Outcomes after Thoracic Surgery for Patients with Lung Cancer

Capstone 2
V. Moore

Introduction

Lung cancer is a disease with a high rate of mortality, though the factors associated with success after surgery for tumor removal are unclear, particularly as a patient's health history can influence outcomes in ways that may not be readily apparent. This can be an important problem for patients and caregivers as they consider the risks of surgery.

Since thoracic surgery is not without its own inherent risks, it could be beneficial to patients and care providers to have some insight into expected risk/benefit scenarios as they may relate to a patient's own health history. Health policy organizations and payers could also benefit from such insight for planning for optimal health outcomes.

Data acquisition and preprocessing

Data used for this machine learning project will come from the University of California, Irvine, Machine Learning Repository, accessible here. This dataset includes one-year mortality outcomes for 470 patients, along with data for 16 features in each patient's health history. Patient data were collected from the Wroclaw Thoracic Surgery Centre in Poland from patients who experienced tumor resections for primary lung cancer at the Department of Thoracic Surgery of the Medical University of Wroclaw and Lower-Silesian Centre for Pulmonary Diseases, Poland, from 2007-2011. Data are part of Poland's National Lung Cancer Registry.

Table 1 shows the structure of the dataset used for this analysis. The features are found in all columns except for the final column, "Risk1Y", which represents the target variable. The first feature, "DGN", represents diagnosis codes regarding the primary tumor or tumors for each patient, as combinations of ICD-10 codes, and they are here identified as DGN1, DGN2, DGN3, DGN4, DGN5, DGN6, or DGN8. Forced vital capacity is shown in the "FVC" column, and this is a continuous non-integer numeric variable. Forced expiratory volume is shown in the "FEV1" column, and this, too, is a continuous non-integer numeric variable. FEV1 is measured as a part of the pulmonary function test, described here, used to measure FVC; FVC is the total volumetric capacity with one breath, while FEV1 is the portion involved in expiration over the course of the first second. Values on the Zubrod scale are shown in the next column, and this scale represents a patient's performance status, with 0 being asymptomatic and worsening symptomatic states reflected in higher numbers, with 4 representing a bedbound patient and 5 representing death. This study includes only those patients with Zubrod scores of 0, 1, or 2 (perhaps relating to fitness for major surgery).

Pain represents whether the patient experienced pain prior to surgery, and haemoptysis (generally spelled "hemoptysis" in the United States), dyspnea, cough, and weakness also refer to Boolean values

for presence in a patient prior to surgery. Tumor size refers to size of the original tumor, with OC11 being the smallest, ascending to OC14 for the largest, with OC12 and OC13 being intermediate.

Type II diabetes mellitus ("T2DM"), myocardial infarction ("MI"), peripheral arterial disease ("PAD"), and asthma are all shown as Boolean variables for presence or absence of these medical conditions prior to surgery. Smoking is present as a Boolean as well, and age refers to age at time of surgery.

The target variable, "Risk1Y", is present as a Boolean variable with "T" occurring if the patient died during the one-year period following surgery.

Table 1. Dataset structure for features tested in this analysis

	DGN	FVC	FEV1	Zubrod	Pain	Haemoptysis	Dyspnea	Cough	Weakness	Tumor_size	T2DM	MI	PAD	Smoking	Asthma	Age	Risk1Y
0	DGN2	2.88	2.16	PRZ1	F	F	F	T	T	OC14	F	F	F	T	F	60	F
1	DGN3	3.40	1.88	PRZ0	F	F	F	F	F	OC12	F	F	F	T	F	51	F
2	DGN3	2.76	2.08	PRZ1	F	F	F	T	F	OC11	F	F	F	T	F	59	F
3	DGN3	3.68	3.04	PRZ0	F	F	F	F	F	OC11	F	F	F	F	F	54	F
4	DGN3	2.44	0.96	PRZ2	F	T	F	T	T	OC11	F	F	F	T	F	73	T

The dataset originally contained no column names, so those were added. There were no missing values for these variables in this population of n=470. The Boolean variables appearing as "T" or "F" were replaced with a 0 for each "F" entry and a 1 for each "T" entry. The values that contained a combination of letters and numbers had the letters stripped from them so that all remained in those columns were the numerical portions of the values. All features became set as integers apart from FVC, FEV1, and age, which were all cast as floats.

Exploratory data analysis

The first step in analysis of this dataset involving thoracic surgery outcomes for patients with lung cancer was to examine histograms of all feature and target data in order to assess distributions and be able to understand the overall nature of the data. Many distributions were somewhat imbalanced, and, importantly, the target data about mortality were imbalanced. Figure 1 shows a thumbnail image of the histograms from this dataset.

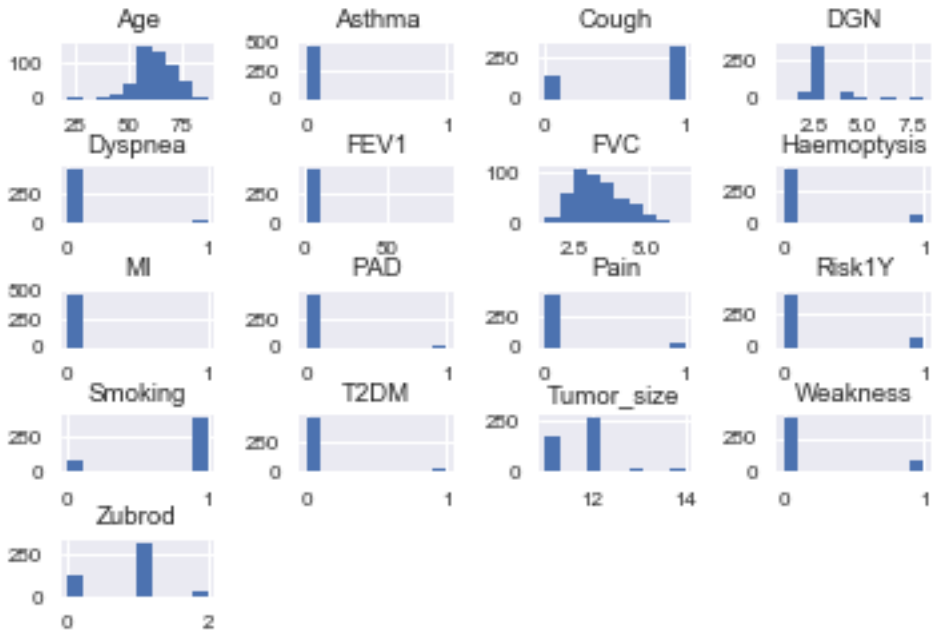

Figure 1. Histograms of feature and target data from the thoracic surgery dataset.

Among the features of this dataset, the one most closely related to the pathology of lung cancer is tumor size, so we may wish to immediately probe whether there are any patterns among tumor size classes versus mortality data. Figure 2 shows this relationship, with patients in the “False” category for “Risk1Y” being those who survived one year after thoracic surgery, with those who did not survive appearing in the “True” category. Tumor size is shown in classes “OC11” through “OC14”, with size classes increasing incrementally.

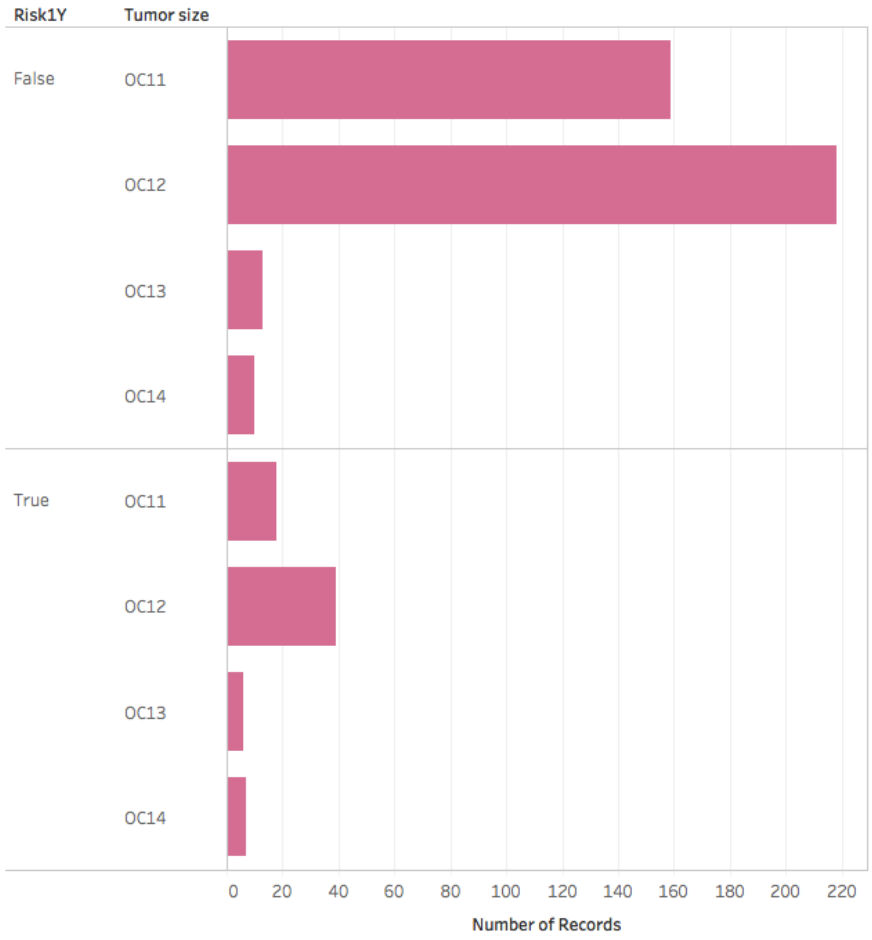

Figure 2. Frequency of tumor size classes divided into survival groups.

In general, most tumors are represented by the smaller two size classes, though the smallest class is not the most abundant among these patients who underwent surgical resection. More patients from each size class survived than did not, but relatively more of the patients from the larger tumor size classes did not survive. Figure 3 provides the same data in a different arrangement to underscore that point. Smaller size classes show a much greater proportion of survivors than do larger tumor size classes.

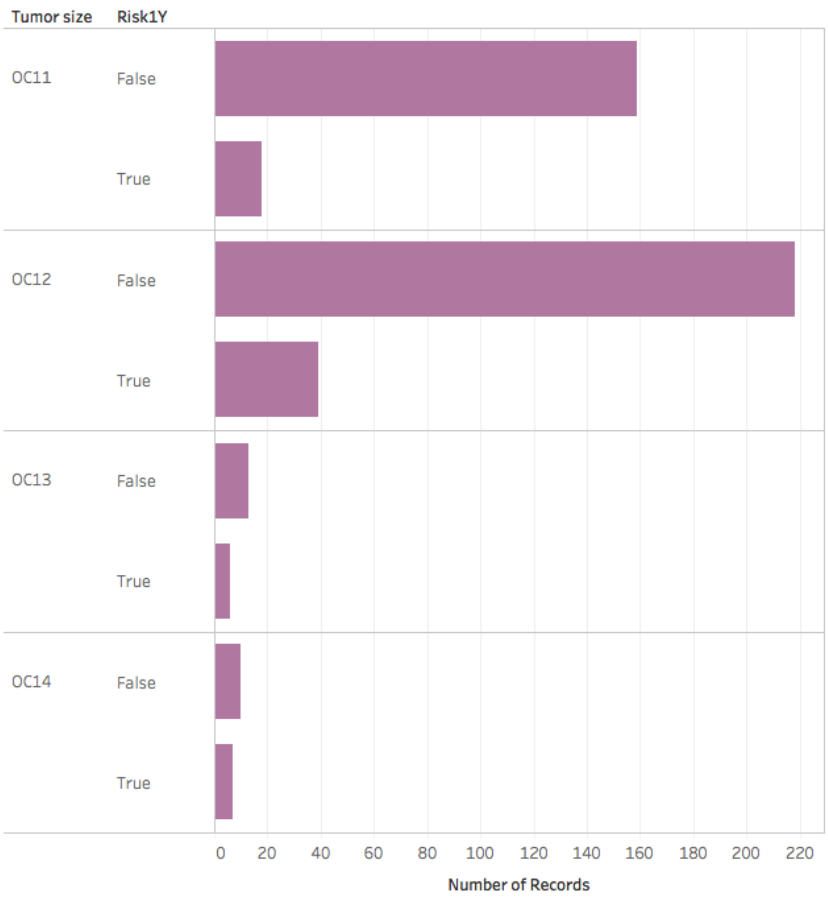

Figure 3. Survival patterns among tumor size classes.

When age is considered, slight differences appear in survival. Figure 4 shows the median ages of patients whose tumors appear in each size class and with respect to the patient's survival information. From the smallest tumor size class ("OC11"), the ages of survivors and non-survivors do not appear very distinct, and a similar pattern exists for patients with tumors in the second smallest class ("OC12"), too, with slightly older patients from this group passing away. However, from the second largest tumor class ("OC13"), the age among those who did not survive appears considerably younger than for other groups. For the largest tumor category ("OC14"), older patients, similar to what is shown for "OC12", are among the non-survivors, while somewhat younger patients from this group do survive. It is unclear why the second largest tumor size class is associated with greater mortality among younger patients.

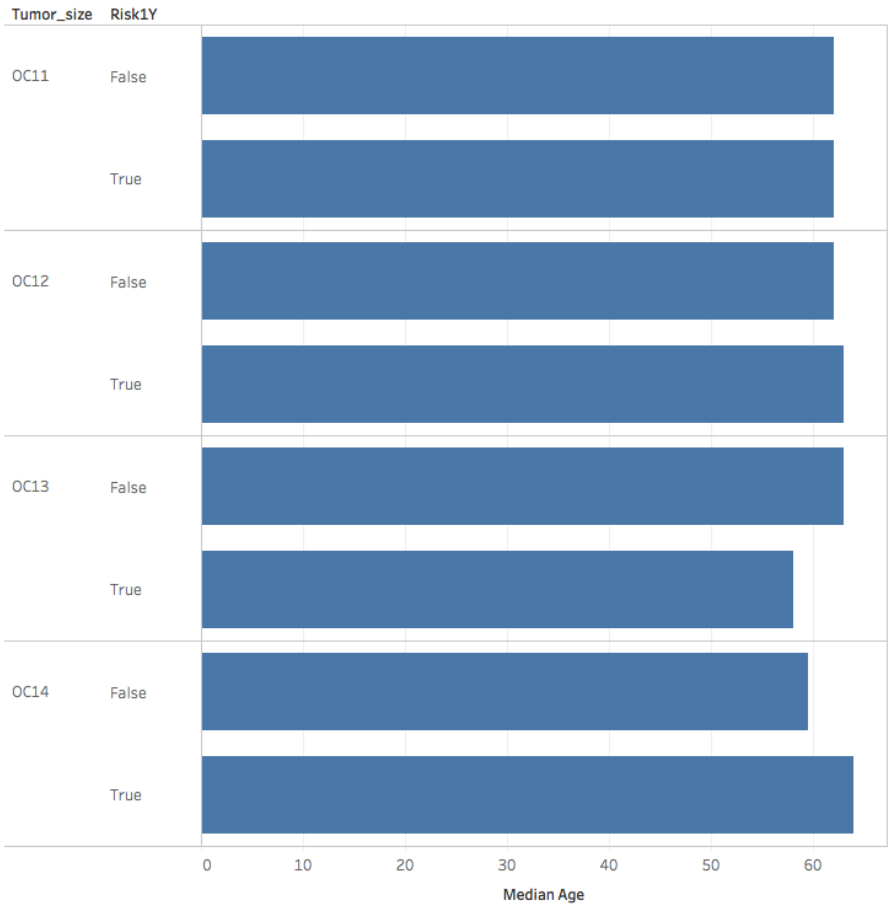

Figure 4. Survival and tumor size classes by patient age.

Three columns of data contain continuous variables, being Age, FVC, and FEV1, though FEV1 does not appear as such in Figure 1, in which it appears to have an extraordinarily long x-axis that may be indicative of outliers.

Figures 5 and 6 show simple regression plots of age and FVC or age and FEV1, respectively, with discrimination between survival (0, blue) and mortality at one year (1, green) shown for each.

From Figure 5 we can see that there is a tremendous amount of scatter within the data between age and FVC, and the trendline through these data for patients who passed away during the study shows a very large confidence interval. For both categories of survival data, with advancing age the FVC values trend slightly downward, though they perhaps started at a lower level for those who were younger who passed away during the one-year period following surgery.

From Figure 6, which plots age versus FEV1 values, it is apparent that there are approximately 14 obvious outliers in the FEV1 data. It is highly unlikely that for a feature for which most patients vary from each other by a few integers and with most data points below 10, that the data points that appear above, for instance, 40 along the y-axis are accurate if scaled with the same units as the rest of the data are.

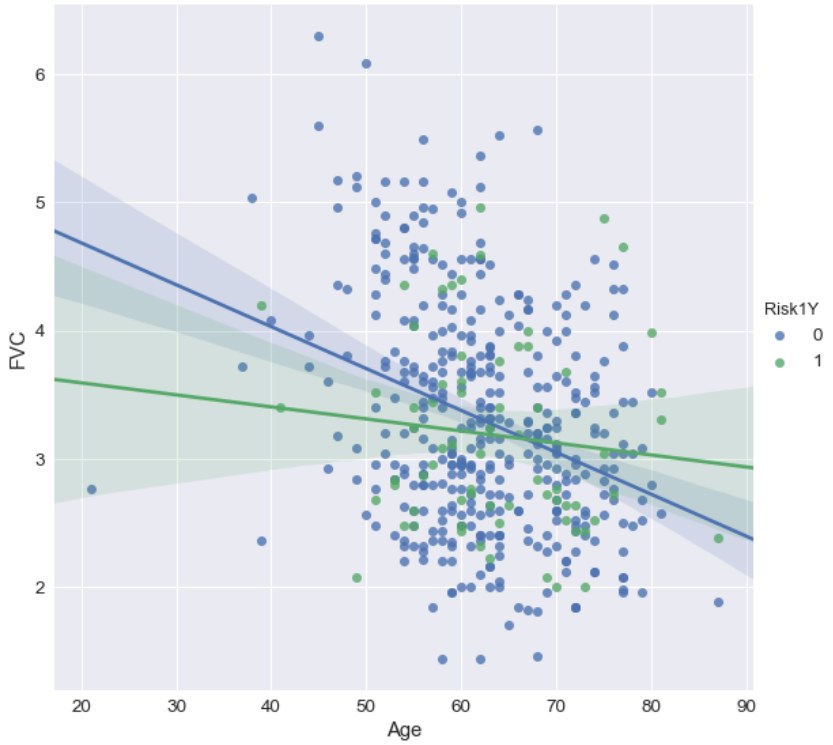

Figure 5. Age versus FVC, by one-year mortality.

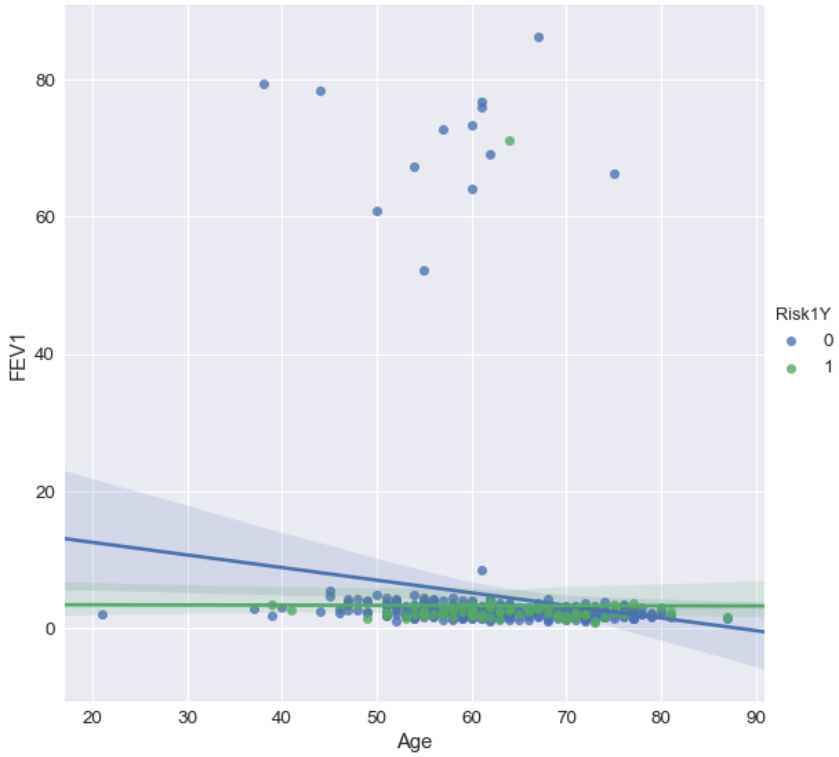

Figure 6. Age versus FEV1, by one-year mortality.

Analyzing the relationship between age and FEV1 further through ordinary least-squares regression using Python's Statsmodels package reveals that these 14 high-FEV1 data points fall quite far from their predicted range (Figure 7) and that their residual values depart greatly from the line of best fit (Figure 8).

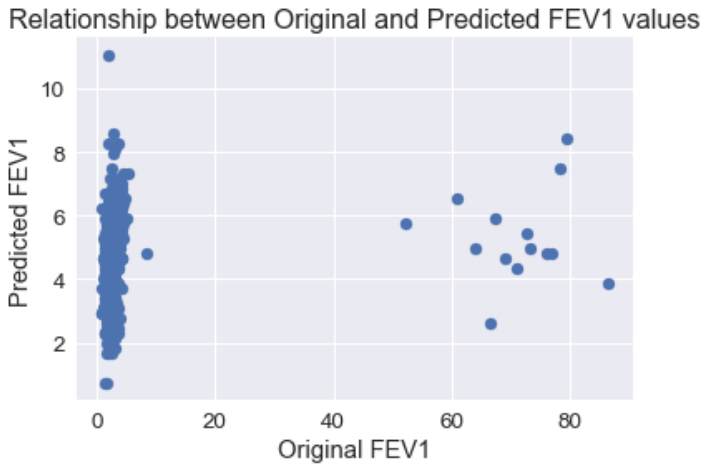

Figure 7. Actual (original) FEV1 values versus predicted FEV1 values.

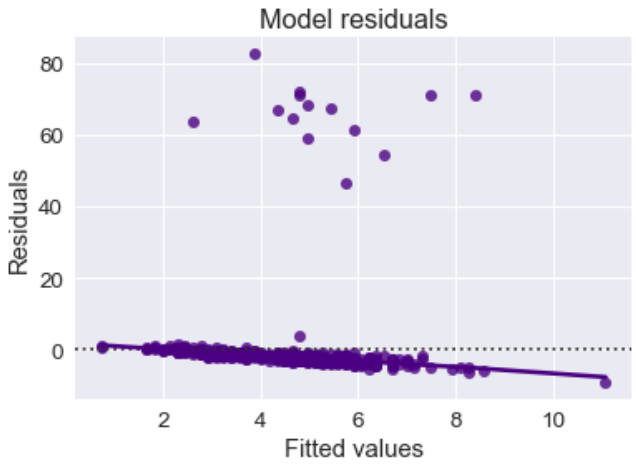

Figure 8. FEV1 residuals versus the model line of best fit.

In a separate ordinary least-squares regression analysis of mortality data fitted against FEV1 data, it appears at least one of the data points (which is for an entry marked with an FEV1 of 71.1 in a patient who passed away within a year of surgery) imparts an outsized influence on this regression (Figure 9).

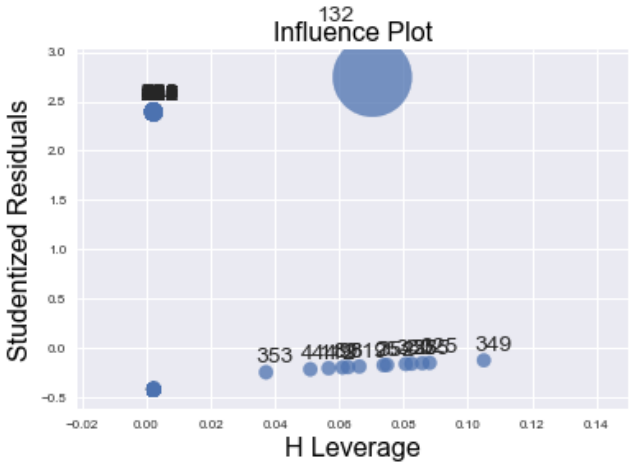

Figure 9. Influence plot (based on Cook's distances for each entry) of residuals for mortality data regressed with FEV1 data.

Assuming that the 14 FEV1 likely outliers exist in the dataset in error, it may be prudent to remove them from the dataset prior to predictive modeling. Incorrect information is not helpful in designing a predictive model, and the dataset of n=470 is large enough that removal of 14 entries should not cripple model development. The next step in treatment of this dataset was to remove these entries with FEV1 values above 40 units, resulting in n=456 for further analysis. All interpretations presented from this point forward do not contain data using the 14 identified outliers.

Figure 10 shows a new residual plot for the FEV1 data regressed against age after removal of outliers.

Figure 10. FEV1 residuals versus the model line of best fit after removal of outliers.

Figure 11 shows another presentation of age versus FEV1 data, but this time with the described outliers removed. The pattern for these data is nearly identical to what we saw in Figure 7 for age versus FVC.

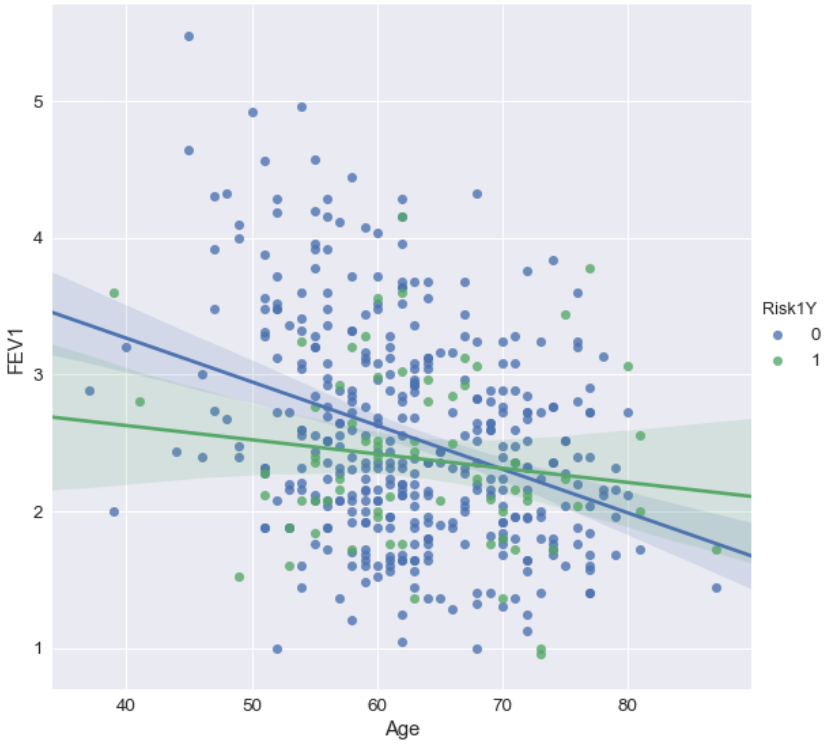

Figure 11. Age versus FEV1 with outliers removed.

While tumor size appears in the dataset as a discrete variable, we can examine it versus age, as in Figure 12 to get a rough idea of any patterns.

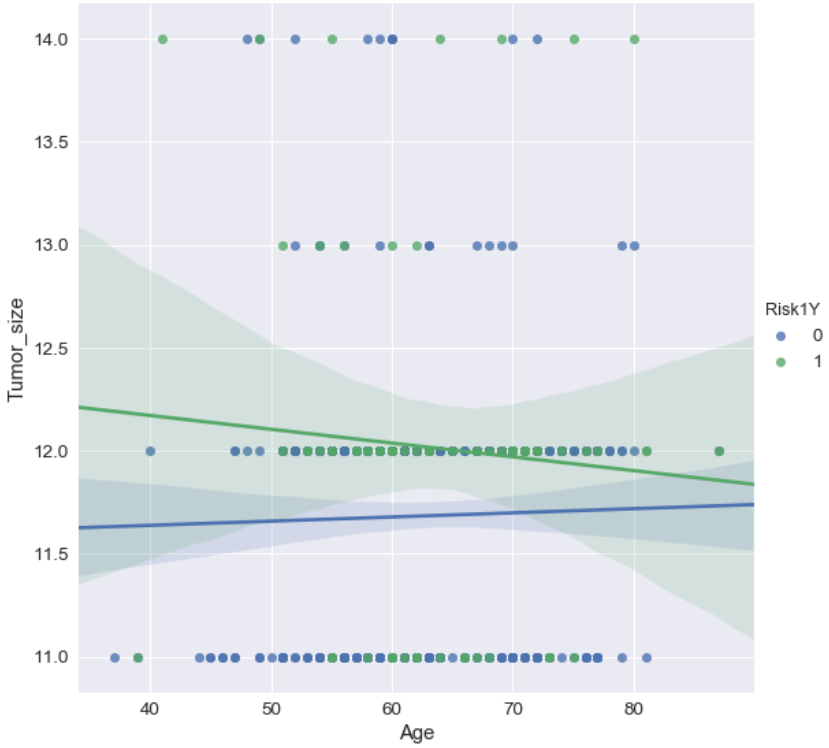

Figure 12. Age versus tumor size, by one-year mortality.

Most tumors in this dataset appeared to be in the smaller size groups (categories 11 and 12), and survivors overall showed smaller tumors, though there is considerable overlap in confidence intervals. Across ages, tumor size did not appear to vary much for survivors. For those patients who passed away, younger patients seem to have had larger tumors, and mortality with smaller tumors increased as patients were older.

Overall the dataset seems to be low in correlations among variables with each other, as shown in the heatmap of Pearson's correlation coefficients in Figure 13. Most variables are categorical, so Pearson's correlation coefficient is not the best statistics for measuring their relationships, but it can give a quick view of relationships.

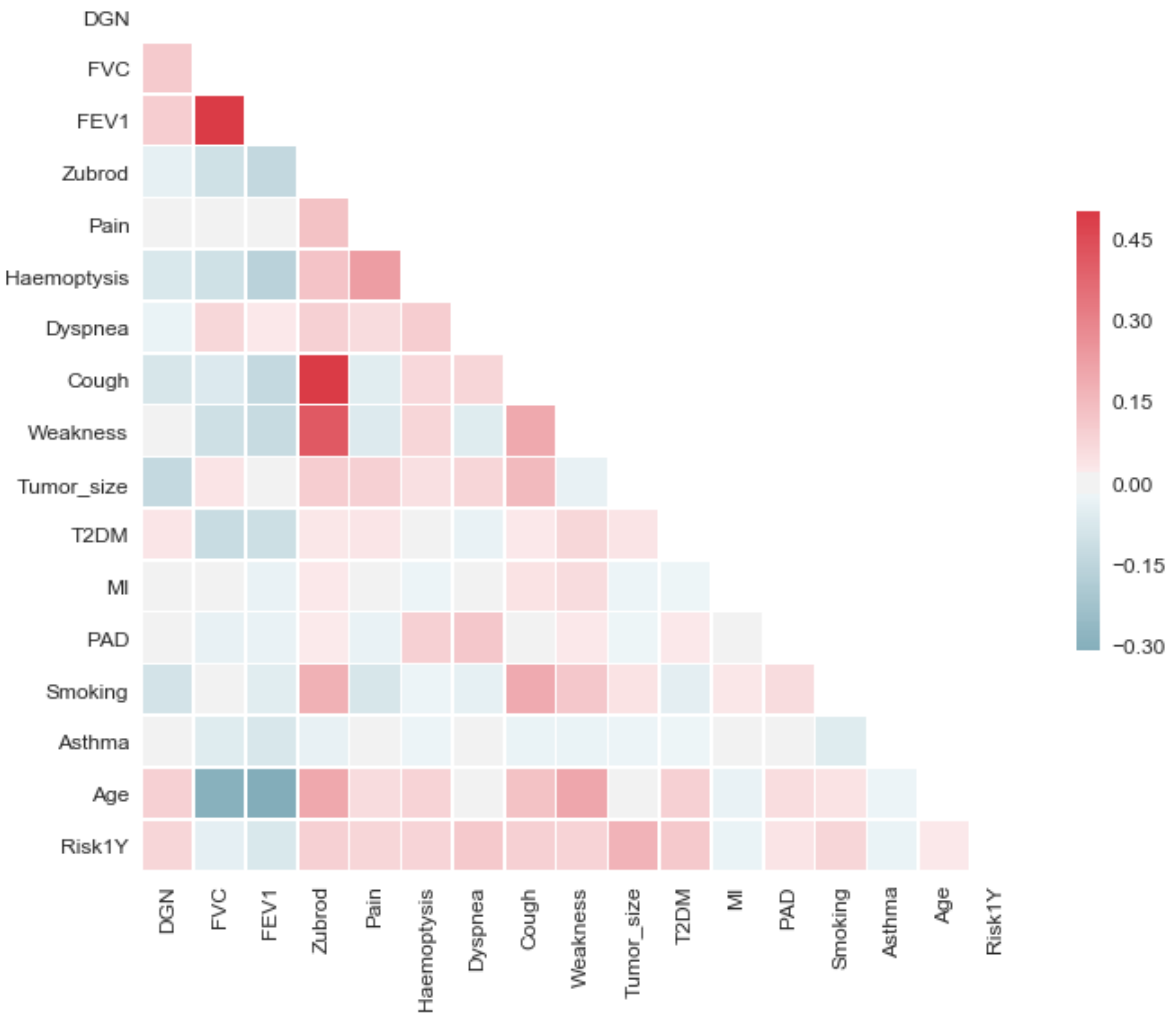

Figure 13. Heatmap of correlations among variables of this dataset.

Most relationships in Figure 13 show correlation statistics near 0, with the exceptions of positive relationships to be spotted for FVC plotted against FEV1 and cough and weakness each plotted against Zubrod score. None of these are surprising. FEV1 and FVC are both very similar metrics that measure breathing capacity, so they would be expected to covary. As Zubrod score is a metric of a person's level of function, weakness would be expected to appear with a worse Zubrod score. Severity of cough seems

to relate to Zubrod score more strongly, but this, too, might be expected for patients who are suffereing with a more severe lung condition.

Negative relationships appear for both age with FVC and age with FEV1. It is again unsurprising that FVC and FEV1 would trend the same way in this case, but it may not necessarily be obvious that age would, on its own, have a negative effect on a person's respiratory capacity.

More quantitative analyses of relationships among variables show that the Pearson correlation coefficients are all highly statistically significant among continuous variables in this study, with FVC and FEV1 showing a Pearson coefficient of 0.888 (p=0.0). Age shows Pearson coefficients of -0.299 (p=0.0) and -0.031 (p=0.0) for FVC and FEV1, respectively.

Chi-squared statistics of relationships between categorical variables in this study show most variables to not covary, with a few exceptions. Relationships that are present a level of $p<0.001$ exist for Zubrod score with cough ($\chi^2 = 259.27$), Zubrod with weakness ($\chi^2 =103.59$), Zubrod with smoking ($\chi^2 =16.53$), pain with haemoptysis ($\chi^2 =20.38$), pain with tumor size ($\chi^2 =16.48$), cough with weakness ($\chi^2 =16.93$), cough with smoking ($\chi^2 =16.01$), and myocardial infarction with asthma ($\chi^2 =27.63$).

Chi-squared statistics that are statistically significant at the level of $0.001 \leq p < 0.05$ include diagnosis code with one-year mortality ($\chi^2 =21.56$), Zubrod score with pain ($\chi^2 =8.21$), Zubrod score with haemoptysis ($\chi^2 =7.76$), Zubrod with tumor size ($\chi^2 =14.86$), dyspnea with tumor size ($\chi^2 =10.24$), dyspnea with one-year mortality ($\chi^2 =4.5$), cough with tumor size ($\chi^2 =10.8$), tumor size with one-year mortality ($\chi^2 =14.28$), and peripheral arterial disease with asthma ($\chi^2 =6.27$). The remaining 61 relationships tested with chi-squared analysis revealed no close associations.

Overall, this appears to be a dataset for which predictive modeling should be able to proceed without serious concerns of covariance among variables. It is perhaps unnecessary for FVC and FEV1 to both be included, but this can be considered if there are difficulties in generating a model with good accuracy metrics.

Analytical approach for finding the best predictive model

The target variable to be predicted in our machine learning model is mortality assessed at one year post-surgery, which is organized within the data as a binary variable of survival versus mortality. We can choose a binary classification predictive approach based on labeled data, and a supervised classification model (such as random forest or XGBoost) would be a prudent approach to predict probability of categorization into either class. The dataset shows imbalance in mortality data and some other features, so we can apply SMOTE or another resampling technique to the dataset to generate synthetic samples through interpolation prior to modeling. We can use SHAP values to determine feature importance collectively or individually. As appropriate, based on patterns that emerge from these values, we can also re-examine the model to focus on complex features individually against the rest of the features. Most features are present as Boolean values or categorical variables representing diagnosis codes or classes of tumor size. However, some features, such as age, forced vital capacity, and forced expiration volume, are continuous. All potentially can contribute to results in a complex fashion, however.

Even among supervised classification model types there are several options, including classifier and hyperparameter options. The dataset does not included missing data, so there is no need

to consider imputing data or restricting analysis to a model that can handling missing data. Prior to SMOTE, data were split into training and testing subsets (75% of data for training, 25% for testing).

Figure 14 shows a series of receiver-operating curves for extra trees, random forest, and XGBoost models applied to this dataset's test data, based on training with SMOTE-treated data.

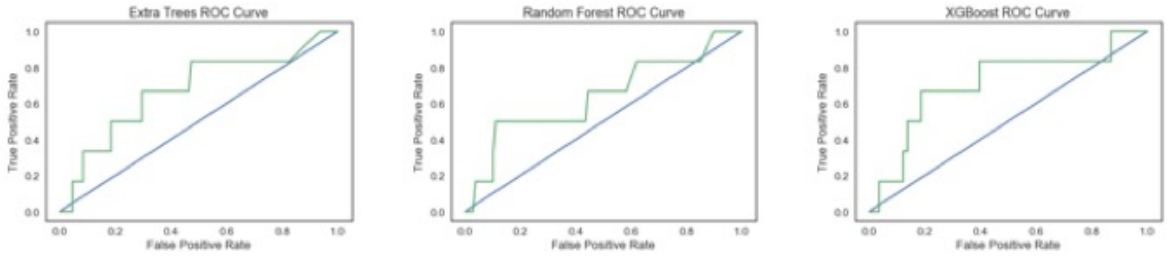

Figure 14. ROCs for the top three model classifiers for this study using a SMOTE-treated training set: extra trees, random forest, and XGBoost.

For each of the models examined in Figure 14, there appears to be greater accuracy than what is expected by randomness, and in fact it could be very difficult to predict mortality from a condition within a year, but Table 2 reveals a possible issue with model development so far.

Table 2. Classifiers and accuracy, precision, and recall scores from SMOTE-treated training data.

	Accuracy	Class	Precision	Recall	F1-score	Support
Extra trees:						
Training	1.0					
Testing	0.877					
AU-ROC	0.6736					
		0	0.95	0.92	0.93	108
		1	0.10	0.17	0.12	6
		Avg/total	0.91	0.88	0.89	114
Random forest:						
Training	1.0					
Testing	0.904					
AU-ROC	0.6404					
		0	0.95	0.94	0.95	108
		1	0.14	0.15	0.15	6
		Avg/total	0.91	0.91	0.91	114
XGBoost:						
Training	1.0					
Testing	0.868					
AU-ROC	0.7083					
		0	0.95	0.91	0.93	108
		1	0.09	0.17	0.12	6
		Avg/total	0.91	0.87	0.89	114

While accuracy (sum of true positives and true negatives over the total) and precision (sum of true positives over predicted positives) show each of these tests to perform well for identifying those patients who survive over the one-year period, each analysis shows low scores for precision and recall

for those who do not survive one year. This likely relates to the imbalance in data, which SMOTE treatment was expected to assist with, but it does not seem that this approach worked well here.

SMOTE is not the only method by which sampling of data can balance classes within a dataset for predictive analysis. While SMOTE is based on sampling based on nearest-neighbor distances between points in order to bolster the number of data points, another approach that can be taken is to resample data points from the minority class. In the case of the training partition of this dataset, there are 277 members in the survivor class, compared with 63 members in the non-surviving class. A resampling approach to balance these class is to "upsample" the minority class (non-survivors in this case) so that each class includes 277 samples. The next analyses show modeling results for data that have been upsampled in such a manner after data were split into training and test sets (75% of data for training, 25% for testing). Table 3 shows accuracy, precision, and recall using upsampled training data (just as performed to generate Table 2).

ROCs give the impression that, on the test data, XGBoost may perform the best of these three models chosen here. However, precision-recall numbers reveal that the original problem of difficulty with correctly classifying the minority class (non-survivors) persists with this small testing dataset. With 6 samples here, there is little room for error, but in all cases the models struggle to predict mortality.

Table 3. Classifiers and accuracy, precision, and recall scores from upsampled training data.

	Accuracy	Class	Precision	Recall	F1-score	Support
Extra trees:						
Training	1.0					
Testing	0.904					
AU-ROC	0.6127					
		0	0.94	0.95	0.95	108
		1	0.00	0.00	0.00	6
		Avg/total	0.90	0.90	0.90	114
Random forest:						
Training	1.0					
Testing	0.921					
AU-ROC	0.7014					
		0	0.95	0.96	0.98	108
		1	0.20	0.17	0.18	6
		Avg/total	0.91	0.92	0.92	114
XGBoost:						
Training	1.0					
Testing	0.86					
AU-ROC	0.733					
		0	0.95	0.90	0.92	108
		1	0.08	0.17	0.11	6
		Avg/total	0.91	0.86	0.88	114

Low precision and recall with mortality with up- or resampled data

For this dataset it is not difficult to generate a model that can report high accuracy and even a high AUC for ROC. It is even possible with few optimizations to generate a model with a level of accuracy that is

higher than the prevalence of survival is in the population; in this dataset about 85% of the patients survived the first year after surgery, so any model that reports greater accuracy than 85% should in theory perform better than random guessing would. However, precision and recall values associated with the target class (1) that refers to patients who did not survive the first year post-surgery from these models are low.

While for some imbalanced datasets SMOTE, up- or downsampling pretreatment can aid in artificially balancing target classes to improve model accuracy, in this case these methods may not enhance predictive utility outcome. Another option, setting the weights of target classes prior to model training, may work better in this case.

Instead of SMOTE or upsampling training data prior to introducing data to classifiers, the train-test split was next constructed with stratification so that both sets contained the same ratio of target classes. Also, class weights were balanced for classification with tree-based methods in Scikit-Learn. Additionally, probabilities of predicting each class were accounted for in isotonic and sigmoid calibrations applied to each model, in comparison with uncalibrated models. This in many cases allowed for more balanced precision-recall scores between target classes. While F1-scores tended to go down with calibrations (Table 4), this represented slightly lower precision and recall values for the majority class while the minority class gained significantly in these values.

Scores for areas under the curve of ROC curves (AU-ROCs) (Table 4) tended to be higher with isotonic calibration for XGBoost and gradient boosting classifier models, but not sigmoid calibration. XGBoost, however, showed overall lower AU-ROC in Table 4 than in Table 3 with upsampled data.

For tree-based classifiers, uncalibrated models performed generally about as well as isotonically calibrated models did, and sigmoid calibration was not helpful. The decision tree classifier was alone among classifiers in producing higher precision and recall values for the minority class, in addition to a relatively high AU-ROC among all models tested without calibration.

Even with higher AU-ROC values and lower Brier scores for some classifiers in Table 4, F1-scores at times fell due to trade-offs with more balanced precision and recall scores between the majority and minority classes. For the random forest model, isotonic calibration tended to result in improved precision, recall, and Brier scores over the uncalibrated model, but the AU-ROC score tended to be lower, so the uncalibrated values are presented in Table 4 for this model, assuming that AU-ROC score is the metric to prioritize. For the decision tree classifier, both isotonic and sigmoid calibrations showed slightly lower Brier scores than did the absence of calibration, but at a serious cost to the AU-ROC score and with worse precision, recall, and F1-scores, so this classifier is presented in Table 4 also without calibration.

With fairly similar AU-ROC scores among each of the classifiers shown in Table 4, it can be difficult to decide that they show dramatically different performance, so for models showing generally similar AU-ROC scores, precision and recall scores can provide additional information into how well each model allocates predictions.

The isotonic calibration applied to some models is geared toward factoring in the probability of any sample being associated with either class. Since 85% of samples belong to the survivor class, it is expected that 85% of predictions would be correct if just assigning patients to that class. With calibration, the classes are weighted by the probability associated with the class. In general, among models (and including those with data not shown here), isotonic calibration showed higher precision

and recall scores with the non-survivor class, though the decision tree classifier without calibration presented a relatively high AU-ROC score (0.722) in the presence of higher precision and recall values for the non-survivor class, while maintaining high precision and recall for the survivor class, and with an overall average F1-score of 0.83. It was not necessarily anticipated that the decision tree classifier would rival or outdo random forest, which is generally more successful than a single decision tree is, especially if there is any instability. However, for this dataset, the decision tree classifier by itself tended to perform well.

Table 4. Test data metrics with sample weighting, stratification, and calibration (calib.).

	Score	Class	Precision	Recall	F1-score	Support
Extra trees (isotonic calib.):						
Brier score	0.229					
AU-ROC	0.7216					
		0	0.94	0.60	0.73	97
		1	0.25	0.76	0.38	17
		Avg/total	0.83	0.62	0.68	114
Random forest (no calib.):						
Brier score	0.324					
AU-ROC	0.7247					
		0	0.85	1.00	0.92	97
		1	0.00	0.00	0.00	17
		Avg/total	0.72	0.85	0.78	114
XGBoost (isotonic calib.):						
Brier score	0.233					
AU-ROC	0.7089					
		0	0.91	0.73	0.81	97
		1	0.28	0.59	0.38	17
		Avg/total	0.82	0.71	0.75	114
Decision tree (no calib.):						
Brier score	0.277					
AU-ROC	0.722					
		0	0.92	0.86	0.89	97
		1	0.42	0.59	0.49	17
		Avg/total	0.85	0.82	0.83	114

The uncalibrated (Uncal. DT) decision tree model applied to this dataset shows a pronounced elbow in its ROC curve (Figure 15), which is plotted with curves for both isotonic calibration (IC DT) and sigmoidal calibration (SC DT).

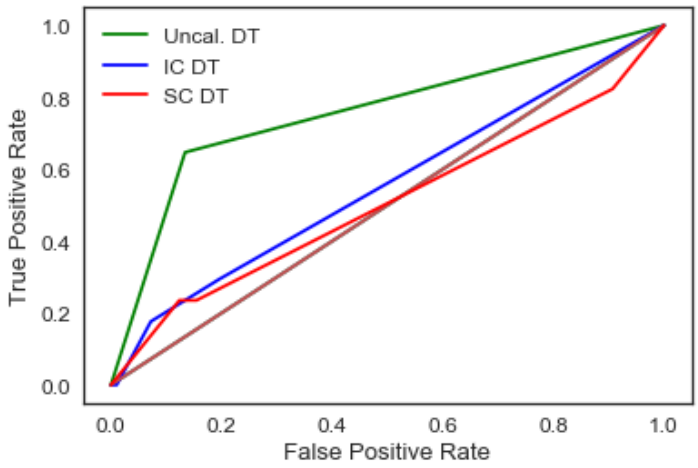

Figure 15. Decision tree (DT) classifier ROC curve presenting uncalibrated (Uncal.), isotonically calibrated (IC), and sigmoidally calibrated (SC) models.

The decision tree generated with results shown in Table 4 is fairly elaborate, with 23 levels to its hierarchy and a portion of those shown in Figure 16. The very first decision point is whether a patient has a diagnosis code of below or equal to 4.5 (referring to "DGN" codes 1,2,3, or 4). If true, then the next decision along one path is the individual's smoking status. If not a smoker, then the decision is whether the patient's tumor is in one of the two smaller size categories, with either result here leading to a Gini impurity of 0.0.

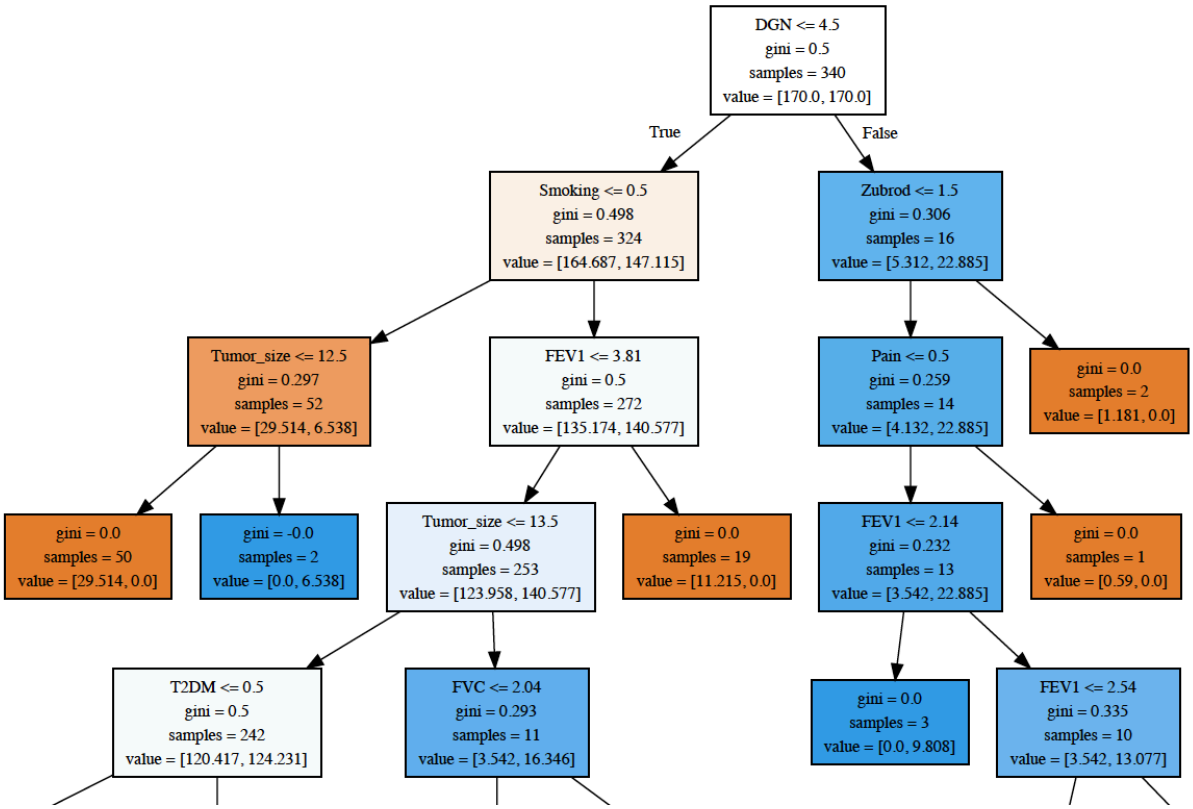

Figure 16. Top levels of the decision tree hierarchy described in this study.

Feature importance

In order to interpret a model's results and also to ensure that a model is as computationally efficient as possible (the latter being more important for datasets with enormous feature sets), it is helpful to know the relative importance of each feature. If the decision tree in Figure 16 provides any hint as to feature importance, it is that there is significance to the patient's diagnosis associated with the tumor for which he/she undergoes resection. Also, FVC and FEV1 frequently appear in the tree, which is indicative of the importance of these two variables. However, decision trees can branch in varying ways, which is why random forests often outperform them, as they combine the results of multiple trees.

Principal component analysis (PCA) can be used to determine how variance relates to the complexity of the feature set. The number of components analyzed by PCA translates to the number of features in the dataset, and overall contribution to variance by each translates to how complex the dataset is or how little covariance there is.

Figure 17 shows a PCA plot of variances for each component measured using the dataset from this study, with data normalized using Scikit-Learn's Standard Scaler, since the scale for age swamps the range of any other variable in the dataset. There is one component that shows the strongest signal, but 12 other components that also seem to stand out in nearly equal measure to most components. Two components appear relatively unimportant, and this is not a surprise as there are two features (myocardial infarction (MI) and asthma) that are not prominent in the dataset. However, this does not mean that they are not important. What PCA does show, though, is that most features appear to independently contribute influential information.

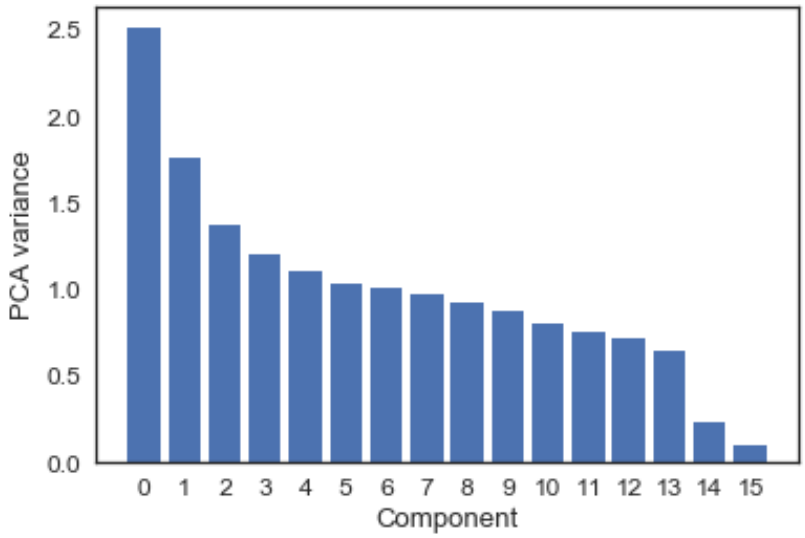

Figure 17. Principal component analysis of the dataset in this study using normalized data.

A comparison of feature importances calculated for each of the four classifiers discussed here is shown in Figure 18, with code adapted from here. Here the classifiers are not calibrated, as this is a straight comparison between each using the full dataset, though the extra trees and random forest classifiers use 300 estimators each, and both of these in addition to the decision tree classifier are parameterized to balance class weight.

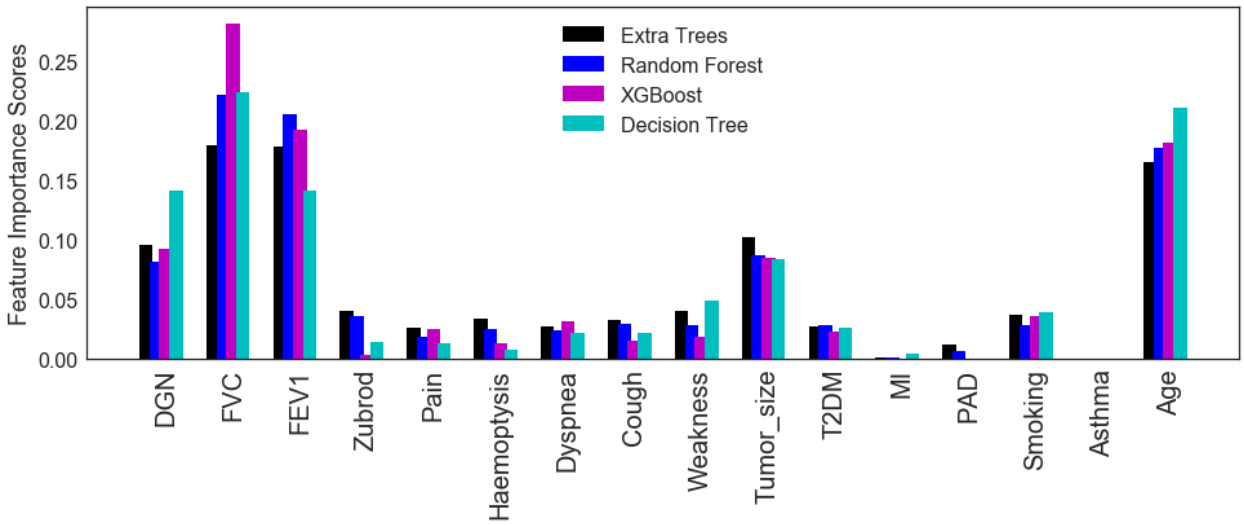

Figure 18. Feature importance scores for the full dataset shown for four classifiers.

From Figure 18 we can see that for feature importance each included classifier shows a lot of agreement on the whole, with slight variations. The qualities of FVC, FEV1, and age stand out as important. FVC and FEV1 are closely related, so it is not a surprise that they would show similar levels of importance. Diagnosis and tumor size appear next in importance. Myocardial infarction (MI) and asthma barely register on the scale, though this is not surprising due to their lack of abundance in the dataset. The other qualities of smoking, weakness, type II diabetes (T2DM), Zubrod score, pain, dyspnea, cough, and peripheral arterial disease (PAD) rank lower in importance for each classifier. Among classifiers there does not seem to be a trend for allocation of feature importance by classifier type, with the tree-based methods (which are also weighted by target class) not diverging from the boosting method. XGBoost does stand out for its level of importance given to FVC, but this is a highly important feature for each classifier shown here. Zubrod is also considerably less important to the XGBoost model than to extra trees or random forest models, though this feature appears at a relatively low importance-level for each.

Model interpretation

Since it may be desired to ascertain risk levels for particular patients associated with thoracic surgery, it may be helpful to have an interpretable model that can take into account each patient's risk factors. It is inherently difficult to predict whether a person will survive for a year or not using a predictive model, and the decision to undergo pulmonary surgery is not trivial. Pulmonary dysfunction itself could potentially make recovery from such a surgery more difficult if the patient has less capacity to breathe adequately in order to limit general post-surgical complications.

No classifier in this study produced a model with close to 100% precision or recall on test data with consideration of the minority target class. However, each of the metrics presented in Table 4 of this report suggest that this dataset may provide useful insight into risk factors.

To determine individual risk levels and the component variables that influence these, it is possible to use the SHAP Python package with some classifiers. This package allows viewing of general trends

throughout a global population with a hover tool explaining important features for each individual, as well as a view of the individual's feature make-up. Figure 19 shows a view of the global SHAP output for decision tree and random forest classifiers applied to test data. (The extra trees classifier has no SHAP deployment at this time, and XGBoost showed less accuracy than the other models, so it is excluded here.)

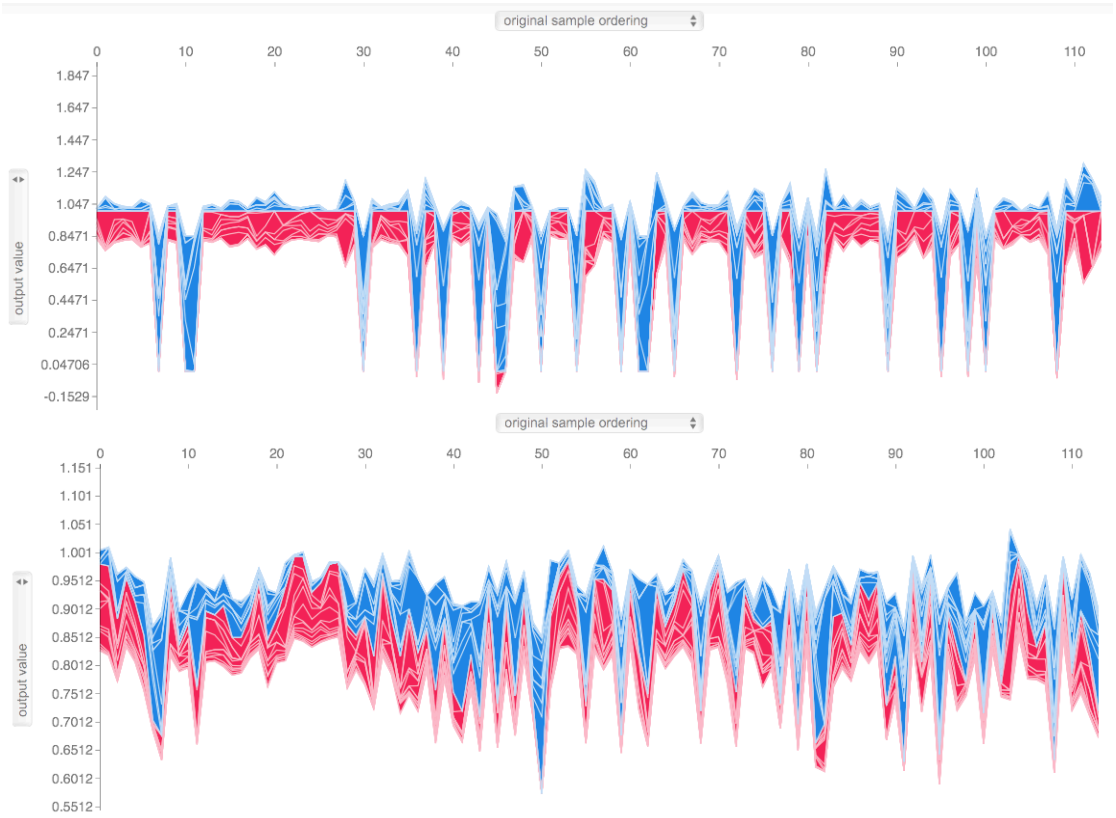

Figure 19. SHAP global output for test data with decision tree (top) and random forest (bottom) models.

With the random forest and decision tree models, AU-ROC scores are easy to compute and are 0.7383 for this random forest model and 0.7462 for this decision tree classifier. Figure 19 additionally shows that the range of predictive values produced for the decision tree classifier (y-axis) is broader than for the random forest model.

For each classifier it is possible to view the influence of component features for each individual whose data appear in the dataset. Figure 20 shows an individual's SHAP output for a survivor, as assessed using the decision tree model, and Figure 21 shows the same for a patient who did not survive.

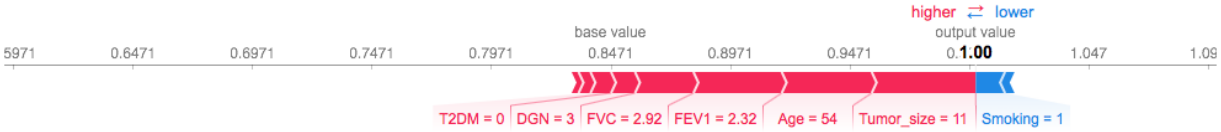

Figure 20. Decision tree-based SHAP output for a correctly predicted surviving patient.

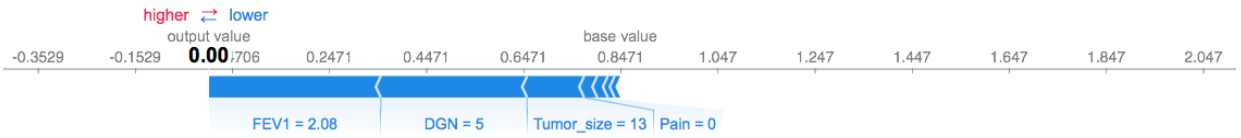

Figure 21. Decision tree-based SHAP output for a patient correctly predicted to not survive.

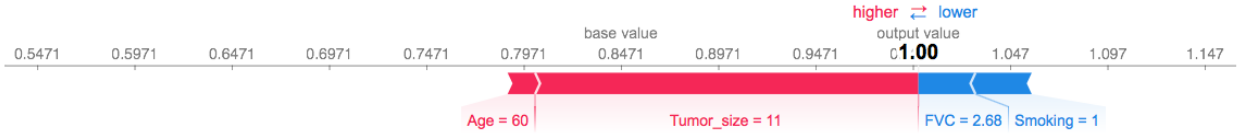

Figure 22. Decision tree-based SHAP output, failing to identify a patient who did not survive.

Unlike with the decision tree classifier, the random forest predictions (not shown here) did not fall on 1.00 or 0.00 as much, so that is not a strict feature of the KernelExplainer method, but perhaps is a feature of the decision tree classifier. In Figures 20 and 21, even though the decision tree classifier seems to perform well at prediction, not all feature assignments seem logical. For the surviving patient in Figure 20, the values appear sensible, but in Figure 21 it is not clear why absence of pain is associated with non-survival. Even though the decision tree classifier overall performs fairly well here, various oddities like these appear among the individual outputs, and this may be a product of the small minority class in this dataset being more difficult to learn from. However, the medium-large tumor and low FEV1 scores, in addition to the diagnosis code (survivors on average had a lower number for diagnosis code), contribute sensibly to the prediction.

In Figure 22 is output for a patient that the model failed to identify as not surviving. From the individual SHAP output it makes sense that presence of smoking may be contrary to survival, and the FVC value is a little low. The small tumor size appears to be driving a lot of the prediction of survival for this patient, but the patient did not survive the year.

In Figure 23 is shown a SHAP-based feature importance plot that incorporates information from every single data point analyzed and what effect each has had on the SHAP prediction score. The AU-ROC value for this model is 0.7462. In the case of this decision tree-based plot, a higher SHAP value reflects predictive survival. High values for tumor size and type II diabetes appear to strongly limit predicted survival. Many features, such as pain and those below it along the vertical axis have very little, if any, effect on survival. Lack of smoking seems to raise risk of survival, though for some patients this is shown here to have the opposite effect, which reflects either the model's imperfection or the presence of an unknown factor to consider. Presence of smoking appears to lower predicted survival, although this does not show a pronounced effect. Conversely, lack of type II diabetes is slightly associated with an increased prediction of survival. Many of the most important features, including FEV1, FVC, and age, show some complexity, with some high and low values for each being associated with different survival predictions. This is not necessarily problematic, as it is possible, for instance, that a younger patient with the same FVC value as an older patient may be more at risk of mortality. A difference between the decision tree classifier and a random forest model is that variance tends to be higher with the decision

tree; variance tends to be trimmed down with a random forest. Still, the accuracy and AU-ROC scores trend higher for the decision tree classifier than for random forest with this dataset.

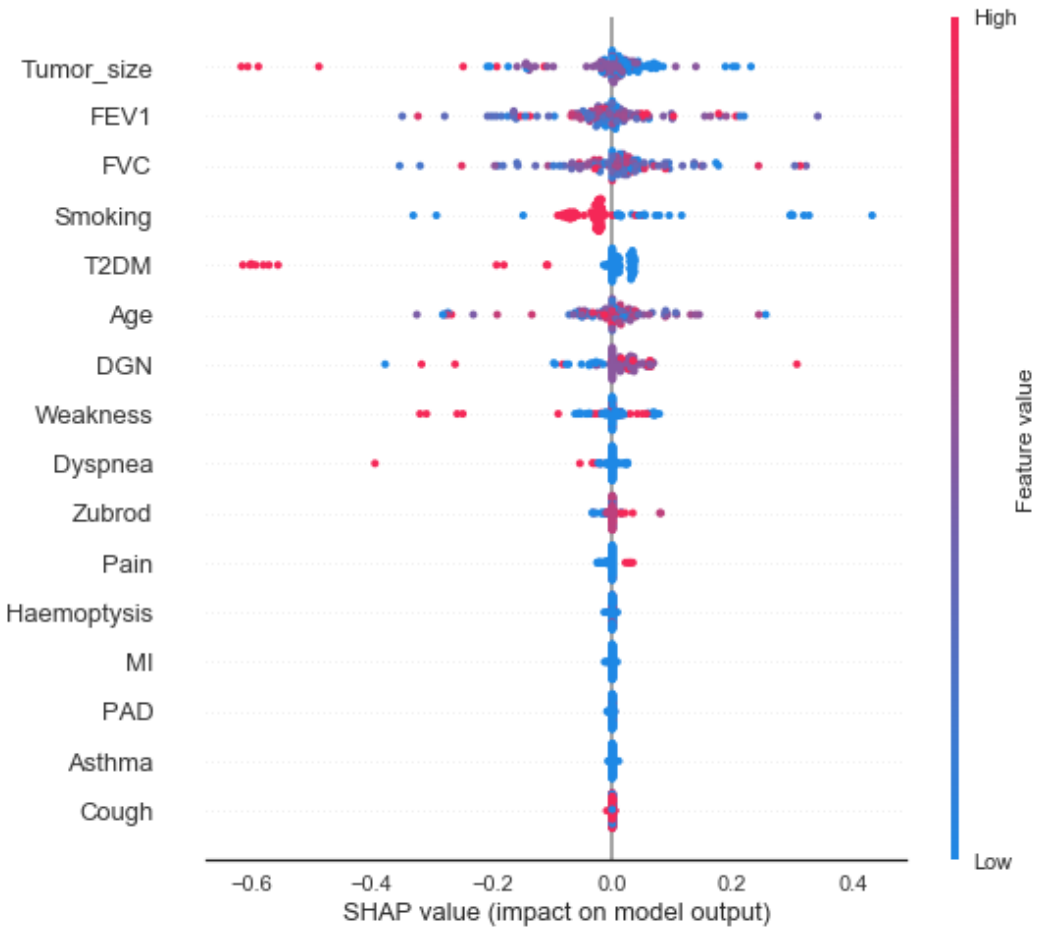

Figure 23. Decision tree-based SHAP global feature summary plot.

Figure 24 shows a similar plot as in Figure 23 but developed from a random forest model instead of the decision tree classifier. The AU-ROC value is 0.7383, which is slightly lower than that from the decision tree model used to generate the plot in Figure 23, but the two scores are very close. Overall the scale of SHAP values is smaller than in Figure 23, possibly due to the fact that that a random forest is an averaging of many decision trees, so any extreme values would likely become lost as values conform to central tendencies. However, the AU-ROC for this random forest model is lower than that for the decision tree classifier, as before.

The random forest feature summary in Figure 24 shows some differences in the order of influential features versus the decision tree classifier, with, for example, presence of cough taking on a more prominent role than shown in Figure 23. In Figure 24, cough is placed in a medium position relative to other features, and it shows impacts on SHAP values, with lack of cough tending to reflect a prediction of survival and presence of cough showing a slightly increased prediction of non-survival. In Figure 23, this feature is ranked at the bottom. Smoking, according to Figure 24, more unambiguously affects survival; lack of smoking clearly here skews a prediction toward survival, while presence of smoking

clearly shows a slightly negative impact on the prediction. The apparent trend with smoking in Figure 24 may be due to decreased variance in the random forest model versus that of a single decision tree.

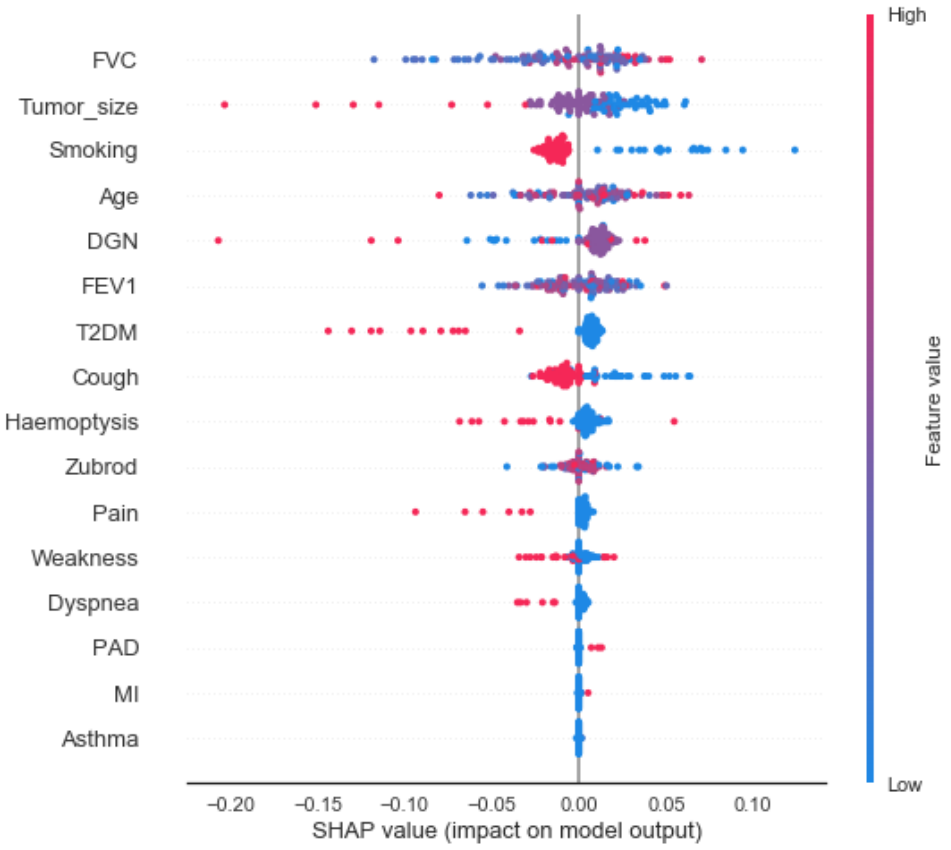

Figure 24. Random forest-based SHAP global feature summary plot.

Feature reduction

Even if a feature's importance can be ranked based on Gini impurity, it is not immediately clear how much of an effect a feature has on AU-ROC or accuracy. Individually, each of the features was removed from the total set and reexamined with extra trees, random forest, XGBoost, and decision tree classifiers. Based on AU-ROC scores, some were removed in combination and re-examined. Table 5 shows the resulting AU-ROC values for each test.

In many cases removal of a feature did not clearly improve a model's AU-ROC. Tumor size is a useful feature, for instance, across classifiers, as indicated by the substantially lower scores upon its removal. With almost every classifier but XGBoost, removal of age results in increased AU-ROC values, up to 0.7884 with the isotonically calibrated extra trees classifier and 0.7859 with the uncalibrated decision tree classifier. With removal of haemoptysis, the isotonically calibrated extra trees classifier showed an AU-ROC value of 0.7878, and the random forest AU-ROC was 0.782.

Removal of both age and haemoptysis improves the AU-ROC score for the uncalibrated decision tree classifier to 0.7911. Removal of MI, PAD, and asthma in some cases increases AU-ROC scores, though

the prevalence of each in the dataset is extremely low, with MI and asthma occurring only in two patients each, and only much larger survivor class for these. Realistically they seem to be not useful for this analysis.

Table 5. AU-ROC scores with test data for models with one or more features removed, with calibration type noted in parentheses as UC (uncalibrated), IC (isotonic), or SC (sigmoid).

Removed feature(s)	Extra trees AU-ROC	Random forest AU-ROC	XGBoost AU-ROC	Decision tree AU-ROC
Age	0.7884 (IC)	0.7508 (IC)	0.6683 (SC)	0.7859 (UC)
Tumor size	0.6316 (IC)	0.6892 (IC)	0.6331 (UC)	0.6043 (UC)
Diagnosis code	0.7195 (UC)	0.7362 (UC)	0.6771 (IC)	0.6528 (UC)
FVC	0.7295 (IC)	0.6883 (UC)	0.7283 (SC)	0.5992 (UC)
FEV1	0.6898 (IC)	0.6947 (IC)	0.6413 (IC)	0.6874 (UC)
Smoking	0.738 (IC)	0.7623 (SC)	0.7429 (IC)	0.6058 (UC)
Cough	0.6922 (IC)	0.7389 (UC)	0.6689 (UC)	0.722 (UC)
Haemoptysis	0.7878 (IC)	0.782 (SC)	0.6743 (IC)	0.7514 (UC)
Type II diabetes	0.7053 (UC)	0.7347 (UC)	0.6546 (UC)	0.6416 (IC)
Zubrod score	0.7277 (IC)	0.7492 (IC)	0.6998 (IC)	0.722 (UC)
Pain	0.721 (IC)	0.7638 (UC)	0.7029 (SC)	0.6856 (IC)
Dyspnea	0.7044 (IC)	0.732 (UC)	0.6998 (IC)	0.7116 (UC)
Weakness	0.7226 (IC)	0.7216 (IC)	0.6886 (IC)	0.6886 (IC)
MI	0.7814 (IC)	0.7456 (UC)	0.7089 (IC)	0.7359 (UC)
PAD	0.7662 (UC)	0.7195 (SC)	0.7089 (IC)	0.7168 (UC)
Asthma	0.7638 (IC)	0.7338 (UC)	0.7089 (IC)	0.7565 (UC)
Age and haemoptysis	0.7483 (UC)	0.7762 (IC)	0.6692 (SC)	0.7911 (UC)
Age and MI	0.7771 (IC)	0.7671 (UC)	0.6683 (SC)	0.7911 (UC)
Age and PAD	0.7738 (IC)	0.7838 (IC)	0.6683 (SC)	0.7911 (UC)
Age and asthma	0.775 (IC)	0.7632 (UC)	0.6683 (SC)	0.7911 (UC)
Age, haemoptysis, MI	0.7571 (IC)	0.762 (IC)	0.6692 (SC)	0.7962 (UC)
Age, haemoptysis, PAD	0.7435 (SC)	0.7693 (UC)	0.6692 (SC)	0.7911 (UC)
Age, haemoptysis, asthma	0.7632 (IC)	0.7744 (IC)	0.6692 (SC)	0.7911 (UC)
Haemoptysis, asthma, MI, PAD	0.7893 (IC)	0.7544 (UC)	0.6743 (IC)	0.7462 (UC)
Age, haemoptysis, asthma, MI, PAD	0.7859 (IC)	0.7641 (UC)	0.6692 (SC)	0.7962 (UC)

While age is considered important, it is not clear why removal of this feature tends to improve a model's AU-ROC. Perhaps more aggressive tumors afflict younger patients more. In Figure 4 it was shown that the patients who died with tumors from the second largest size class were younger than for the two smaller size classes. The largest size class did not show such a phenomenon, but it may also be the case that older patients are more prone to fatality as a consequence of surgery or other causes, including differences in cancer treatment. One detail of the dataset is that diagnosis codes are not defined, so indolence or aggressiveness of cancer types is not available for analysis. (Replacing the "DGN" variable with categorical dummy variables reduced AU-ROC (not shown).)

Being a continuous variable, any unclear patterns with age may hinder model accuracy. One solution is to divide ages into categories, create new binary dummy variables for predictive analysis, and then train the model with these in place of the continuous age variable. Figures 25-27 show decision tree-based SHAP outputs for three individuals whose data were tested after training the model on binned ages, with three equally sized bins representing younger, medium-aged, and older patients. The exact boundaries of bins designed in the manner applied here could vary with additional data, but with the current dataset the cutoff between younger and medium-aged patients is in the early 50's years of age, while the cutoff between medium-aged and older patients is about 70-years-old.

While multiple combinations of analyses have been run on these data (not all shown), the focus of Figures 25-28 is on output from the decision tree classifier, using binned ages and also with MI, PAD, and asthma removed from analysis. Removal of MI, PAD, and asthma had little effect on model metrics, but these factors have not seemed to contribute meaningful information to this analysis.

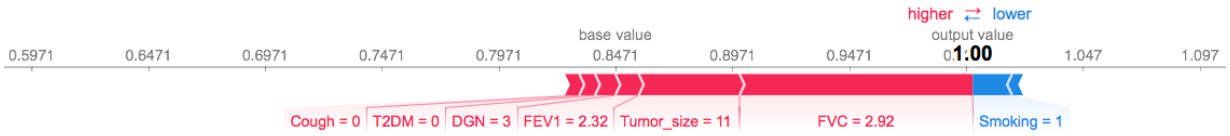

Figure 25. Decision tree-based SHAP output for a correctly predicted surviving patient, using binned ages and modeled without MI, PAD, and asthma.

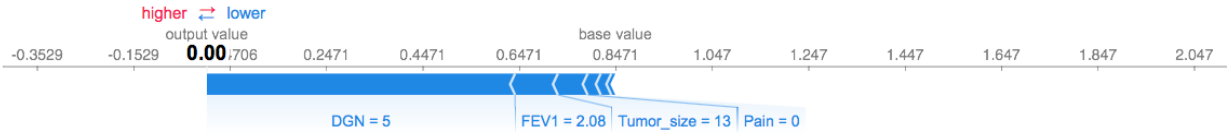

Figure 26. Decision tree-based SHAP output for a patient correctly predicted to not survive, using binned ages and modeled without MI, PAD, and asthma.

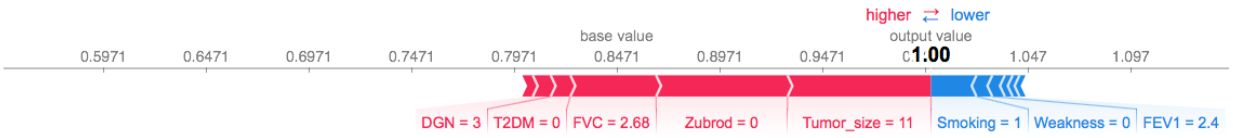

Figure 27. Decision tree-based SHAP output, failing to identify a patient who did not survive, using binned ages and modeled without MI, PAD, and asthma.

The decision tree classifier typically showed the highest AU-ROC score among classifiers for multiple analyses, and here varying between 0.7808 and 0.7911 for this set of features, with a test accuracy of 0.833. The average F1-score for the model presented in Figures 25-28 was 0.86. Precision and recall were 0.94 and 0.88, respectively, for the majority (survivor) class, versus 0.50 and 0.71, respectively, for the minority (non-survivor) class. In total, 84/97 survivors were correctly identified from test data, while

12/17 non-survivors were correctly identified. The random forest classifier showed a greater tendency to classify all patients as surviving and usually carried a lower AU-ROC score, even if accuracy was about the same as with the decision tree classifier.

In Figure 25 output is very similar to what is in Figure 20 for the same patient, though without the continuous variable age shown, the absence of cough shows up in the output as a feature contributing to survival. FVC shows a greater impact here. Figure 26 looks essentially the same as Figure 21 for the same patient, in both cases identified correctly as not surviving (and again with a lack of pain showing up as a feature not conducive to survival). Figure 27 still shows a patient who did not survive but was incorrectly expected to survive, as with Figure 22. This patient has many characteristics that are similar to those in Figure 25, not all appearing in the highlighted SHAP output, and many of which tending to be associated with a better outcome. It remains unclear why this patient did not survive.

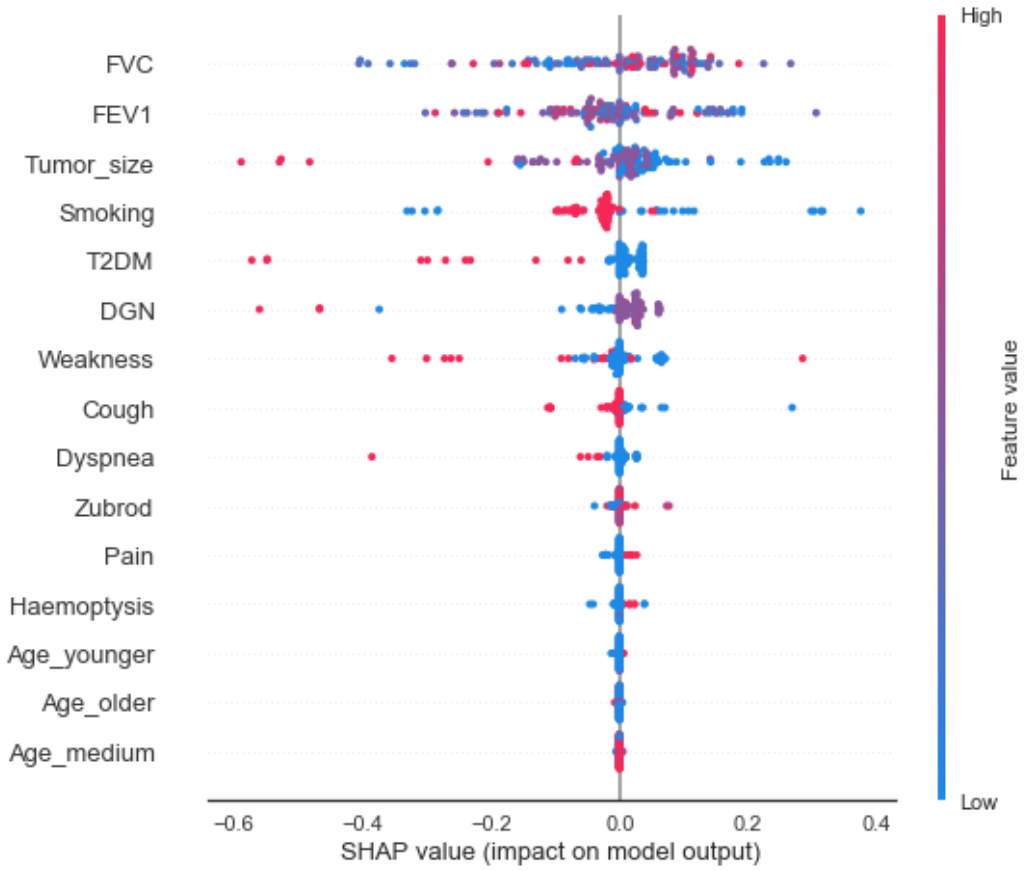

Figure 28. Decision tree-based SHAP global feature summary plot with binned ages and without MI, PAD, or asthma.

In Figure 28, with the continuous age variable replaced by bins, the importance of each age group appears very low compared with the other variables (and MI, PAD, and asthma are not included). In contrast to Figure 24, FVC and FEV1 are not at the top of the list, with tumor size just below. Overall the structure of the plot is similar to that in Figure 24, though the effects of individual feature values on SHAP output have changed slightly for some individuals.

A trend throughout models in this report has been that FVC and FEV1 are very important features in relation to survival among the patients involved in this dataset. These are the physiological characteristics most associated with pulmonary capacity. With every model the exact influence of any given FVC or FEV1 value is complex, as seen in Figure 28 in which some higher values for these features appear associated with survival, while in other individuals they are shown as associated with a prediction of non-survival, and vice versa. This complexity may arise from the statistically significant negative association between FVC and age (Figures 5, 13), with low FVC values in younger patients slightly more associated with mortality in this study (Figure 5). Also, it is unclear whether the given FVC and FEV1 values are averages of multiple samples per patient; since this testing requires patient intervention, it is possible some numbers to not represent a patient's true respiratory potential. Another possible indicator of respiratory function to consider that could be instructive to model development may be blood oxygen concentration, which requires less patient intervention.

Tumor size in Figure 28 clearly relates to survival, with larger tumor sizes contributing to lower survival and vice versa. Smoking generally poses a risk to survival in this patient population, though with the decision tree (unlike the overall less accurate random forest), a few data points in Figure 28 show the opposite pattern. Some lung cancer diagnostic codes may be more associated with smoking than others are, and perhaps the lack of smoking in a patient with an aggressive tumor that is typically associated with cancer puts the patient at a greater risk than most of the non-smoking cohort is and vice versa.

Presence of type II diabetes also shows a clear relationship to mortality in this study, while lack of this condition is positively associated with survival. Whether this relationship is stronger than in the general population amongst the age groups included here is unclear. Weakness, cough, and dyspnea are all mostly associated with poorer survival, while the lack of these conditions is marked by an improved prediction of survival. This is not surprising. Zubrod, pain, and haemoptysis, however, all show a trend opposite to expectations, with poorer performance and presence of pain or haemoptysis being associated with a prediction of survival. However, these three features bear very little predictive influence relative to others.

In a similar analysis of patient survival following thoracic surgery, 15,183 patients were observed in hospitals throughout France between 2002 and 2005 (Falcoz et al., 2013). Of these, 2.2% died during the hospital stay immediately following surgery. The research team developed a scoring system based on logistic regression, with multiple influential features. Important features that overlapped with the analysis presented here include age, dyspnea, pain, performance score, and diagnosis code. They examined FEV1 but found that less influential for their analysis. Important features Falcoz et al. identified that are not included in the dataset of the current study were sex, surgical priority, type of procedure, and comorbidity score (though some comorbidities are included here). However, Falcoz et al. examined immediate, in-hospital mortality risk with thoracic surgery, rather than evaluating the risk over the course of a year.

Perspective

It is inherently challenging to accurately predict mortality over the course of a year in patients who are not bed-bound and who may be subjected to a variety of stressors over the period of time. However, in this report predictive models with AU-ROC scores of around 0.78-0.79 and 83% test accuracy were able attainable. The decision tree classifier produced the greatest metrics overall, with the AU-ROC and accuracy scores stated here, but also with the best precision and recall scores for each target class among the classifiers tested. With both global and individual SHAP analyses, it is possible for the

components of a patient's risk to be evaluated. Since the model makes incorrect predictions approximately 20% of the time, it should not be solely relied upon, and patient caregivers should use their own judgment to make informed decisions regarding risks and benefits to a patient undergoing thoracic surgery or lung tumor resection. Some characteristics that are not included here, but which may be relevant, are other vitals that may indicate pulmonary function and other comorbidities beyond the few considered here.

While the decision tree classifier is emphasized here, the extra trees classifier also showed high AU-ROC scores for many tests. The SHAP python package is not tuned to this classifier yet, but if it becomes compatible with this classifier, it would be worthwhile to use SHAP to examine individual feature contributions for this classifier. Also, in various tests different features were removed or adjusted, and AU-ROC scores were fairly similar among many tests, so some flexibility exists around which model may be desired based on available features or interest.

References

Falcoz, PE, M Conti, L Brouchet, et al. 2007. The thoracic surgery scoring system (Thorascore): risk model for in-hospital death in 15,183 patients requiring thoracic surgery. *The Journal of Thoracic and Cardiovascular Surgery* 133:325-332.

Jose, BM. 2017. "Find Fraud in Enron Data." Web. Accessed: 7/12/18. https://bibinmjose.github.io/Enron_find_fraud/ .

Lundberg, S. "A unified approach to explain the output of any machine learning model." Web. Accessed: 7/12/18. https://github.com/slundberg/shap .

Miller, MR, J Hankinson, V Brusasco, et al. 2005. Standardisation of spirometry. *European Respiratory Journal* 26:319-338. Also available on the web: https://www.thoracic.org/statements/resources/pfet/PFT2.pdf .

University of California, Irvine, Machine Learning Repository. Thoracic Surgery Data Set. Web. Accessed: 6/20/18. http://archive.ics.uci.edu/ml/datasets/Thoracic+Surgery+Data . Citation associated with dataset: Zieba, M, JM Tomczak, M Lubicz, & J Swiatek. 2014. Boosted SVM for extracting rules from imbalanced data in application to prediction of the post-operative life expectancy in the lung cancer patients. *Applied Soft Computing* 14A:99-108. Principal investigator information for component datasets used in this study: M Lubicz (1), K Pawelczyk (2), A Rzechonek (2), and J Kolodziej (2):
-- (1) Wroclaw University of Technology, wybrzeze Wyspianskiego 27, 50-370, Wroclaw, Poland
-- (2) Wroclaw Medical University, wybrzeze L. Pasteura 1, 50-367 Wroclaw, Poland