performance. A continuous outcome derived from the Martingale residuals has been used in other imaging biomarker studies[11] but has not been previously used in radiomic feature selection studies.

METHODS AND MATERIALS

Data Source

Our institutional database was retrospectively reviewed for patients with OPC who were treated at The University of Texas MD Anderson Cancer Center (Houston, TX) from 2005 to 2013 after institutional review board approval. Eligible patients who were diagnosed with OPC that was pathologically confirmed by either a biopsy or a surgical excision and who received their treatment on a curative intent were eligible. All patients with OPC were nonsurgically managed with radio-therapy with or without systemic therapy either in neoadjuvant or concurrent settings. Demographic, clinical, toxicity, and outcome data were collected for these patients.

For imaging data, contrast-enhanced CT scans at initial diagnosis before any active local or systemic treatment were retrieved to our commercially available contouring software (Velocity AI v3.0.1; Varian Medical Systems, Palo Alto, CA). The volumes of interest, including the gross primary tumor volumes, were manually segmented by a radiation oncologist in three-dimensional fashion and then inspected by a second radiation oncologist. The generated volumes of interest and CT images were exported in the format of DICOM and DICOM-RTSTRUCT to be used for radiomics features extraction.

Radiomics Analysis

The gross primary tumor volumes were contoured on the basis of the International Commission on Radiation Units and Measurements 62/83 definition.[12] Radiomics analysis was performed using the freely available open-source software Imaging Biomarker Explorer, which uses the Matlab platform (Mathworks, Natick, VA). We extracted features that represent the intensity, shape, and texture. The categorization of these features was ranked as first, second, and higher texture features on the basis of the applied method from pixel to pixel.[13] More details about this process can be found in the Appendix and Appendix Table A1.

Data Processing

Figure 1 shows the overall processing pipeline, including the procedures for feature selection and evaluation. Features that have a high Spearman rank correlation (≥ 99%) with at least one other feature or with little variability were removed because they are not helpful in outcome prediction. Highly skewed features were log transformed.

To further select the radiomic features, we considered eight feature selection and extraction algorithms. Each feature selection algorithm assumes that the outcome of interest is binary, continuous, or time to event with censoring indicator. Preprocessing the data for these three different outcomes permits the use of many different feature selection algorithms beyond those for right-censored data. We considered three different ways of preprocessing the time-to-event outcome to be used in feature selection algorithms:

1. Binary outcome. The outcome is dichotomized on the basis of whether the event was experienced before 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.

2. Censor-incorporating outcome. The Martingale residual is computed from a Cox proportional hazards model. The Martingale residual can be thought of as the variability in the time-to-event outcome that is not explained by the clinical covariables included in the model. The multivariable Cox model adjusted for sex, age, tumor subsite, T stage, N stage, American Joint Committee on Cancer (seventh edition) stage, human papillomavirus (HPV) status, and smoking status. The Martingale residual is a continuous outcome.

3. Censored-aware outcome. The follow-up time and censoring indicator are used directly.

Each type of outcome was used with at least two different feature selection algorithms from the machine learning literature. A total of eight feature selection and extraction methods were applied to the data set, most of which are supervised. We also considered an unsupervised method, principal component analysis (PCA), as a result of its popularity in high-dimensional data analysis. For completeness, we also compared the performance of the feature selection methods with randomly selecting 10 radiomic features and using clinical features only. The feature selectors include minimum redundancy maximum relevance (MRMR), Wilcoxon rank sum test (Wilcoxon), RF, RReliefF, random regression forest (RRF), incremental association Markov blanket (IAMB), RSF, and PCA. Table 1 summarizes the algorithms and the type of outcome variable they require (binary, continuous, or time to event). Some of the methods can also be used as predictive models (noted in Table 1). More details can be found in the Appendix.

The selected radiomics features identified using the feature selection algorithms together with other relevant clinical features were used for outcome prediction. For this study, we considered estimating 5-year OS and RFS. All the patients included in this study had complete radiomic data for the primary tumor but may have been missing other clinical data. Missing values were imputed using multivariable imputation by chained equations[26] before evaluation. A third level, labeled unknown, was used for missing HPV values.

The predictive models used to estimate 5-year OS and RFS included logistic regression, Cox proportional hazards,[27] RF and RSF,[19] logistic elastic net,[28] and an ensemble (ie, 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 to extend machine learning methods for survival analysis.[10]

The following three different metrics were used to evaluate model performance: Harell's C-index, area under the curve (AUC), and Nam-D'Agostino calibration test statistic.[29] If the models are well calibrated, the calibration test statistic follows a χ^2 distribution with 8 *df*. Test statistics greater than 15.5 would indicate that the models are significantly miscalibrated at the $P = .05$ significance level. The AUC and C-index are measures of the predictive power of the learning model, where higher values indicate better predictive power. Ten-fold cross-validation was used for evaluation.

RESULTS

Data

Table 2 lists the demographic characteristics, clinical features, and OS and RFS outcomes of the 529 patients. The cohort was predominately male (87%), and the median age was 58 years (range, 21 to 88 years). Most cancers (87%) were stage IV according to American Joint Committee on Cancer staging. More than half of the cohort (58%) was HPV positive. Twenty percent of patients died during follow-up, and 18% experienced a relapse. Median follow-up time was 70 months. More than 3,800 radiomic features accompanied the clinical data for the patients.

Model Performance

Figure 2 shows heatmaps displaying the performance metrics for each feature selector and model pair in predicting OS and RFS. Comparing the different risk prediction methods, the ensemble model resulted in the best performance for many feature selectors across all metrics. Often AUC and C-index were greater than 0.7, and the calibration test statistic was less than 20. For all feature selectors with the ensemble model, the average C-index, AUC, and calibration scores for OS were 0.719, 0.696, and 16.69, respectively; the average scores for RFS were 0.655, 0.686, and 18.82, respectively. Logistic regression and RF models tended to result in poor calibration test statistics (typically > 20) independent of the feature selection method. The RSF predictions were comparable to RF predictions in AUC and C-index scores, with both following the ensemble, but calibration scores were poor for some selectors, more frequently for RFS than OS. Elastic net and Cox models had more acceptable values for calibration in general (with elastic net being better more consistently); however, the Cox model’s AUC and C-index scores were lower and comparable to those of the logistic model.

Considering the different radiomic feature selector algorithms, compared with only clinical data, all supervised methods selected a subset of features that generally resulted in higher AUC and C-index scores for both RFS and OS. For reference, using only clinical features resulted in AUC and C-index scores less than 0.70 for OS and 0.65 for RFS. From here on, we discuss the results of the different feature selection algorithms in the context of the ensemble risk prediction model because this was consistently the best-performing model. Random selection performed better than PCA and occasionally achieved fairly high AUC and C-index scores; however, it rarely achieved acceptable calibration scores, especially for RFS. The feature selection procedures that use a binary outcome (MRMR, Wilcoxon, and RF) all had similar C-index scores (0.66) and AUC scores (0.69 to 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 (0.72 to 0.73) and AUC (0.70 to 0.72). In general, the RF selectors (RF, RRF, and RSF) achieved the best overall scores. Occasionally, an RF selector would tie with another selector or be beat by a small margin (< 0.01 for AUC and C-index), and sometimes calibration scores were high, as was the case for RSF in predicting OS. Regardless, an RF-based selector always achieved the highest AUC and C-index scores with a reasonable calibration for both outcomes. RRF performed the best for OS (AUC, 0.75; C-index, 0.76; calibration, 8.7), and RSF performed the best for RFS (AUC, 0.71; C-index, 0.68; calibration, 19.1).

Effect of Censored Outcome Transformation

Table 3 lists the number of features selected by each method for the OS and RFS outcomes and a list of the features selected by at least two methods. The number of features selected ranged from one to 24 (mean, 10.1 features), with RF methods selecting the largest number of features. There was no significant overlap between the features selected by the different methods. The number of features selected was slightly larger for the RFS outcome (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. Neighbor Intensity Difference 25

Complexity 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.

Figure 3 shows the AUC and calibration metrics for the ensemble model using RF as the feature selection algorithm for binary outcome (RF), continuous outcome (RRF), and censored-aware outcome (RSF) for both OS and RFS. For comparison, we also include random selection of 10 features, PCA, and clinical only results. As can be seen, for OS, the RRF exhibits both the best calibration and best AUC. For RFS, RSF shows the best AUC with a reasonable calibration (albeit > 15.5).

DISCUSSION

The ensemble model consistently made the best predictions across all selectors considering all three metrics for both outcomes. Combining clinical features with radiomic features improves predictions; however, unsupervised feature selection results in miniscule improvement, bad calibration, or both. In general, RF-based selectors 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) achieved the highest predictive power. The different outcomes, OS and RFS, did not significantly affect the number of features selected in total by all of the methods; however, the RFS outcome resulted in more overlap in features selected between the feature selection methods, although neither outcome resulted in a large overlap in total.

We observed similar results to other previous studies. A few of the selected features (eg, F29. IntensityDirectGlobalMax) were also selected in another study,[2] which indicates that these features in particular have predictive value and may be enhanced by the inclusion of other nonredundant radiomic features. As discussed elsewhere,[7] where only the binary outcome was considered, MRMR and Wilcoxon performed fairly well, with MRMR performing slightly better depending on the model. As presented in Leger et al,[6] where only the right-censored outcome was considered, RF feature selectors performed well in predicting outcome 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.

Supervised methods should be preferred over unsupervised ones such as PCA because the metric scores are consistently better and resulting features can be interpreted more easily. There was significant variability in the number of selected features among the supervised methods; a few methods selected few features. Those feature selectors that selected fewer features, such as incremental association Markov blanket and RReliefF, which both selected less than five features, performed consistently worse than the other selectors. Among all feature selectors and considering all prediction models, the RF, RRF, and RSF feature selectors provided the best predictions for OS and RFS (and consistently selected higher numbers of features). In particular, of the three, RRF and RSF selected the most predictive radiomic signatures. In addition, 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.

Limitations of this study include the use of 10-fold cross-validation for evaluation. Features were selected on the same set of patients for which the learning models were applied because of the small number of patients with radiomic data and large number of radiomic features. However, because we compared the performance among the feature selectors, and each feature selector was informed by the same set of patients, we do not expect that any feature selector had an advantage over others. In addition, the proportion of censored patients was high for both outcomes. As more patient data are collected and the amount of censoring changes, we will be able to evaluate outcome transformation across various amounts of censoring.

Finally, in the initial pruning of radiomic features, we kept features that were less than 99% correlated with another feature. The value of 99% was chosen as a conservative threshold; however, the number of selectable features remained large, and many highly correlated features still remained in the data set. Although different feature selection algorithms select different features for PFS and OS, these may be in fact highly correlated features. With removal of features with ≥ 90% correlation, more than 90% of the raw features can be removed. However, this could result in lower performance if relevant features are pruned. Combining the pruned features can minimize this information loss.

This study highlights how different feature selection algorithms can be used when the outcome of interest is right-censored and creates a framework for future radiomic applications. Although the development of new methods for extracting quantitative data from imaging features is beyond the scope of this article, the methods evaluated herein can guide the selection of the most informative features for a particular outcome. In particular, the RF group exhibited the highest predictive scores, with RRF (continuous outcome) for OS and RSF (censored-aware outcome) for RFS exhibiting the best predictive performance. For data sets with a substantial amount of censoring, we advocate using the censor-incorporating and censored-aware outcomes with RRF and RSF, respectively.

Acknowledgments

Support

Supported in part by the National Institutes of Health (NIH) National Cancer Institute (NCI) Big Data to Knowledge Program under Grant No. R01CA214825 and a joint National Science Foundation/NIH Initiative on Quantitative Approaches to Biomedical Big Data (Grant No. R01CA225190). Also supported by the Andrew Sabin Family Foundation; C.D.F. is a Sabin Family Foundation Fellow. C.D.F. is supported by the NIH, including the National Institute for Dental and Craniofacial Research Award (Grant No. 1R01DE025248–01/R56DE025248–01), NCI Early Phase Clinical Trials in Imaging and Image-Guided Interventions Program (Grant No. 1R01CA218148–01), an NIH/NCI Cancer Center Support Grant (CCSG) Pilot Research Program Award from The University of Texas MD Anderson Cancer Center CCSG Radiation Oncology and Cancer Imaging Program (Grant No. P30CA016672), and an NIH/NCI Head and Neck Specialized Programs of Research Excellence Developmental Research Program Award (Grant No. P50 CA097007–10). G.E.M. is supported by National Science Foundation Award No. CNS-1625941, NIH National Library of Medicine Grant No. R01LM012527, and The Joseph and Bessie Feinberg Foundation. H.E. is supported in part by the philanthropic donations from the Family of Paul W. Beach to G. Brandon Gunn, MD. B.A.E. is on an Egyptian-American conjoint PhD program funded by the Egyptian Cultural and Educational Bureau.

Appendix

Radiomics Analysis

Each patient's head and neck contrast-enhanced computed tomography (CT) image was identified and individually checked. For each patient, the primary tumor volume was identified by two expert radiation oncologists to whom the relevant clinical data were not released. The gross tumor volumes (GTVs) were contoured on the basis of the International Commission on Radiation Units and Measurements 62/83 definition of the gross tumor representing the gross demonstrable extent and location of the tumor.[12] The common ontology used to represent the primary (p) tumor volumes was GTVp. The manual segmentation of the GTVp was done using the commercial treatment planning software Velocity AI 3.0.1 (powered by VelocityGrid; Varian Medical Systems, Palo Alto, CA). In addition, the contours were done with the guidance of the findings from the physical examination, endoscopic examination, and other radiology such as magnetic resonance imaging and positron emission tomography. Slices with metal artifacts that did not allow an accurate identification of the GTVp were omitted. Then, CT images with GTVp generated were extracted in the format of Digital Imaging and Communications and Medicine (DICOMRT). Radiomics analysis was performed using the freely available open-source software Imaging Biomarker Explorer (IBEX), which was developed by The University of Texas MD Anderson Cancer Center and uses the Matlab platform (Mathworks, 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 on the basis of the applied method from pixel to pixel.[13] The intensity values (Hounsfield units) and shape of the region of interest 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 gray 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 co-occurrence matrix, gray level run length matrix, and neighbor intensity difference.[13] 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, and z dimensions of the voxel size were set to 0.488, 0.488, and 1 mm, respectively. The calculation of intensity-based features was preceded by applying the Laplacian of Gaussian filter. The standard deviation (sigma) of the Laplacian of Gaussian filter ranged from 0.5 to 2.5 voxels for a total of five iterations (Ganeshan B, et al: Clin Radiol 67:157–164, 2012). The Butterworth smoothing preprocessing filter was applied before extracting the intensity and texture features to calculate the impact of smoothing and noise removal on the radiomics features. The regions of interest were fitted to 512 × 512 pixels when applying the Butterworth filters. The uniformity of voxel size was ensured by applying the two-dimensional Butterworth filters with the three-dimensional voxel size before the smoothing process. More of these statistical texture features, along with their relevant equations, were illustrated by Davnall et al (Insights Imaging 3:573–589, 2012). More details describing the data can be found in a report by the M.D. Anderson Cancer Center Head and Neck Quantitative Imaging Working Group.[2] Appendix Table A1 lists the radiomic features, their categories, and their definitions.

Feature Selection

Here, we briefly describe each feature selection method and any relevant parameters and implementation details. The minimum redundancy maximum relevance feature selector is frequently used in gene expression experiments.[14] It seeks to find a subset of features that 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, however, is maximized and is defined by the following 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 covariables or a covariable 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[15] is used for selecting the features. We specify 10 features as the number of features to select because this is near the average number of features selected by the other methods.

The Wilcoxon rank sum test, also known as the Mann-Whitney U test, is a statistical test for feature importance and does not assume a normal distribution of the data.[16] We run Wilcoxon with 10 different splits of the data set. We select the features by splitting the data set 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 that 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 five times, feature B appears four times, feature C appears one time, and all other features appear zero times, then only features A and B will be selected because the largest jump in number of occurrences happens between B and C. Boulesteix[17] provides the R package, WilcoxCV, which we use for the Mann-Whitney U test with cross-validation.

A random forest[18] 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 data set, 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[18] and random survival forests,[24] 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[20] is used for feature selection with all of the random forests. The random forests are over five Monte Carlo iterations. All of the random forests are implemented with the R package randomForestSRC.[19]

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 that fall in the same and different classes.[21] 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 is able to calculate feature importance on the basis of 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:

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

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.[21]

The incremental association Markov blanket (IAMB)[22] feature selector finds a subset of features that excludes those independent of the target outcome. IAMB works in the following two phases: a growth phase and shrink phase. The growth phase adds independent features on the basis of 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 use the R package MXM,[23] 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. PCA transforms the set of features into a set of components that are uncorrelated and thus can reduce dimensionality.[25] We do not desire every component because 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 small number of components compared with the actual number of features. Because with this dimensionality reduction technique, the feature space is transformed, it is not as clear which features are indicative of the outcome; thus, interpretation of feature importance is not as straightforward when using PCA compared with the other methods, which return a subset of the original features. No features are log transformed before extraction, as in the other methods; however, all features are scaled and centered.

Table A1.

Computed Tomography–Derived Intensity Histogram, Shape, and Texture Features Set

Feature Category and Feature	Definition
Gray level co-occurrence matrix 25	
Auto-correlation	The correlation texture measures the linear dependency of gray levels on those of neighboring pixels[*]
Cluster prominence	A measure of the skewness or asymmetry[*]
Cluster shade	A measure of the skewness or asymmetry[*]
Cluster tendency	Assess whether nonrandom structure exists in the data by measuring the probability that the data are generated by a uniform data distribution[*]
Contrast	Returns a measure of the intensity contrast between a pixel and its neighbor over the whole image[†‡]
Correlation	Returns a measure of how correlated a pixel is to its neighbor over the whole image[†‡]
Difference entropy[3][†]	
Dissimilarity[*]	
Energy[†‡]	
Entropy[*]	
Homogeneity[3][†‡]	
Homogeneity 2[3][†‡]	
Information measure correlation 1[3][†‡]	
Information measure correlation 2[3][†‡]	
Inverse diff moment norm[3][†‡]	
Inverse diff norm[3][†‡]	
Inverse variance[3]	
Max probability[*]	
Sum average[3][†‡]	
Sum entropy[3][†‡]	
Sum variance[3][†‡]	
Variance[3]	
Gray level run length matrix 25[§]	
Gray level nonuniformity	
High gray level run emphasis	
Long run emphasis	
Long run high gray level emphasis	
Long run low gray level emphasis	
Low gray level run emphasis	

Feature Category and Feature	Definition
Run length nonuniformity	
Run percentage	
Short run emphasis	
Short run high gray level emphasis	
Short run low gray level emphasis	
Neighbor intensity difference 25[∥]	
Busyness	
Coarseness	
Complexity	
Contrast	
Texture strength	
Intensity direct[3]	
Energy	
Global entropy	The intensity entropy among all the voxels
Global max	The intensity maximum among all the voxels
Global mean	The intensity mean among all the voxels
Global median	The intensity median among all the voxels
Global min	The intensity minimum among all the voxels
Global std	The intensity standard deviation among all the voxels
Global uniformity	The intensity uniformity among all the voxels
Interquartile range	The interquartile range of the intensity values among all the voxels
Kurtosis	Measures the peakedness of all the voxels' intensities
Local entropy max	First, at each voxel, compute entropy in its neighborhood region. Then, compute the maximum among all the voxels' entropies calculated from step 1.
Local entropy mean	First, at each voxel, compute entropy in its neighborhood region. Then, compute the mean among all the voxels' entropies calculated from step 1.
Local entropy median	First, at each voxel, compute entropy in its neighborhood region. Then, compute the median among all the voxels' entropies calculated from step 1.
Local entropy min	First, at each voxel, compute entropy in its neighborhood region. Then, compute the minimum among all the voxels' entropies calculated from step 1.
Local entropy std	First, at each voxel, compute entropy in its neighborhood region. Then, compute the standard deviation among all the voxels' entropied calculated from step 1.
Local range max	First, at each voxel, compute range value (Max Value-Min Value) in its neighborhood region. Then, compute the maximum among all the voxels' range values calculated from step 1.
Local range mean	First, at each voxel, compute range value (Max Value-Min Value) in its neighborhood region. Then, compute the mean among all the voxels' range values calculated from step 1.
Local range median	First, at each voxel, compute range value (Max Value-Min Value) in its neighborhood region. Then, compute the median among all the voxels' range values calculated from step 1.
Local range min	First, at each voxel, compute range value (Max Value-Min Value) in its neighborhood region. Then, compute the minimum among all the voxels' range values calculated from step 1.

Feature Category and Feature	Definition
Local range std	First, at each voxel, compute range value (Max Value-Min Value) in its neighborhood region. Then, compute the standard deviation among all the voxels' range values calculated from step 1.
Local std max	First, at each voxel, compute standard deviation in its neighborhood region. Then, compute the maximum among all the voxels' standard deviation values calculated from step 1.
Local std mean	First, at each voxel, compute standard deviation in its neighborhood region. Then, compute the mean among all the voxels' standard deviation values calculated from step 1.
Local std median	First, at each voxel, compute standard deviation in its neighborhood region. Then, compute the median among all the voxels' standard deviation values calculated from step 1.
Local std min	First, at each voxel, compute standard deviation in its neighborhood region. Then, compute the minimum among all the voxels' standard deviation values calculated from step 1.
Local std std	First, at each voxel, compute standard deviation in its neighborhood region. Then, compute the standard deviation among all the voxels' standard deviation values calculated from step 1.
Mean absolute deviation	The mean absolute deviation of the intensity values among all the voxels
Median absolute deviation	The median absolute deviation of the intensity values among all the voxels
Percentile	Percentiles of the intensity values among all the voxels
Quantile	Quantiles of the intensity values among all the voxels
Range	The intensity range (Max Value-Min Value) among all the voxels
Root mean square	
Skewness	Measures the asymmetry of all the voxels' intensity
Variance	
Intensity histogram[3]	
Interquartile range	The interquartile range of the occurrence probability values in the histogram
Kurtosis	Measures the peakedness of the occurrence probability values in the histogram
Mean absolute deviation	The mean absolute deviation of the occurrence probability values in the histogram
Median absolute deviation	The median absolute deviation of the occurrence probability values in the histogram
Percentile	Percentiles of the occurrence probability values in the histogram
Percentile area	Percentiles of values in the accumulative histogram
Quantile	Quantiles of the occurrence probability values in the histogram
Range	Measures the range (Max Value-Min Value) of the occurrence probability values in the histogram.
Skewness	Measures the asymmetry of the occurrence probability values in the histogram
Shape	
Compactness 1	Compactness1 = (Volume)/(sqrt(pi) *(SurfaceArea)^(2/3))[3]
Compactness 2	Compactness2 = 36 *pi *(Volume^2)/((SurfaceArea)^3)[3]
Convex	Measures the proportion of the pixels in the convex hull that are also in the region[3]
Convex hull volume	The mean volume of the two-dimensional convex hulls that are the convex envelopes of each slice's binary mask[3]
Convex hull volume 3D	Three-dimensional volume of the convex hull that is the convex envelope of binary mask[3]
Mass[3]	

Feature Category and Feature	Definition
Max 3D diameter	Largest pairwise Euclidean distance between voxels on the surface of the tumor volume[3]
Mean breadth	Denotes integral of mean curvature[3]
Number of voxels	The number of voxels treating the edge voxels differently[3]
Orientation	Measures the angle between the x-axis and the major axis of the ellipse in two dimensions[3]
Roundness	Measures how much the binary mask is close to circle in two dimensions[3]
Spherical disproportion[3]	
Sphericity[3]	
Surface areail[¶]	The surface area of the binary mask
Surface area density	Surface Area Density = (surface area of the binary mask)/(volume of the binary mask)[3][¶]
Volume	The physical volume treating the edge voxels differently[¶]

[*] Soh LK, Tsatsoulis C: Texture analysis of SAR sea ice imagery using gray level co-occurrence matrices. IEEE Trans Geosci Remote Sens 37:780–795, 1999

[†] Haralick RM, Shanmugam K, Dinstein I: Textural features for image classification. IEEE Trans Syst Man Cybern 3:610–621, 1973

[‡] Haralick RM, Shapiro LG: Computer and Robot Vision. Boston, MA, Addison-Wesley Longman Publishing, 1992

[§] Tang X: Texture information in run-length matrices. IEEE Trans Image Process 7:1602–1609, 1998

[‖] Amadasun M, King R: Textural features corresponding to textural properties. IEEE Trans Syst Man Cybern 19:1264–1274, 1989

[¶] Legland D, Kiêu K, Devaux M-F: Computation of Minkowski measures on 2D and 3D binary images. Image Anal Stereol 26:83–92, 2007

REFERENCES

1. Panth KM, Leijenaar RT, Carvalho S, et al.: Is there a causal relationship between genetic changes and radiomics-based image features? An in vivo preclinical experiment with doxycycline inducible GADD34 tumor cells. Radiother Oncol 116:462–466, 2015 [PubMed: 26163091]
2. Anderson MD Cancer Center Head and Neck Quantitative Imaging Working Group: Investigation of radiomic signatures for local recurrence using primary tumor texture analysis in oropharyngeal head and neck cancer patients. Sci Rep 8:1524, 2018 [PubMed: 29367653]
3. Aerts HJ, Velazquez ER, Leijenaar RT, et al.: Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat Commun 5:4006, 2014 [PubMed: 24892406]
4. Huang Y, Liu Z, He L, et al.: Radiomics signature: A potential biomarker for the prediction of disease-free survival in early-stage (I or II) non-small cell lung cancer. Radiology 281:947–957, 2016 [PubMed: 27347764]
5. Wong J, Kanwar A, Mohamed AS, et al.: Radiomics in head and neck cancer: From exploration to application. Transl Cancer Res 5:371–382, 2016 [PubMed: 30627523]
6. Leger S, Zwanenburg A, Pilz K, et al.: A comparative study of machine learning methods for time-to-event survival data for radiomics risk modelling. Sci Rep 7:13206, 2017 [PubMed: 29038455]
7. Parmar C, Grossmann P, Rietveld D, et al.: Radiomic machine-learning classifiers for prognostic biomarkers of head and neck cancer. Front Oncol 5:272, 2015 [PubMed: 26697407]
8. Royston P, Altman DG: External validation of a Cox prognostic model: Principles and methods. BMC Med Res Methodol 13:33, 2013 [PubMed: 23496923]
9. Therneau TM, Grambsch PM, Fleming TR: Martingale-based residuals for survival models. Biometrika 77:147–160, 1990

10. Vock DM, Wolfson J, Bandyopadhyay S, et al.: Adapting machine learning techniques to censored time-to-event health record data: A general-purpose approach using inverse probability of censoring weighting. J Biomed Inform 61:119–131, 2016 [PubMed: 26992568]
11. Wang H, Schabath MB, Liu Y, et al.: Semiquantitative computed tomography characteristics for lung adenocarcinoma and their association with lung cancer survival. Clin Lung Cancer 16:e141–e163, 2015 [PubMed: 26077095]
12. 4. Definition of volumes. J ICRU 10:41–53, 2010 [PubMed: 24173326]
13. Haralick RM: Statistical and structural approaches to texture. Proc IEEE 67:786–804, 1979
14. Ding C, Peng H: Minimum redundancy feature selection from microarray gene expression data. J Bioinform Comput Biol 3:185–205, 2005 [PubMed: 15852500]
15. De Jay N, Papillon-Cavanagh S, Olsen C, et al.: mRMRe: An R package for parallelized mRMR ensemble feature selection. Bioinformatics 29:2365–2368, 2013 [PubMed: 23825369]
16. Wang Y, Cheung Y-M, Liu H (eds): Gene selection using Wilcoxon rank sum test and support vector machine for cancer classification, in Computational Intelligence and Security. Berlin, Germany, Springer, 2007, pp 57–66
17. Boulesteix A-L: WilcoxCV: An R package for fast variable selection in cross-validation. Bioinformatics 23:1702–1704, 2007 [PubMed: 17495999]
18. Breiman L: Random forests. Mach Learn 45:5–32, 2001
19. Ishwaran H, Kogalur U: Random Forests for Survival, Regression, and Classification (RF-SRC), r package version 2.5.1. https://cran.r-project.org/package=randomForestSRC
20. Ishwaran H, Kogalur UB, Gorodeski EZ, et al.: High-dimensional variable selection for survival data. J Am Stat Assoc 105:205–217, 2010
21. Robnik-Šikonja M, Kononenko I: Theoretical and empirical analysis of ReliefF and RReliefF. Mach Learn 53:23–69, 2003
22. Tsamardinos I, Aliferis CF, Statnikov A: Algorithms for large scale Markov blanket discovery. FLAIRS Conf 2:376–380, 2003
23. Lagani V, Athineou G, Farcomeni A, et al.: Feature selection with the R package MXM: Discovering statistically equivalent feature subsets. J Stat Softw 80:1–25, 2017
24. Ishwaran H, Kogalur UB, Blackstone EH, et al.: Random survival forests. Ann Appl Stat 2:841–860, 2008
25. Wold S, Esbensen K, Geladi P: Principal component analysis. Chemom Intell Lab Syst 2:37–52, 1987
26. van Buuren S, Groothuis-Oudshoorn K: mice: Multivariate imputation by chained equations in R. J Stat Softw 45:1–67, 2011
27. Therneau TM: A Package for Survival Analysis in S, version 2.38. https://CRAN.R-project.org/package=survival
28. Friedman J, Hastie T, Tibshirani R: Regularization paths for generalized linear models via coordinate descent. J Stat Softw 33:1–22, 2010 [PubMed: 20808728]
29. Steyerberg EW, Vickers AJ, Cook NR, et al.: Assessing the performance of prediction models: A framework for traditional and novel measures. Epidemiology 21:128–138, 2010 [PubMed: 20010215]

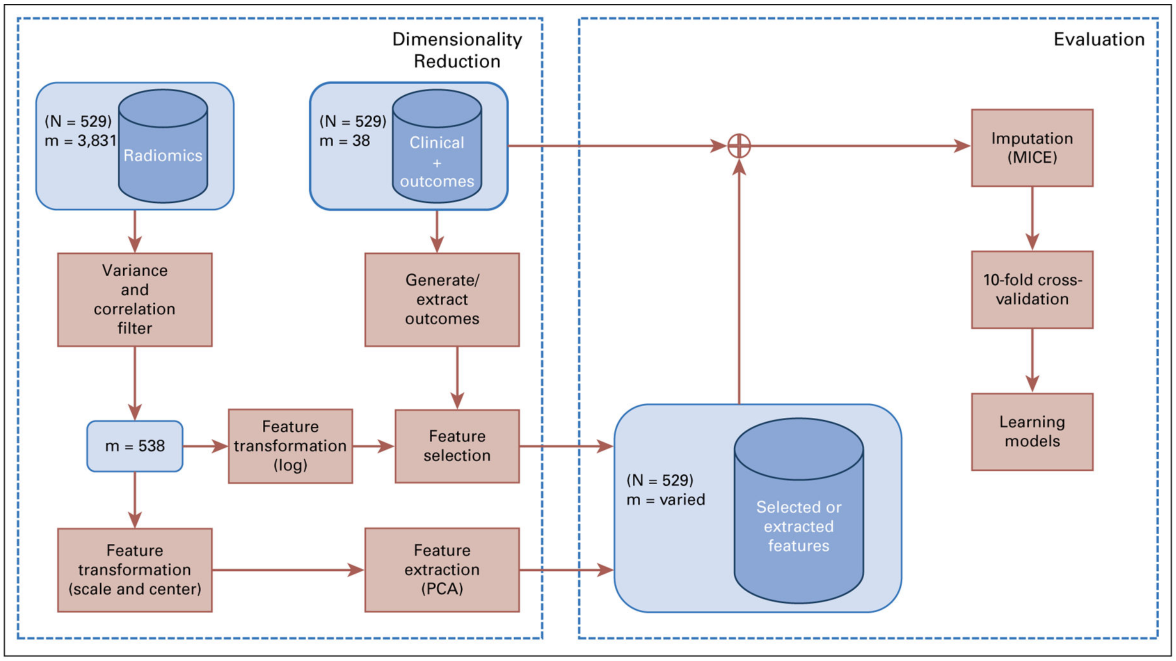
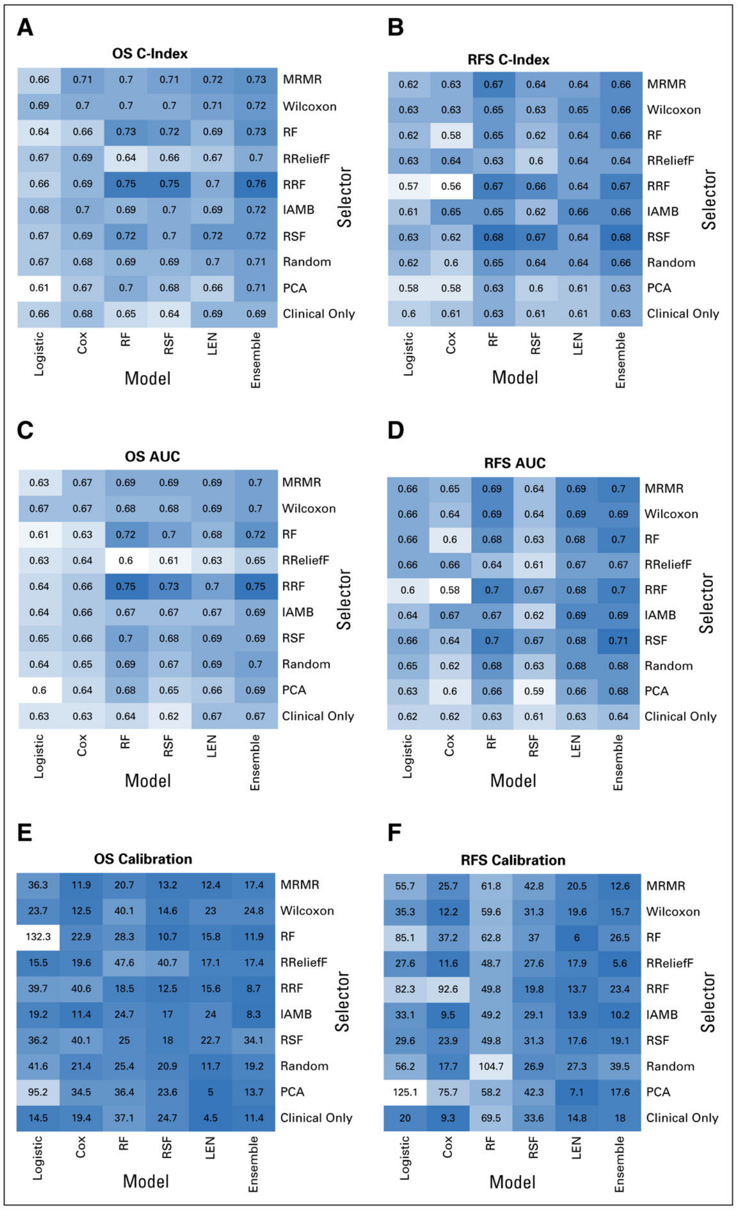


Fig 1.

Pipeline for identification of radiomic signatures. Low-variance features and high-correlated features are pruned. Clinical features are appended to selected features, and missing data are imputed before evaluation with the learning models. m, radiomic features; MICE, multivarible imputation by chained equations; PCA, principal component analysis.



A OS C-Index

Selector	Logistic	Cox	RF	RSF	LEN	Ensemble
MRMR	0.66	0.71	0.7	0.71	0.72	0.73
Wilcoxon	0.69	0.7	0.7	0.7	0.71	0.72
RF	0.64	0.66	0.73	0.72	0.69	0.73
RReliefF	0.67	0.69	0.64	0.66	0.67	0.7
RRF	0.66	0.69	0.75	0.75	0.7	0.76
IAMB	0.68	0.7	0.69	0.7	0.69	0.72
RSF	0.67	0.69	0.72	0.7	0.72	0.72
Random	0.67	0.68	0.69	0.69	0.7	0.71
PCA	0.61	0.67	0.7	0.68	0.66	0.71
Clinical Only	0.66	0.68	0.65	0.64	0.69	0.69

B RFS C-Index

Selector	Logistic	Cox	RF	RSF	LEN	Ensemble
MRMR	0.62	0.63	0.67	0.64	0.64	0.66
Wilcoxon	0.63	0.63	0.65	0.63	0.65	0.66
RF	0.62	0.58	0.65	0.62	0.64	0.66
RReliefF	0.63	0.64	0.63	0.6	0.64	0.64
RRF	0.57	0.56	0.67	0.66	0.64	0.67
IAMB	0.61	0.65	0.65	0.62	0.66	0.66
RSF	0.63	0.62	0.68	0.67	0.64	0.68
Random	0.62	0.6	0.65	0.64	0.64	0.66
PCA	0.58	0.58	0.63	0.6	0.61	0.63
Clinical Only	0.6	0.61	0.63	0.61	0.61	0.63

C OS AUC

Selector	Logistic	Cox	RF	RSF	LEN	Ensemble
MRMR	0.63	0.67	0.69	0.69	0.69	0.7
Wilcoxon	0.67	0.67	0.68	0.68	0.69	0.7
RF	0.61	0.63	0.72	0.7	0.68	0.72
RReliefF	0.63	0.64	0.6	0.61	0.63	0.65
RRF	0.64	0.66	0.75	0.73	0.7	0.75
IAMB	0.64	0.66	0.67	0.67	0.67	0.69
RSF	0.65	0.66	0.7	0.68	0.69	0.69
Random	0.64	0.65	0.69	0.67	0.69	0.7
PCA	0.6	0.64	0.68	0.65	0.66	0.69
Clinical Only	0.63	0.63	0.64	0.62	0.67	0.67

D RFS AUC

Selector	Logistic	Cox	RF	RSF	LEN	Ensemble
MRMR	0.66	0.65	0.69	0.64	0.69	0.7
Wilcoxon	0.66	0.64	0.69	0.64	0.69	0.69
RF	0.66	0.6	0.68	0.63	0.68	0.7
RReliefF	0.66	0.66	0.64	0.61	0.67	0.67
RRF	0.6	0.58	0.7	0.67	0.68	0.7
IAMB	0.64	0.67	0.67	0.62	0.69	0.69
RSF	0.66	0.64	0.7	0.67	0.68	0.71
Random	0.65	0.62	0.68	0.63	0.68	0.68
PCA	0.63	0.6	0.66	0.59	0.66	0.68
Clinical Only	0.62	0.62	0.63	0.61	0.63	0.64

E OS Calibration

Selector	Logistic	Cox	RF	RSF	LEN	Ensemble
MRMR	36.3	11.9	20.7	13.2	12.4	17.4
Wilcoxon	23.7	12.5	40.1	14.6	23	24.8
RF	132.3	22.9	28.3	10.7	15.8	11.9
RReliefF	15.5	19.6	47.6	40.7	17.1	17.4
RRF	39.7	40.6	18.5	12.5	15.6	8.7
IAMB	19.2	11.4	24.7	17	24	8.3
RSF	36.2	40.1	25	18	22.7	34.1
Random	41.6	21.4	25.4	20.9	11.7	19.2
PCA	95.2	34.5	36.4	23.6	5	13.7
Clinical Only	14.5	19.4	37.1	24.7	4.5	11.4

F RFS Calibration

Selector	Logistic	Cox	RF	RSF	LEN	Ensemble
MRMR	55.7	25.7	61.8	42.8	20.5	12.6
Wilcoxon	35.3	12.2	59.6	31.3	19.6	15.7
RF	85.1	37.2	62.8	37	6	26.5
RReliefF	27.6	11.6	48.7	27.6	17.9	5.6
RRF	82.3	92.6	49.8	19.8	13.7	23.4
IAMB	33.1	9.5	49.2	29.1	13.9	10.2
RSF	29.6	23.9	49.8	31.3	17.6	19.1
Random	56.2	17.7	104.7	26.9	27.3	39.5
PCA	125.1	75.7	58.2	42.3	7.1	17.6
Clinical Only	20	9.3	69.5	33.6	14.8	18

Fig 2.

Heatmaps for each of the different feature selection and learning models showing C-index for (A) overall survival (OS) and (B) relapse-free survival (RFS), area under the curve (AUC) for (C) OS and (D) RFS, and calibration for (E) OS and (F) RFS. Darker color indicates a better score for all metrics. IAMB, incremental association Markov blanket; LEN, logistic elastic net; MRMR, minimum redundancy maximum relevance; PCA, principal component analysis; RF, random forest; RRF, random regression forest; RSF, random survival forest.

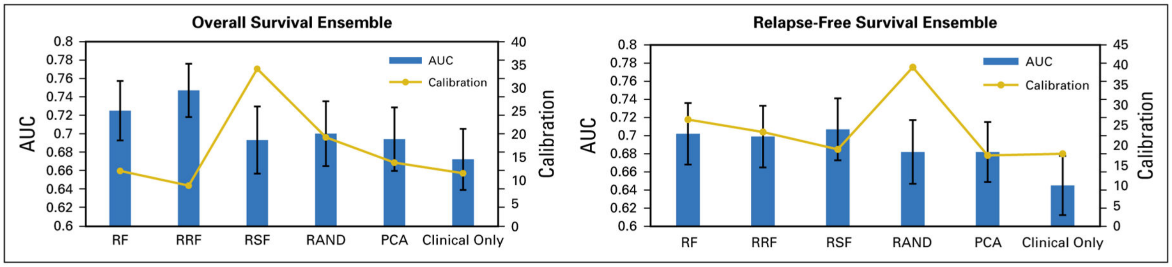


Fig 3.

Area under the curve (AUC) values with SEs and calibration values for the ensemble model when different feature selectors are used for overall survival and relapse-free survival. PCA, principal component analysis; RAND, random selection of features; RF, random forest; RRF, random regression forest; RSF, random survival forest.

Table 1.

Feature Selection Methods Summary

Feature Selector or Extractor	Abbreviation	Supervised	Outcome Type	Classifier
Minimum redundancy maximum relevance[14,15]	MRMR	Y	B	N
Wilcoxon rank sum test[16,17]	Wilcoxon	Y	B	N
Random forest[18–20]	RF	Y	B	Y
RReliefF[21]	RReliefF	Y	R	N
Random regression forest[18–20]	RRF	Y	R	Y
Incremental association Markov blanket[22,23]	IAMB	Y	C	N
Random survival forest[19,20,24]	RSF	Y	C	Y
Principal component analysis[25]	PCA	N	—	N

Abbreviations: B, binary; C, censored aware; N, no; R, continuous; Y, yes.

Table 2.

Demographic and Disease Characteristics of the Patients With Oropharyngeal Cancer

Characteristic	Patients (N = 529)
Sex, no. (%)	
Male	462 (87)
Female	67 (13)
Age at diagnosis, years	
Median (range)	58.1 (21–88)
25th-75th percentile	52–65
T category, No. (%)	
Tl/2	329 (62)
T3/4	200 (38)
N category, No. (%)	
<N2b	120 (23)
≥ N2b	409 (77)
AJCC stage (seventh edition), No. (%)	
I	2 (<1)
II	8 (2)
III	59 (11)
IV	459 (87)
Smoking, packs per year	
Median (range)	5 (0–120)
25th-75th percentile	0–30
Smoking status, No. (%)	
Former	191 (36)
Current	117 (22)
Never	221 (43)
Tumor subsite, No. (%)	
Tonsil	199 (38)
Base of tongue	285 (54)

Characteristic	Patients (N = 529)
Other	45 (8)
HPV status, No. (%)	
Positive	307 (58)
Negative	49 (9)
Unknown	173 (33)
Cancer treatment, No. (%)	
Concurrent chemoradiation	282 (53)
Induction chemotherapy followed by concurrent chemoradiation	141 (27)
Induction chemotherapy followed by radiation alone	53 (10)
Radiation alone	53 (10)
Vital status at last follow-up, No. (%)	
Alive	423 (80)
Deceased	106 (20)
Overall survival, months	
Median (range)	70.5 (1.10–148.37)
25th-75th percentile	47.37–99.77
Relapse-free survival at last follow-up	
Yes	435 (82)
No	94 (18)
Relapse-free survival, months	
Median (range)	64 (1.10–144.37)
25th-75th percentile	40.57–97.80
Local control at last follow-up	
Yes	483 (91)
No	46 (9)
Local control, months	
Median (range)	67.47 (1.10–148.37)
25th-75th percentile	44.03–98.37

Abbreviations: AJCC, American Joint Committee on Cancer; HPV, human papillomavirus.

Table 3.

Features Selected by the Different Methods for OS and RFS and Features Selected by at Least Two Different Methods

	No. of Features Selected		
Method and Outcome Type	**OS**	**RFS**	**Features Selected by Two or More Methods**
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	

Abbreviations: B, binary; C, censored aware; IAMB, incremental association Markov blanket; MRMR, minimum redundancy maximum relevance; OS, overall survival; R, continuous; RF, random forest; RFS, relapse-free survival; RRF, random regression forest; RSF, random survival forest.