JOURNAL OF INSURANCE MEDICINE

J Insur Med 2017;47:31–39

RESEARCH NOTES

Random Forest

Steven J. Rigatti, MD, DBIM, DABFM

For the task of analyzing survival data to derive risk factors associated with mortality, physicians, researchers, and biostatisticians have typically relied on certain types of regression techniques, most notably the Cox model. With the advent of more widely distributed computing power, methods which require more complex mathematics have become increasingly common. Particularly in this era of "big data" and machine learning, survival analysis has become methodologically broader. This paper aims to explore one technique known as Random Forest. The Random Forest technique is a regression tree technique which uses bootstrap aggregation and randomization of predictors to achieve a high degree of predictive accuracy. The various input parameters of the random forest are explored. Colon cancer data (n = 66,807) from the SEER database is then used to construct both a Cox model and a random forest model to determine how well the models perform on the same data. Both models perform well, achieving a concordance error rate of approximately 18%.

Address: 1295 State St, Springfield, MA 01111; ph: 413-744-8539; srigatti@massmutual.com

Correspondent: Steven J. Rigatti, MD, DBIM, DABFM, VP, Chief Medical Director, MassMutual.

Received: October 31, 2016

Accepted: November 30, 2016

The Random Forest Algorithm for Survival Analysis

In the clinical and insurance medicine, it is often desired to analyze a set of data in order to determine the relative relationships of a given vector of attributes (**X**) to an outcome (y). When that outcome is time to death or "failure" of a different sort, the most popular statistical technique is the Cox model.[1] The Cox model is popular because it does not make assumptions about the distribution of the predictor variables (is "non-parametric"), and it is able to deal with right-censored data – that death/failure need not have occurred for every individual in the study. Further, the Cox model is patterned after linear regression and its output is in the form of coefficients which may be easily transformed (exponentiated) to arrive at the hazard ratio; that is the expected multiplicative change in the hazard of the outcome per one-unit change in the predictor. Thus, the value of each predictor to the overall risk of the outcome is readily calculated.

However, there are some drawbacks to a Cox model. One is the proportional hazards assumption,[2] which states that the hazard of a given risk factor is proportional to the baseline hazard throughout the time of the study. This limitation is why the Cox model is often described as semi-parametric rather than non-parametric. Also, interactions between correlated predictors must be explicitly included in the Cox formula. For instance, if BMI and systolic blood pressure are both included in a formula to predict time to a cardiac event, an interaction (product) term should be included (SBP*BMI) to account for the fact that these variables may

have a more-than-additive effect on risk. If there are many variables, the number of interaction terms can be large and may include 3-way or higher level interactions. Also, nonlinear terms and/or splines need to be explicitly included and are warranted for those predictors with U or J-shaped relation to the outcome of interest.[3] If there are several such terms in each model, then the ease of interpretability may be lost. Finally, Cox models are not robust to missing data and some sort of imputation of data, or elimination of cases with missing data is often needed.

In the era of "big data", various other statistical techniques have been developed to analyze survival data. These include support vector machines (SVM), neural nets (NN) and random forests™ (RF).[4,5] The purpose of this article is to describe the random forest algorithm and to demonstrate its implementation on a survival analysis problem, while comparing its performance to that of a Cox model.

The Random Forest™ Algorithm

In much the way that Cox models are a further refinement of linear regression, RF is a further refinement of classification and regression tree (CART) models.[1] CART models are straightforward, non-parametric, and may be used in either classification (eg, determining whether someone has prostate cancer based on PSA, age and other variables) or regression (eg, determining the probability that someone has prostate cancer). The output of a CART model is often referred to as a decision tree. Such trees are familiar in the clinical literature as a way to guide diagnostic and treatment decisions. Mathematically, the CART algorithm selects that optimal decision threshold for a variable based on recursive partitioning, that is, testing the potential value of different thresholds for each predictor and implementing the most valuable until further splitting no longer improves discrimination. This is the way the tree is "grown." Botanical puns abound in the terminology of CART models – each decision node is a "branch," each terminal node a "leaf" and each starting node a "root."

Single-tree CART models are not generally up to the task of analyzing complex data with many interacting predictors. However, an assemblage of many such trees using different sets of predictors may, when aggregated, produce very good prediction results.

The random forest, developed by Breiman, uses randomization to create a large number of decision trees. The output of these trees is aggregated into a single output using voting for classification problems or averaging for regression problems. Randomization is implemented in 2 ways. First, the data set is sampled with replacement (bootstrap sampling). The process of aggregating a new sample this way is called "bootstrap aggregation" or "bagging." For instance, if there are 6 subjects in a study, 6 die rolls may be used to pick the new, randomized sample. There is no guarantee that every subject will appear in the new sample and some may appear more than once. In a large data set with n subjects, the probability of being left out of a bootstrap sample of size n converges on 1/e or about 37%. These left-out or "out-of-bag" subjects make up a useful set of data to test the decision tree developed from the in-sample subjects. The second randomization occurs at the decision nodes. At each node, a certain number of the predictors are chosen. For a set with p predictors a typical number is the rounded square root of p – though this parameter may be chosen by the analyst. The algorithm then tests all possible thresholds for all selected variables and chooses the variable-threshold combination which results in the best split – the split which most effectively separates cases from controls, for instance. This random selection of variables and threshold testing continues until either "pure" nodes are reached (containing only cases or controls), or some other pre-defined endpoint. This whole tree-growing process is repeated (commonly 100 to 1000 times) to grow the random forest.

Perhaps the greatest strength of the random forest is its ability to discover meaningful interactions and non-linear effects of the predictors. The nature of the interaction or non-linearity need not be specified ahead of time like is required for Cox models or other parametric survival models.

Once a random forest model is constructed, it can be used for prediction. Each new case is evaluated by each of the trees in the algorithm and the predicted outcome is, in the case of classification, the majority class (the one receiving the most "votes" from the individual trees) or the average of all predictions in the case of regression. Because the model is fairly complex in this way, it lacks the easy interpretability of a linear or Cox model. Because there are so many trees in the forest, it can be difficult to tell which variables are most strongly influencing the predictions. Fortunately, there are several mathematical constructs used to measure the impact of the variables in the model, so-called "variable importance measures." There are several such measures, and they may be based on the Gini coefficient, a measure related to the area under the curve (AUC).[7] Other methods test the model with the individual variables left out to see how the error rate is affected.

The error rate of a random forest is generally the average error rate of all the individual trees when the out-of-bag (OOB) cases are used as test cases. For classification problems, this is simple enough, since an erroneous class is easily identified. For regression, it is less simple, since a point prediction for a continuous value is, mathematically, always wrong. In this instance, the concordance (Harrell's C) is used. This compares the model output using pairs of cases. For instance, consider a model that predicts weight from height, zip code, lipids, and other variables. We then evaluate all pairs of test cases, and calculate the proportion of cases in which the model correctly predicted the heavier of the two individuals. This proportion is the concordance, and its complement (1-C) is the concordance error rate. This concordance error rate can be measured for Cox models as well, and can serve as a comparison between RF and Cox models evaluating the same data.

Survival Data

Random forests can be applied to right-censored survival data in much the same was as a Cox model. Recall that with right-censored data, not all the survival times are known. Right-censored cases survived to the end of the study, were lost to follow-up, or reached another defined end-point such as death from a cause other than that of interest.

There are a few special considerations when evaluating this type of data with a random forest model. One is that the splitting at each node is evaluated not by purity, but by the log-rank test, which evaluates the difference in two survival distributions. Another consideration is that, effectively, a new forest is fit for each unique survival time. This makes it computationally advantageous to round survival times to the largest useful interval – months instead of days, for instance. Finally, since there is no measure of node purity, the algorithm needs another indication of when to stop growing each tree. This can be chosen as a fixed number of cases or a fixed number of outcomes (deaths).

Tuning Parameters

When implementing a random forest, the analyst has a number of decisions to make regarding tree construction. These decisions can impact accuracy as well as computational intensity. One decision is how much data to feed in. Some studies have shown that random survival forests perform best when the number of deaths is approximately equal to the number of right-censored cases. Since most survival data in insurance industry in-force blocks contains many more survivors than deaths, it may be necessary to sub-sample the survivors. Another choice is how many variables to test at each split. The

default for this *mtry* variable is the aforementioned square root of the number of predictors. The number of trees, *ntree,* is directly relevant to computational intensity. On hundred to one thousand trees are often used, but smaller forests may still perform well. The maximum number of splits to try at each variable, *nsplit,* may be useful when there are many continuous predictors. This is because the default algorithm evaluates every possible split. So, for our weight prediction example, every unique height would be tested for its value as a threshold. For a large dataset, this may consume a great deal of computing time. Setting *nsplit* to a small, finite number such as 10, 50 or 100 can speed things up considerably. Finally, the variable importance measures rely on multiple iterations of the forest with and without each variable – so collecting these measures is intensive. One approach is to collect these measures during initial model construction and to leave them out after that.

Note that the names of the parameters above refer specifically to the Random Forest package[8] for the R statistical programming language. Other software packages or implementations may use different terminology.

Uses

The random forest algorithm has demonstrated robust utility across many fields of study, including medical research and "big data" applications. It is a common choice in the winning submissions for data analysis competitions such as those run by Kaggle (kaggle.com). For instance, the Mayo Clinic and University of Pennsylvania recently posted a challenge to data analysts on Kaggle: develop an algorithm which can detect seizures from intracranial EEG electrodes.[9] This algorithm would then be implemented in a responsive neurostimulation device to halt detected seizures. The winner of this competition, Michael Hills, used a random forest classifier to take the brain-wave pattern information and classify it into epileptiform vs non-epileptiform categories.

There are several examples in the medical literature where the random forest algorithm is used to evaluate stroke risk,[10] predict Alzheimer's onset,[11] automate the differentiation between melanoma and dysplastic nevi based on images.[12] and to detect bladder cancer based on urine metabolomic profiling.[13]

Example—*Titanic*

This simple example illustrates both simple CART analysis and random forests. In April 1912, the RMS *Titanic,* the largest passenger ship afloat the oceans, collided with an iceberg and sank off the coast of Newfoundland. She carried 2244 passengers and crew; 1500 of them were lost. The purpose of this data analysis example is to develop a prediction algorithm to correctly sort survivors from those lost at sea based on information available at the time of departure. The data is taken from the Kaggle web site (this example fairly closely follows the tutorial also to be found there).[14] The data includes an indicator of outcome (survived vs. lost), the passenger class (1, 2 or 3, where 3 = steerage), the embarkation point (S = Southampton, C = Cherbourg, and Q = Queenstown), age, sex, the price of the ticket, the total number of the passengers' siblings/spouses aboard, the total number of the passengers' children and parents aboard and the cabin number.

For a simple first prediction attempt, let us predict that all the women survive and all the men do not ("women and children first"). This yields a prediction accuracy of 81%. Next, a backward stepwise logistic regression model is fit. The result is a model with terms for age, sex, passenger class, embarkation point and the siblings/spouse variable. A prediction using this model is slightly worse than the "all women survive" model with a prediction accuracy of 80%. Next, a single classification tree is fit, see Figure 1. This utilized age, sex, passenger class, ticket fare, and the siblings/spouse variable – improving

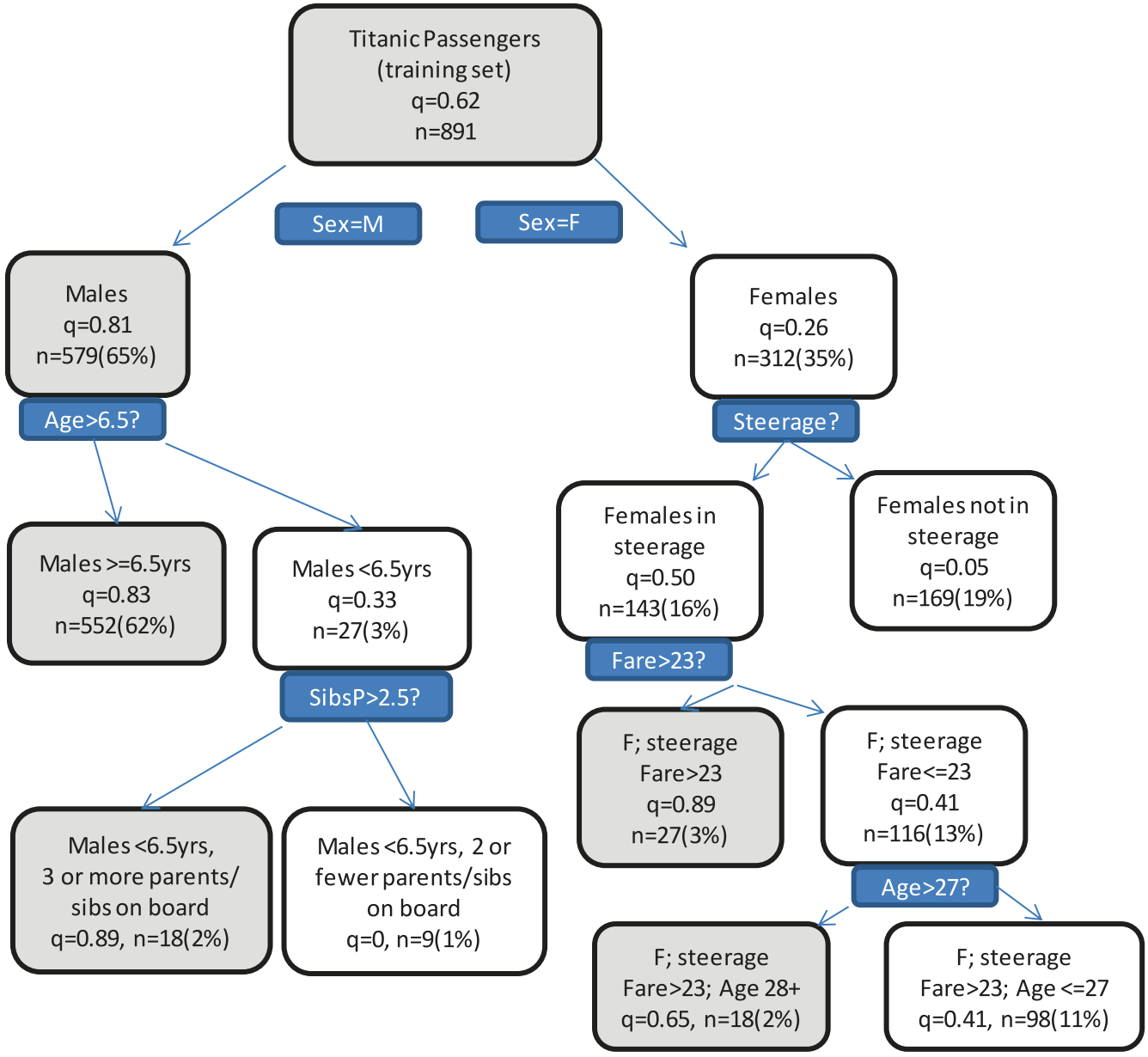

Figure 1. *Classification tree for determining survival aboard the RMS Titanic. For each decision point, the left branch represents a "Yes" answer to the decision criterion. Shaded boxes indicate a mortality of >50% for that group - and thus in the test set, this model would predict death for that group.*

the accuracy to 84%. Finally, a random forest is fit using 500 trees, trying 3 variables at each node, and testing a maximum of 5 thresholds for each variable. This improves the prediction accuracy to 89%. The variable importance testing shows that sex is the most important variable, followed by age, passenger class, ticket fare, parents/children aboard, siblings/spouse aboard, and the embarkation point. The implication here is that for a classification task, a large group of randomized decision trees can outperform a single "best" tree.

Random Survival Forest vs Cox: Colon Cancer

To evaluate the performance of the random forest method vs a Cox model on survival data, SEER data was downloaded using the

Table 1. Baseline Characteristics

n	66807
Sex (%M)	0.556
Age, mean (sd)	60.3 (5.6)
Year of dx:	2004-2011
AJCC 6th ed. Stage:	
0	2804(4.2%)
I	15781(23.6%)
IIA	13171(19.7%)
IIB	2263(3.4%)
IIIA	2179(3.3%)
IIIB	8975(13.4%)
IIIC	5923(8.9%)
IV	14256(21.4%)
Unknown	1455(2.2%)
T-Stage:	
T0	100(0.1%)
Tis	2884(4.3%)
T1	11758(17.6%)
T2	7485(11.2%)
T3	30408(45.6%)
T4	9492(14.2%)
TX	4620(7.0%)
N-Stage:	
N0	37353(55.9%)
N1	15070(22.6%)
N2	10566(15.8%)
NX	3758(5.6%)
M-stage:	
M0	50703(76.0%)
M1	14256(21.4%)
MX	1788(2.7%)
Grade:	
I	6169(9.2%)
II	41200(61.7%)
III	10240(15.3%)
IV	1064(1.6%)
Unknown	8134(12.2%)
Race:	
Black	10289(15.4%)
White	50189(75.1%)
Other	5691(8.5%)
Unknown	638(1.0%)

SEER*Stat program. Data were from the years 2004 through 2011 and included both males and females with colon cancer. There were no restrictions on stage, grade or histology. The characteristics of this group are listed in Table 1. The data was split into training (n = 61,807) and test (n = 5000) sets.

Table 2. Cox Model

	HR	Coef	Z	Pr(>|Z|)
Age	1.03	0.027	5.52	<0.0001
Sex (M vs. F)	1.22	0.200	3.76	0.0002
Year of Dx	0.99	−0.012	−0.87	0.383
Race:				
Black			ref	
Other	0.56	−0.578	−5.14	<0.0001
Unknown	0.19	−1.642	−2.31	0.021
White	0.73	−0.309	−4.62	<0.0001
T-stage:				
Tis			ref	
T0	7.55	2.021	5.15	<0.0001
T1	2.27	0.822	3.24	0.0012
T2	1.66	0.508	1.88	0.0596
T3	2.25	0.810	3.25	0.0012
T4	4.07	1.403	5.55	<0.0001
Tx	4.63	1.533	5.88	<0.0001
N-stage:				
N0			ref	
N1	1.38	0.326	4.37	<0.0001
N2	2.05	0.716	8.84	<0.0001
NX	1.85	0.615	4.99	<0.0001
M-stage:				
M0			ref	
M1	5.70	1.740	25.83	<0.0001
MX	1.50	0.403	2.61	0.009
Grade:				
I			ref	
II	1.32	0.277	2.21	0.0271
III	1.67	0.515	3.8	0.0001
IV	2.40	0.874	4.04	<0.0001
Unknown	2.17	0.775	5.5	<0.0001

A Cox model was run on the training data, using all available variables including age, sex, race, stage and grade. Stage was broken up into T-stage, N-stage and M-stage, and these were treated as separate variables. No interaction or non-linear terms were included in the model specification. The output of the Cox model is shown in Table 2. Note that when a factor variable is included, one level of that variable must be specified as the baseline or referent. For example, the M-stage variable has 2 levels, M0 and M1. The M0 level is the referent, and the hazard ratio associated with M1 is 5.7, implying that those with metastases die at a rate that is 5.7 times higher than those with no metastases,

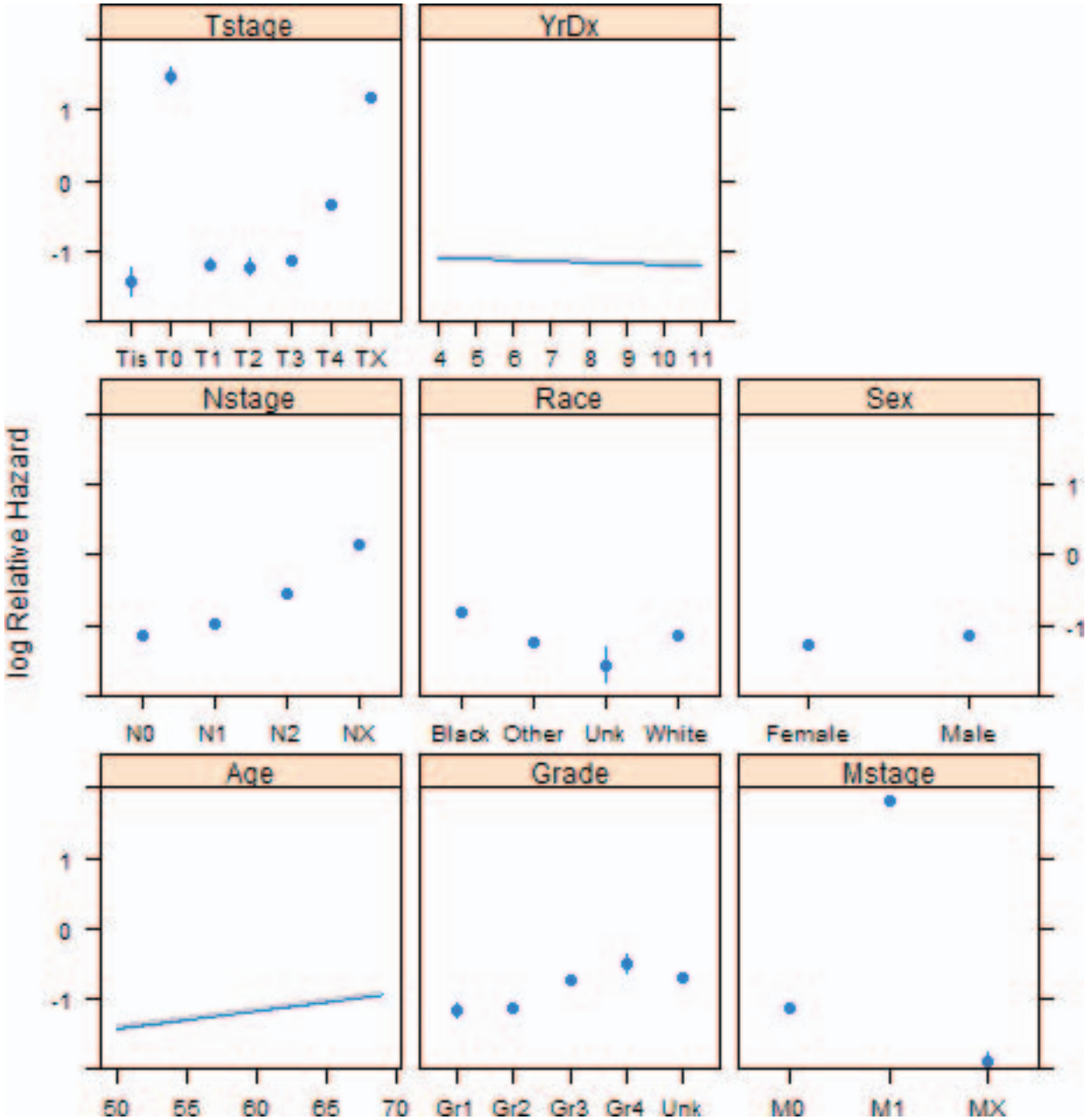

Figure 2. *Partial effects plot.*

controlling for age, sex, race, T-stage, N-stage and grade. Also, note that T0 is associated with a rather high risk, since such cases would necessarily be node-positive or metastatic – that is, the primary tumor was occult. See the partial effects plot in Figure 2.

Next, a random forest model was fit to the same training data with 500 trees, 5 variables at each node and 5 possible splits for each variable. There are a few different methods for determining how good a particular split is. For this experiment, 2 methods were tested, the log-rank test previously described and the log-rank score, which is a faster version of the other. The methods turned out to be nearly equivalent, so only the log-rank score results are reported here. The variable importance measures are displayed in Table 3.

The random forest variable importance is

Table 3. RSF Variable Importance

Variable	
M0	0.0803
M1	0.0741
N0	0.0235
T3	0.0226
T4	0.0180
T1	0.0143
Tx	0.0126
Race: Black	0.0107
Grade2	0.0098
N2	0.0093
Grade3	0.0092
T2	0.0089
Race: White	0.0087
Grade: Unknown	0.0080
N1	0.0077
Age	0.0067

based on how well a random forest without a particular variable performs – so there is some redundancy when a factor has only 2, or just a few levels. For example, M0 and M1 have similarly high variable importance since they are mutually exclusive and thus contain the same information. Note also that the actual variable importance number is essentially meaningless; it does not provide information about the direction of the variable's influence. We cannot tell from this measure whether M1 is associated with higher or lower mortality, just that it is very influential.

To compare the 2 methods, we rely on the concordance error rate of the test set. For the random forest algorithm, the concordance error rate was 18.43%; while for the Cox model, it was 18.19%. This would suggest that the 2 methods are roughly equivalent predictors of mortality risk for this data. However, Cox has clear benefits over RSF in the areas of interpretation, simplicity and ease of calculation.

DISCUSSION

The random forest method of analysis is an increasingly common technique used to evaluate survival data. Though it is capable of highly accurate predictions, its implementation is both computationally intense and difficult, given the dependence on several input parameters. One other drawback of the method is the "black box" nature of the output. Whereas a Cox model or a parametric survival model such as a Poisson or Gompertz model generates coefficients and other parameters that lend insights into the partial effects of the predictors, the random forest provides no such parameters, and it is, therefore, much more difficult to glean understanding of how the predictors may be individually or collectively influencing the prediction.

However, these hurdles are not insurmountable. With thorough testing of a given random forest model, one could generate partial effect plots. Though, with very large data sets and many predictors, it may be hard to choose which variables to plot. Also, if one attempts to hold all non-plotted variables constant, some of the flexibility of the random forest algorithm is lost.

In the sample problem evaluated here, the Cox and random forest models performed similarly, so the interpretability of the Cox model makes it the preferred method in this circumstance. However, this may have been due to the nature of the data. With only a few predictors and no obvious interactions or nonlinear effects, the random forest model could not display its strengths.

Files

The data and R files used to analyze the data are available from the author upon request.

REFERENCES

1 Cox, DR. Regression Models and Life-Tables. *J Royal Statistical Society*. 1972;Series B 34(2):187–220.
2 Brant, R. Assumptions of the Cox Model. University of British Columbia Department of Statistics. Available at: http://www.stat.ubc.ca/~rollin/teach/643w04/lec/node69.html. Accessed March 24, 2004.

3 Wesley D, Cox HF. Modeling Total Cholesterol as Predictor of Mortality: The Low-Cholesterol Paradox. *J Insur Med*. 2011;42:62-75.
4 Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning.* New York, NY: Springer New York Inc. 2001.
5 Breiman, Leo. Statistics Department, University of California, Berkeley. *Random Forests.* Available at: https://www.stat.berkeley.edu/~breiman/randomforest2001.pdf. Accessed September 2016.
6 Decision Tree Learning. Available at: https://en.wikipedia.org/wiki/Decision_tree_learning. Accessed September 2016.
7 Receiver Operating Characteristic. Available at: https://en.wikipedia.org/wiki/Receiver_operating_characteristic. Accessed September 2016.
8 Breiman L, Cutler A, Liaw A, Wiener M. Available at: https://cran.r-project.org/web/packages/randomForest/randomForest.pdf. Accessed September 2016.
9 Available at: https://www.kaggle.com/c/seizure-prediction. Accessed September 2016.
10 Chen R, Deng Z, Song Z. The prediction of malignant middle cerebral artery infarction: a predicting approach using random forest. *J Stroke Cerebrovasc Dis*. 2015;24(5):958-964.
11 Lebedev AV, Westman E, Van Westen GJ, et al. Alzheimer's Disease Neuroimaging Initiative and the AddNeuroMed consortium. Random Forest ensembles for detection and prediction of Alzheimer's disease with a good between-cohort robustness. *Neuroimage Clin.* 2014;6:115-125.
12 Rastgoo M, Garcia R, Morel O, Marzani F. Automatic differentiation of melanoma from dysplastic nevi. *Comput Med Imaging Graph*. 2015;43:44-52. Epub 2015 Mar 7.
13 Wittmann BM, Stirdivant SM, Mitchell MW, et al. Bladder cancer biomarker discovery using global metabolomic profiling of urine. *PLoS One*. 2014;9(12).
14 Titanic: Machine Learning from Disaster. Available at: https://www.kaggle.com/c/titanic/data. Accessed July 2015.